Lecture Notes in Computer Science

The LNCS series reports state-of-the-art results in computer science research, development, and education, at a high level and in both printed and electronic form. Enjoying tight cooperation with the R&D community, with numerous individuals, as well as with prestigious organizations and societies, LNCS has grown into the most comprehensive computer science research forum available.

The scope of LNCS, including its subseries LNAI, spans the whole range of computer science and information technology including interdisciplinary topics in a variety of application fields. The type of material published traditionally includes

- proceedings (published in time for the respective conference)
- post-proceedings (consisting of thoroughly revised final full papers)
- research monographs (which may be based on outstanding PhD work, research projects, technical reports, etc.)

More recently, several color cover sublines have been added featuring, beyond a collection of papers, various added-value components; these sublines include

- tutorials (textbook-like monographs or collections of lectures given at advanced courses)
- state-of-the-art surveys (offering complete and mediated coverage of a topic)
- hot topics (introducing emergent topics to the broader community)

In parallel to the printed book, each new volume is published electronically in LNCS Online at *http://link.springer.de/series/lncs/*.

Detailed information on LNCS can be found at the series home page *http://www.springer.de/comp/lncs/*.

Proposals for publication should be sent to the
LNCS Editorial, Tiergartenstr. 17, 69121 Heidelberg, Germany
E-mail: lncs@springer.de

Volume di riferimento
AZ 1216

ISSN 0302-9743

Lecture Notes in Computer Science

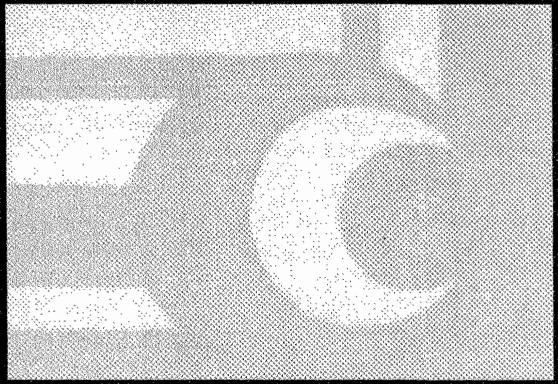
Gelbukh (Ed.)

LNCS 2004

Alexander Gelbukh (Ed.)

Computational Linguistics and Intelligent Text Processing

Second International Conference, CICLing 2001
Mexiko City, Mexico, February 2001
Proceedings



References

1. Apresian, Y.D., Cinman L.L. *Computer-aided periphrasing*. In: Semiotics and Informatics, Issue 36, Moscow, 1998.
2. Bolshakova, E.I., Vasilieva N.E. *On the problem of computer-aided literary and scientific editing*. Proceedings of International Workshop on Computational Linguistics and its Applications Dialog2000. Russia, Protvino, 2000.
3. Meyer, I. *Knowledge Management for Terminology-Intensive Applications: Needs and Tools*. In: Lexical semantics and Knowledge Represantation, First SIGLEX Workshop, J. Pustejovsky and S. Bergler (Eds.), Springer, 1991.
4. Olson, J., Rueter H. Extracting expertise from experts: methods for knowledge acquisition. Expert systems, Vol. 4, 1987, No. 4.
5. Paice, C.D., Jones P.A. *The Identification of Important Concepts in Highly Structured Technical Papers*. In: Proc. of the Sixteenth Annual Int. ACM SIGIR Conference on Research and Development in Information Retrieval, Pittsburg, 1993.
6. Pearson, J. *Terms in Context*. In: Studies in Corpus Linguistics. John Benjamins, Amsterdam, Vol. 1, 1998.
7. Penagos, C.R. *Extraction of Knowledge about Terms from Indications of Metalinguistic Activity in Texts*. Proceedings of Int. Conf. On Intelligent text processing and Computational Linguistics CICLing-2000, Mexico, 2000.
8. Pshenichnaya, L.E., Corenga O.N. *Scientific Term in a Dictionary and in a Text*. Nauchno-Texnicheskaya Informatciya (Scientific and Technical Information), 1991, No. 12.
9. Sager, J.C. *A Practical Course in Terminology Processing*. John Benjamins Publ. House, 1990.
10. Salton, G. *Automatic Text Processing: the Transformation, Analysis, and Retrieval of Information by Computer*. Reading, MA: Addison-Wesley, 1998.
11. Senkevich, M.P. *Style of scientific speech and literary editing of scientific works*. Moscow, Vysshaia Shkola, 1976.
12. Skorokhod'ko, E.F. *Semantic Complexity of Word (Term): Network Parameters and Communicative Characteristics*. Nauchno-Texnicheskaya Informatciya (Scientific and Technical Information), 1995, No. 2.
13. Zobel, J. *Writing for Computer Science*. Springer, 1997.

Lexical-Semantic Tagging of an Italian Corpus

Nicoletta Calzolari [1], Ornella Corazzari [2], Antonio Zampolli [1]

[1] Istituto di Linguistica Computazionale (ILC), Via Alfieri, 1
56010 Pisa, Italy
{glottolo, pisa}@ilc.pi.cnr.it

[2] Consorzio Pisa Ricerche (CPR), Piazza A. D'Ancona, 1
56100 Pisa, Italy
corazzar@ilc.pi.cnr.it

Abstract. Semantically tagged corpora are becoming an urgent need for training and evaluation within many applications. They are also the natural accompaniment of semantic lexicons, for which they constitute both a useful testbed to evaluate their adequacy and a repository of corpus examples for the attested senses. It is essential that sound criteria are defined for their construction and a specific methodology is set up for the treatment of various semantic phenomena. We present some observations and results concerning the lexical-semantic tagging of an Italian corpus within the framework of two projects: the ELSNET feasibility study, part of a preparatory phase started with Senseval/Romanseval, and an Italian National Project (TAL), where one of the components is the lexical-semantic annotation of larger quantities of texts for an Italian syntactic-semantic Treebank. The results of the ELSNET experiment have been of utmost importance for the definition of the technical guidelines for the lexical-semantic level of annotation of the Treebank.

1 Introduction

In this paper we present some observations and results concerning the manual lexical-semantic tagging of an Italian corpus performed within the framework of the ELSNET project and of an Italian National Project (TAL).

The ELSNET experimental project, as a feasibility study, was part of a preparatory phase started with the SENSEVAL/ROMANSEVAL initiative [4]. Given its preparatory nature, we decided to focus the lexical-semantic annotation on the predicate-argument part of the sentences, which can be considered the core of a sentence and is crucial for semantic interpretation. The ELSNET corpus is composed of 1000 contexts of 20 verbs (50 contexts for each verb) extracted from the journalistic section of the Italian PAROLE corpus [16]. The semantic annotation of both verb senses and their argument heads allows i) to consider the disambiguation task from the perspective of the semantic relations holding between verb and arguments (e.g. to what extent the disambiguation of one of the two has an impact on the disambiguation of the other), ii) to analyze different aspects of verb semantics (e.g. the possibility to draw a list of typical semantic subjects vs. objects of a verb objects which combine with it and vice-versa; the adequacy of the "semantic types" of the reference lexicon with respect to the previous tasks, etc.).

In TAL, a multi-level annotated corpus, the Italian Syntactic-Semantic Treebank (ISST) [15], is now being created. The final and tested version of ISST will be available in 2001. ISST has a three-level structure ranging over syntactic and semantic levels of linguistic description. Syntactic annotation is distributed over two different levels, the constituent structure level and the functional relations level: constituent structure is annotated in terms of phrase structure trees reflecting the ordered arrangement of words and phrases within the sentence, whereas functional annotation provides a characterisation of the sentence in terms of grammatical functions (i.e. subject, object, etc.). The third level deals with lexical-semantic annotation, carried out in terms of sense tagging augmented with other types of semantic information. The three annotation levels are independent of each other, and all refer to the same input, namely a morpho-syntactically annotated (i.e. pos-tagged) text which is linked to the orthographic file with the text and mark-up of macrotextual organisation (e.g. titles, subtitles, summary, body of article, paragraphs). The final resource will be available in XML coding. The multi-level structure of ISST shows a novelty with respect to other treebanks: it combines within the same resource syntactic and lexical-semantic annotations, thus creating the prerequisites for corpus-based investigations on the syntax-semantics interface (e.g. on the semantic types associated with functional positions of a given predicate, or on specific subcategorization properties associated with a specific word sense). ISST corpus consists of about 300,000 word tokens reflecting contemporary language use. It includes two different sections: 1) a "balanced" corpus, testifying general language usage, for a total of about 210,000 tokens; 2) a specialised corpus, amounting to 90,000 tokens, with texts belonging to the financial domain. Finally, information stored in ISST will be used, for "external" evaluation within the TAL project, to improve an automatic Italian-English translation system.

In this paper we focus on: i) the methodology for lexical semantic tagging and the strategies for the treatment of some phenomena relevant to this level of annotation (such as titles, proper nouns, idioms etc.); ii) some interesting aspects emerged from the analysis of the annotated verbs and their argument heads (e.g. the usefulness of using a lexicon enriched with semantic types). Finally, some observations are provided about the limits of lexical-semantic annotation, in other words, about what cannot be expressed through lexical tagging.

2 A Brief Description of the Annotation Strategies

2.1 The ELSNET Experiment

The ELSNET experiment was performed through different steps:

- verb selection: verbs were selected to represent different semantic fields (e.g. speech acts (*chiedere, chiamare*), movement (*entrare, portare*), perception (*vedere*), etc.), and various subcategorization properties (transitive, intransitive, reflexive, etc.);

- verb context selection: contexts were selected to illustrate the different meanings of a verb, and to display a significant variety of argument heads for each verb sense;
- corpus annotation at three different levels of description: morpho-syntactic, functional, lexical-semantic.

At the lexical-semantic level, the corpus annotation was manually performed and consisted in both sense-tagging and semantic-tagging [12]. By sense-tagging we mean the assignment, to corpus occurrences, of the appropriate sense taken from the ItalWordNet(IWN)/EuroWordnet lexicon [2]. By semantic-tagging we mean the assignment, to corpus occurrences, of the appropriate semantic type/concept (such as "human, animal", etc.) as defined within the SIMPLE lexicon [13]. The combined use of both the IWN lexicon and the SIMPLE ontology of semantic types was decided in order to allow future comparisons of the two types of annotation and evaluation of the disambiguating power of the SIMPLE semantic types.

2.2 The ISST Annotation Methodology

In ISST, lexical-semantic annotation consists in the assignment of semantic tags, expressed in terms of attribute/value pairs, to full words or sequences of words corresponding to a single unit of sense (e.g. compounds, idioms). Annotation is restricted to nouns, verbs and adjectives and corresponding multi-word expressions.

ISST semantic tags convey three different types of information:

- sense of the target word(s) in the specific context: IWN is the reference lexical resource used for the sense tagging task;
- other types of lexical-semantic information not included in the reference lexical resource, e.g. for marking of figurative uses;
- information about the tagging operation, mainly notes by the human annotator about problematic annotation cases.

It is worth noting that, through the taxonomic organisation of IWN, an implicit assignment is made also of the semantic types of the IWN top-ontology. In this way, ISST sense tagging can also be seen as implicit semantic tagging, and an evaluation can later be done of the level of granularity needed in an ontology e.g. in order to discriminate between different senses of the same word or to express selection preferences on arguments.

Starting from the assumption that senses do not always correspond to single lexical items, the following typology of annotation units is identified and distinguished:

- **us**: sense units corresponding to single lexical items (nouns, verbs or adjectives);
- **usc**: semantically complex units expressed in terms of multi-word expressions (e.g. compounds, support verb constructions, idioms);
- **ust**: title sense units corresponding to titles of any type (of newspapers, books, shows, etc.). Titles receive a two-level annotation: at the level of individual components and as a single title unit.

As to the annotation methodology, in order to ensure that polysemous words and usc are tagged consistently, the annotation is manually performed 'per lemma' and no arbitrary sense assignments are avoided by resorting to under-specification, expressed in terms of disjunction/conjunction over different IWN senses.

3 Treatment of Some Problematic Cases

It is obviously of utmost importance to set up an annotation strategy for semantic phenomena such as idiomatic expressions, compounds, etc., when a sense does not correspond to one single orthographic word. The ELSNET experiment was useful to highlight the issues to be considered and solved for semantic tagging of the larger ISST corpus. In the specifications of ISST criteria are given for idioms, compounds, figurative uses, evaluative suffixation, proper nouns, foreign words, titles, etc. [18]. For the treatment of these phenomena sense assignment is augmented through specification of lexical-semantic tags conveying information not explicitly included in the reference lexicon.

3.1 Compounds

Compounds are treated as a single unit. This treatment is justified from a linguistic point of view because in most cases they are not semantically compositional or they are only partially compositional (e.g. *un_filo_di_continuità, professore_d'orchestra, compagnia_di_prosa, ombrello_antimissile, alta_moda*).

3.2 Proper Nouns

Proper nouns are assigned a semantic type (such as "human, artifact, institution"), e.g. *Francia* is tagged as "place". Proper nouns composed by two or more lexical items are treated as one entry (e.g. *Pippo_Baudo, Incisa_della_Rocchetta* (proper noun), *Teatro_Stabile_delle_Erbe* (theatrical company/troupe), *Amici_ della_ farsa* (theatrical company/ troupe).

3.3 Titles

Titles composed by more than one lexical item are compositional sequences, and the single components should be semantically annotated to allow e.g. information retrieval queries not only on the titles as such but also on the internal components of the title. In the experimental phase, titles were marked only as single units in order to simplify the annotation strategy, while in ISST they are annotated both at the level of the single components and as a unique sequence, identified by a specific tag. Their identification at the semantic level is desirable at least for the following reasons: i) for linguistic reasons, to obtain more coherent data (e.g. considering the sentence *pubblicare* (publish) *"I fiori del male"*, if titles were not annotated one could draw the wrong conclusion that *pubblicare* can have as object a "flower/natural kind", in addition to "book/title/semiotic artifact"; ii) for translation purposes, because titles frequently have no literal/equivalent translation or are left in their original language Few corpus examples are: *Ditegli_ sempre_ di_ si* (title of a show) *Si_recita_Feydeau* (title of a show), *Il_Corriere_della_Sera* (title of a newspaper).

3.4 Figurative Uses and Idiomatic Expressions

Figurative uses and idiomatic expressions in general are marked with specific features. Their identification is important at least: i) for machine translation, since in many cases they have no exact lexical and - as far as idioms are concerned - structural equivalents; ii) for linguistic acquisition purposes, to obtain a correct data extraction (e.g. in the sentence *non comprendo la* ***molla*** *di una simile violenza* 'I don't understand the reason of such a violence', the extraction of the objects of *comprendere* 'to understand' would lead to the wrong conclusion that one of its typical objects is an "artifact" (*molla* 'spring') of type "product" (some artifacts indeed can be used in this position: *non comprendo i suoi dipinti/libri*, but they are "artwork/semiotic_artifact"); iii) for lexicographic purposes, to extend existing computational lexicons with new idioms, collocations, lexicalized metaphors, and allow studies on them.

Metaphors. Examples from the ELSNET corpus are: *abbandonare la* ***passerella*** *dell'alta moda* 'abandon the haute couture', *questo tenore…è arrivato fino alle* ***vette*** 'this tenor…is arrived till the top', *abbandonare la* ***strada*** *dello sport* 'abandon the road of the sport'. The distinction between lexicalized and non lexicalized metaphors was ignored in the experimental project, while it is taken into account in ISST, where the figurative uses are marked with a specific feature (FIG.=metaf). Non lexicalized metaphors are always linked to the literal sense.

Metonymy. Metonymy, that raises the same problems of data interpretation as the other figurative uses, is also marked by a specific feature. For example, in the corpus context *conquistare l'argento* 'to win the silver', *argento* is annotated as metonymy.

Evaluative Suffixation. Similar observations hold for semantic modification conveyed through evaluative suffixation: non lexicalized cases are linked to the relevant sense of the stem word, e.g. *porticciolo* 'small port', *borsone* 'large bag'.

Idiomatic Expressions and Complex Units. Idiomatic expressions are treated as a single word and marked with a specific feature, e.g.: *il processo* ***entra nel vivo*** 'to enter into the heart of the process', ***cartone animato*** 'cartoon', ***aprire un*** *nuovo* ***capitolo*** *nell'industria* lit.: 'to open a new chapter in the industry', ***tagliare la testa al toro*** lit.: 'to cut the head of the bull'. Corpus annotation also identifies expressions that are not recognized as such in IWN, but behave as semantically complex units This is the case e.g. of *anni Sessanta* 'the sixties' which, being fully compositional and productive, does not appear in the lexical resource.

ense assignment combined together with the additional lexical-semantic tags make ıe ISST annotated corpus more than a mere list of instantiations of the senses :tested in the reference lexical resource. The corpus becomes a repository of ıteresting semantic information (going from titles and proper nouns to non-xicalized metaphors, metonymies and evaluative suffixation), especially for what ›ncerns non-conventional uses of a word, i.e. those semantic facts which are ɛcluded – either programmatically or just by chance – from the reference lexical ·source. Corpus annotation can also shed light on the variability of multi-word ɛpressions - from compounds to support verb constructions and idiomatic ɛpressions -, that are prone to massive variation. The gray areas spotted by the above :amples in which corpus annotation either diverges from the lexical resource or rther specifies it, can be seen – in perspective – as the starting point for revisions ıd refinements of both the annotated corpus and the lexicon. In this way, the ınotated corpus presents itself as a flexible resource, which is – to some extent – dependent from the specific internal architecture of the selected reference lexicon.

Some Remarks about the Annotated Verbs and Argument Heads

e report here some observations stemming from the semantic annotation of verbs d their arguments, and we touch issues such as the possibility to characterize typical bjects/objects combining with a given verb sense, the usefulness of a lexicon riched with semantic types and/or collocations, criteria for disambiguating senses.

l Typical Semantic Arguments of a Verb

om the analysis of the semantically annotated corpus, it turns out that there are rious ways of describing - in terms of semantic types - a typical argument of a ⁄en verb sense. The arguments combining with a verbal head can be:

semantically restricted: in this case, it is possible to define the specific semantic types which combine with it (*selection restrictions*);

semantically completely unrestricted (*no selection restrictions*);

semantically unrestricted, but it is possible to define which semantic types cannot combine with it (it is particularly relevant when this allows to discriminate between different senses of the same verb) (we could call it a *negative restriction*);

partially semantically restricted: a list of preferences in terms of semantic types can be defined (*selection preferences*).

t us consider as illustrative example the verb *arrestare*.

In the first sense the verb means 'to stop'. According to the annotated corpus its ıical arguments are of the following semantic types:
›j=act; cause_act; natural_substance; purpose_act; time
bj=non_relational_act; change_of_value; movement_of_thought; act; cause_act; ıse_natural_transition; event. In many cases the object has a *negative connotation*.

Summing-up, the sense 'to stop' selects almost unrestricted subjects and direct objects. However, the subject is preferably non-human (it is rarely human, as in *il governo ha arrestato l'inflazione* 'the government stopped inflation'), while the direct object is preferably an "event; act; change;...", but it is usually not a human or human-like (human-group, institution, etc.). Moreover the object has preferably a negative connotation. All this can be broadly expressed in the following way.

Table 1. Arg.s description of *arrestare1*

SUBJ:
preference= non-human;
DOBJ:
preference= event; act; change; phenomenon
preference= negative connotation
negative_restriction= human

2. The second meaning of the verb is 'to arrest'. Semantic types of the arguments are:
subj= human; human_group; institution; profession, with **domain**=military; law
dobj=human; agent_of_temporary_activity; agent_of_persistent_activity; kinship; profession; people. In many cases the direct object has a *negative connotation.*

This sense clearly selects a human or human-like subject and direct object. The subject preferably belongs to the military/law domain, while the object preferably has a negative connotation (not always however, e.g. *arrestare un innocente* 'arrest an innocent').

Table 2. Arg.s description of *arrestare2*

SUBJ:
selection_restriction= human or human-like
preference= domain=law, military
DOBJ:
selection_restriction= human
preference= negative connotation

Another example is the verb *percepire*. Its direct objects can be described as follows:

1. The first meaning is 'to perceive'. This sense is marked as a "perception" type.
dobj= color; group; shape; sign; phenomenon

2. The second sense is 'to receive' and is marked as "change_possession".
dobj= money; convention; number; amount

3. The third is a figurative use ('to perceive with the intuition', 'to perceive something as if it is something else') and is marked as "perception figurative".
dobj= unrestricted

This last sense frequently - but not always - occurs with a complement introduced by *come* (e.g. *l'opinione pubblica percepisce il Servizio sanitario nazionale (Ssn) come poco efficiente* 'public opini

Summing-up, for *percepire* only the second sense seems to have a semantically restricted direct object, while the third meaning is frequently marked by a specific (preferred) syntactic pattern.

4.2 Verb/Arguments Interaction at the Lexical-Semantic Level

The interpretation/disambiguation of the sense of a given argument head may strongly depend on the meaning of the surrounding context, more precisely of the verbal head. Between the verb and its arguments there is a strong interaction from the semantic point of view: the verb meaning may determine (or select) the sense of its subject and/or direct object. For instance *arrestare*, both 'to arrest' and 'to stop', as said above frequently selects direct objects which have themselves, or receive from the verb, a negative connotation, as shown in the corpus examples below.

Table 3. Dobj of the verb *arrestare*

Dobj	Sem.type of Dobj	Conn. Feat.
ladro_1	agent_temp_act	neg
spacciatore_1	agent_temp_act	neg
trafficante_1	agent_temp_act	neg
traffico_2	act	neg
invasione_1	cause_act	neg
massacro_1	cause_nat_trans	neg
inflazione_1	event	neg
pregiudicato_1	human	neg
balordo_1	human	neg
maniaco_1	human	neg
strozzino_1	agent_temp_act	neg

Another example is *comprendere* which, in the meaning of 'to include', selects a specific sense of its subjects. For the lemmas below, the sense marked in the SIMPLE lexicon as "group of entities" (which can therefore 'include' other entities) is selected.

Table 4. Dobj of the verb *comprendere*

Dobj	Sem.type of Dobj
carico_1	group
elenco_1	group
equipaggiamento_2	group
lista_2	group
panorama_1	group
tris_1	group
comune_1	human_group
consiglio_2	human_group
costituente_2	human_group
dossier_1	group

It may also happen that the sense/semantic type of the direct object determines the meaning of the verb. For instance, the semantic type of the objects helps to characterize different possible senses/nuances of meaning of the verb *coprire,* as shown in the tables below.

– *coprire un periodo* 'to cover a period of time':

Table 5. Sem. type of Dobj of *coprire*

Dobj	Sem.type of Dobj
periodo_1	time
1970-1993	time

– *coprire uno spazio* 'to cover a space/a distance':

Table 6. Sem. type of Dobj of *coprire*

Dobj	Sem.type of Dobj
superficie_1	area
territorio_1	area
area_2	area
area_1	area
pista_1	artifactual_area
continente_1	geopolitical_location
80_per_cento	part
35%	part

– *coprire un suono* 'to smother':

Table 7. Sem. type od Dobj of *coprire*

Dobj	Sem.type of Dobj
rumore_1	experience_sound

– *coprire una persona/un reato* 'to hide a crime' :

Table 8. Sem. type of Dobj of *coprire*

Dobj	Sem.type Of Dobj	Conn. Feat.
crimine_1	act	neg
mafioso_2	human	neg
violento_1	human	neg

4.3 Acquisition of Senses and Enhancement of Existing Lexical Resources

The analysis of a semantically tagged corpus allows not only to identify totally new senses, but also to have a more precise and complete which senses to encode for a lemma. Relying on the different semantic types of rgument heads that combine with a given verb, it is easier to identify the most eneral senses of a lemma and to capture the most specific senses or shifts of meaning f the same lemma. One can then decide how to collapse some specific uses into nore general, inclusive senses according to:

- different design requirements of the lexical resource to be created/extended/tuned. Indeed, both the number and type of senses to be encoded may strongly depend on the 'apparatus' (information types and representation means) available to describe them, i.e. semantic nets, frames, selection restrictions, ontology, domain, semantic relations, etc.
- different applications of the lexical resource (e.g. MT, IR, etc.). For instance, in an MT environment (bilingual, multilingual resources), it makes sense to treat as independent meanings those that have a different translation (e.g. also the sense number 9 below, among others, for the language pair Italian/English), but not necessarily for IR, where an excessive granularity may create noise.

is important that corpus analysis does not lead to an excessive granularity of sense istinctions - not desirable for different reasons [4], [9] -, but that it provides ground r decisions based on actual evidence. We give below the example of the verb *bbandonare* 'to abandon/leave', which has at least the following three main senses ccording to current paper dictionaries: 1) to abandon, to leave forever (e.g. a place),) to abandon, to desert (e.g. the children), 3) to give up, to renounce. On the basis of e analysis of the semantic types of direct objects, the following major/minor senses ses) of *abbandonare* come out:

- 'to leave a place': dobj= building, geopolitical_location, area
- 'to get up' (*abbandonare la sedia, un veicolo*): dobj= furniture, vehicle
- 'to abandon someone'': dobj= kinship, animal, human
- 'to give up an activity': dobj= act, purpose_act
- 'to give up an ideology, a dream..': dobj= movement_of_thought, cognitive_fact
- 'to leave a group, a party, a club..': dobj= institution, human_group
- 'to abandon a sector, a domain…(sport, biology)': dobj= domain
- 'to change one's psychological state' (*abbandonare la calma, la prudenza,* lit.: 'to abandon the calm, the caution'): dobj= psych_property
- 'to drop something' (*abbandonò la divisa a casaccio sulla sedia* 'he dropped the uniform at random on the chair'): in this case the direct object is a "concrete/ inanimate entity" (neither human nor animal). It is worth noting that this specific use of *abbandonare* combines with a particular modifier which cannot occur with the other senses of the verb (e.g. **abbandona la moglie a casaccio* *'he abandoned the wife at random').

ese very granular distinctions – even though motivated by textual evidence - n/should be grouped under the three main senses above, however this additional formation on the various semantic type preferences can be encoded within each oad sense and may be necessary for the selection of the correct translation in MT.

5 The Complexity of Word Sense

Word sense disambiguation (WSD) can be performed, in different contexts, through the use of various information types at different levels of linguistic description: morphosyntactic/syntactic/semantic and even multilingual [10]. Other projects, such as DELIS [14], stressed the interaction between e.g. morphosyntactic patterns and word meanings. The following are some syntactic and semantic indicators which can sometimes help in the identification of a word-sense. The problem is that they are not at all sure tests: they have only a partial validity, and are not completely discriminating. Moreover it is not easy to predict when to apply which test. Therefore human judgement has still an important part in WSD.

- A specific syntactic pattern may allow selecting a particular sense [14], [5], [1]. This is the case of *comprendere* which co-occurs with a that-clause when it means 'to understand' (and not when it means 'to include'), or *aprire* which occurs with a PP introduced by *a* and with "human" head when it has the meaning of 'to be ready, open, well disposed towards someone' (e.g. *Cossiga apre a La Malfa*).
- The domain of use can help to select a specific word meaning (e.g. *perseguire un reato* 'to prosecute a crim' (domain=law)).
- Even the presence of a specific modifier more often than one could think selects a particular sense [14], [5]: e.g. *perseguire penalmente* 'to prosecute at the penal level', does not mean 'to pursue (a goal)'; *comprendere benissimo* 'to understand very well', does not mean 'to include'.
- A specific semantic type of subjects and/or direct objects and/or indirect objects, etc. can help to select a particular meaning of a word (e.g. a human subject always selects the meaning 'to understand' of the verb *comprendere*).
- Different synonyms and/or antonyms select different senses [7].
- Two different senses of a lemma cannot be selected simultaneously in the same context [7] (e.g. **Leo arresta sia il colpevole che il corso degli eventi* *'Leo arrests both the criminal and the events').

It is clear that the availability of large quantities of semantically tagged corpora may help i) to better analyze the impact of different clues to perform WSD in different contexts, and ii) to study the interaction of clues belonging to different levels of linguistic description, in order to improve WSD strategies.

6 What Cannot Be Easily Encoded at the Lexical-Semantic Level of Annotation

In a large number of cases, sense interpretation requires appeal to extra-linguistic knowledge (world knowledge, etc.) which cannot be encoded or, to put it that way, captured at the lexical-semantic level of description. We provide below a few examples.

- When the metaphors are not restricted to a single lemma (e.g. *la chiave del* *verde arriva sul tavolo del governo* lit.: 'the green car arrives on the table of the government').

The sequence means that the "topic" of *auto verde* (the car which does not pollute) will be discussed by the government. However, at the lexical level only *auto*, *verde* and *tavolo* can be marked as figurative uses, whereas the metaphorical sense of the whole sentence will come out only from a violation of the selection restrictions. Indeed, a car (type=vehicle) can arrive (type=move) but not on a table (type=furniture), more probably in a "place/location", while the "topic" of *auto verde* can arrive on the table of the government. This complex sense interpretation of the whole expression cannot be characterized through simple lexical-semantic annotation. It is impossible to imagine the assignment of a label "topic" to *auto* and/or *verde* (everything can indeed be a topic).

- When it is the intention of the author that a sequence is actually ambiguous between two meanings, e.g. *[Titolo]: Nina Vinchi* ***entra in scena*** '[Title]: Nina Vinchi starts/comes on stage' *[Sottotitolo]: A 84 anni la signora del Piccolo affronta per 3 ore i giudici.*

In this corpus context the multi-word expression *entrare in scena* has the double meaning of 'to appear/to start' (the idiomatic sense) and 'to come on stage' (the literal sense). In this case, the interpretation of the sequence is based on knowledge about the domain type (domain=theater) and the context type (indeed ambiguities of this kind are frequently used within titles).

- When some words acquire a specific sense, strictly dependent on the context in which they occur, that cannot be encoded at the lexical-semantic level, e.g. *la donna (Pauline Collins), che ha già visto arrestare il marito dai* ***tedeschi***, ...

Arrestare usually combines with a subject belonging to the military/law domain. Also in this case *tedesco* has to be interpreted as 'German soldier' (and not any kind of German people). However, in the computational lexicon *tedesco* is obviously marked as "people" and cannot be otherwise. Another example is the verb *chiamare* which sometimes means (is synonymous of) 'to telephone'. In most cases the identification of this sense strongly depends on a complex process of context interpretation (even if there are few cases in which the disambiguation is easy, e.g. when the direct object is not a human but a phone number or an inanimate entity, as in *chiamare il (numero) 113/Buckingham Palace/l'ambulanza*). Examples of difficult interpretation are the following: *E io* ***chiamo*** *Craxi per 150 miserabili milioni?* 'And I should call Craxi for 150 miserable millions?', *In gran parte sono bambine dai 6 ai 14 anni.* ***Chiamano*** *per lo più da Milano* .. 'In most cases they are girls from 6 to 14 years. They call mostly from Milan ..'.

- At last, we provide the example of *tagliare* 'to cut'. From the table below it is evident the complexity and variety of nuances implied by the verb, according to the type of direct object which co-occurs with it.

Not all these shifts of meanings can/must be captured through lexical-semantic annotation (sense and semantic tagging). For instance, *tagliare il prato* 'to cut the grass' means 'to eliminate/reduce the grass'; *tagliare le gomme* 'to make a hole (to

hair' (not necessarily to shorten them); *tagliare il mantello* 'to divide the mantle'; *tagliare la legna* 'to cut into pieces the wood'; *tagliare le corolle* 'to detach the corolla of flowers' (to separate the corolla from the flower); on the other hand, *tagliare una fettina* 'to cut a small slice' moves the focus from the whole cut entity to the cut part, etc.

Table 9. Dobj of *tagliare*

Dobj	**Sem.type of Dobj**	**Sem.type of the Verbal Head**
prato	area	cause_change_of_state
gomma	artifact	cause_change_of_state
stoffa	artif.material	cause_constitutive_change
lingua	body_part	cause_constitutive_change
testa	body_part	cause_constitutive_change
capello	body_part	cause_change_of_state
mano	body_part	cause_constitutive_change
mantello	clothing	cause_constitutive_change
legna	material	cause_constitutive_change
corolla	part	cause_constitutive_change
fettina	part	cause_constitutive_change
pezzo	part	cause_constitutive_change
cespuglio	plant	cause_constitutive_change
spino	plant	cause_constitutive_change

Also in this case, this collocational information, which not necessarily implies sense distinction, may be of use – as additional, more subtle information of syntagmatic nature encoded within the existing senses - in a multilingual lexicon for translation purposes.

7 Concluding Remarks

Even a rather small experiment of semantic corpus annotation allows to better understand some of the problematic aspects of lexical-semantic corpus annotation and to have a broad overview of possible types of analysis that can be done on a corpus tagged at the lexical semantic level.

The availability of semantically tagged corpora is considered useful, among others, (i) to evaluate the disambiguating power of the semantic types of the lexical resource used for semantic corpus annotation, (ii) to assess the need of integrating computational lexicons with senses and/or phraseology attested in the corpus, (iii) to identify the inadequacy of certain sense distinctions attested in traditional dictionaries or current computational lexicons which are not applicable to actual usage (see [4]), (iv) to check the real frequency of known senses in different text types or *genres* (some of them may in fact be scarcely attested in a specific corpus type, e.g. in a journalistic corpus), to be kept "under control" - and additional, more granular information (often of collocational nature, such as lexical co-occurrence or selection preferences on arguments) which can/must be encoded within the broader senses e.g. to help translation.

References

1. Atkins, B.T., Kegl, J., Levin, B.: Anatomy of a Verb Entry: from Linguistic Theory to Lexicographic Practice. International Journal of Lexicography 1 (1988) 84–126
2. Alonge, A., Calzolari, N., Vossen, P., Bloksma, L., Castellon, I., Marti, T., Peters, W.: The Linguistic Design of the EuroWordNet Database. Special Issue on EuroWordNet. Computers and the Humanities 32 (1998) 2-3, 91–115
3. Busa, F., Calzolari, N., Lenci, A., Pustejovski, J.: Building a Lexicon: Structuring and Generating Concepts. In: Proceedings of the Computational Semantics Workshop. Tilburg (1999)
4. Calzolari, N., Corazzari, O.: Senseval/Romanseval: the framework for Italian. Computers and the Humanities 34 (2000) 1-2, 61–78
5. Calzolari, N., Corazzari, O., Monachini, M., Roventini, A.: Speech Act and Perception Verbs: Generalizations and Contrastive Aspects. In: EURALEX-96 Proceedings. Goteborg (1996) 73–83
6. Corazzari, O.: Phraseological Units. ILC, Pisa (1992)
7. Cruse, D.A.: Lexical Semantics. Cambridge University Press, Cambridge (1986)
8. Fass, D.: A Method for Discriminating Metonymy and Metaphor by Computer. Computational Linguistics 17 (1991) 1, 49–90.
9. Fellbaum, C. (ed.): Wordnet, An Electronic Lexical Database. MIT Press, Cambridge, (1998)
10. Gale, A. W., Church, K.W., Yarowsky, D.: A Method for Disambiguating Word Senses in a Large Corpus. Computers and the Humanities 26 (1992) 415–439.
11. Kilgarriff, A.: Dictionary word sense distinctions: An enquiry into their nature. Computers and the Humanities 26 (1993) 365–387
12. Kokkinakis, D., Kokkinakis, S. J.: Sense-Tagging at the Cycle-Level Using GLDB. Göteborg University (1999)
13. Lenci, A., Busa, F., Ruimy, N., Gola, E., Monachini, M., Calzolari, N., Zampolli, A.: Linguistic Specifications. SIMPLE Deliverable D2.1. ILC and University of Pisa (1999)
14. Monachini, M., Roventini, A., Alonge, A., Calzolari, N., Corazzari, O.: Linguistic Analysis of Italian Perception and Speech Act Verbs. DELIS Working Paper. ILC, Pisa (1994)
15. Montemagni, S., Barsotti, F., Battista, M., Calzolari, N., Corazzari, O., Zampolli, A., Fanciulli, F., Massetani, M., Raffaelli, R., Basili, R., Pazienza, M.T., Saracino, D., Zanzotto, F., Mana, N., Pianesi, F., Delmonte, R.: The Italian Syntactic-Semantic Treebank: Architecture, Annotation, Tools and Evaluation. In: Proceedings of the COLING Workshop on "Linguistically Interpreted Corpora (LINC-2000)". Luxembourg (2000) 18–27
16. PAROLE: Preparatory Action for Linguistic Resources Organization for Language Engineering. LE-4017, Pisa (1996).
17. Rodriguez, H., Climent, S., Vossen, P., Loksma, L., Peters, W., Alonge, A., Bertagna, F., Roventini, A. : The Top-Down Strategy for building EuroWordNet: Vocabulary Coverage, Base Concepts and Top Ontology. Special Issue on EuroWordNet. Computers and the Humanities 32 (1998) 2-3.

Meaning Sort — Three Examples: Dictionary Construction, Tagged Corpus Construction, and Information Presentation System —

Masaki Murata, Kyoko Kanzaki, Kiyotaka Uchimoto, Qing Ma, and Hitoshi Isahara

Communications Research Laboratory, MPT,
2-2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto, 619-0289, Japan,
{murata,kanzaki,uchimoto,qma,isahara}@crl.go.jp,
http://www-karc.crl.go.jp/ips/murata

Abstract. It is often useful to sort words into an order that reflects relations among their meanings as obtained by using a thesaurus. In this paper, we introduce a method of arranging words semantically by using several types of 'is-a' thesauri and a multi-dimensional thesaurus. We also describe three major applications where a meaning sort is useful and show the effectiveness of a meaning sort. Since there is no doubt that a word list in meaning-order is easier to use than a word list in some random order, a meaning sort, which can easily produce a word list in meaning-order, must be useful and effective.

1 Using Msort

Arranging words in an order that is based on their meanings is called a meaning sort (Msort). The Msort is a method of arranging words by their meanings rather than alphabetically. The method used to list the meanings is described in the next section.

For example, suppose we obtain the following data in a research project:[1]

an event

a temple, a formal style, an alma mater, to take up one's post, the Imperial Household, a campus, Japan, the Soviet Union, the whole country, an agricultural village, a prefecture, a school, a festival, the head of a school, an established custom, a government official, a celebration, a Royal family

This is a list of noun phrases (NPs), each followed by the word *gyoji* (an event) in the form NP X *no gyoji* (an event of NP X) in Japanese. To find the most useful way to examine the list, we first arrange the NPs alphabetically: