Encoding Terms from a Scientific Domain in a Terminological Database: Methodology and Criteria

Rita Marinelli* Melissa Tiberi Remo Bindi***

*Istituto di Linguistica Computazionale, C.N.R.
Area della Ricerca Via Moruzzi 1, 56124 Pisa Italy
e-mail: Rita.Marinelli@ilc.cnr.it, Remo.Bindi@ilc.cnr.it

**Biblioteca Nazionale Centrale di Firenze. Collab. est.
Piazza dei Cavalleggeri 1, 50122 Firenze Italy
tiberim77@yahoo.it

Abstract

This paper reports on the main phases of a research which aims at enhancing a maritime terminological database by means of a set of terms belonging to meteorology. The structure of the terminological database, according to EuroWordNet/ItalWordNet model is described; the criteria used to build corpora of specialized texts are explained as well as the use of the corpora as source for term selection and extraction. The contribution of the semantic databases is taken into account: on the one hand, the most recent version of the Princeton WordNet has been exploited as reference for comparing and evaluating synsets; on the other hand, the Italian WordNet has been employed as source for exporting synsets to be coded in the terminological resource.
The set of semantic relations useful to codify new terms belonging to the discipline of meteorology is examined, revising the semantic relations provided by the IWN model, introducing new relations which are more suitably tailored to specific requirements either scientific or pragmatic. The need for a particular relation is highlighted to represent the mental association which is made when a term intuitively recalls another term, but they are neither synonyms nor connected by means of a hyperonymy/hyponymy relation.

1. Introduction

A lexical semantic database of maritime terminology was built by exploiting the computational tools of ItalWordNet (IWN) (Marinelli and Roventini, 2006) and its lexical semantic model EuroWordNet (Vossen, 1999).
Our choice to perform this type of research was determined by the need of a terminological resource that could be a support for the definition and translation (Italian – English) of the terms belonging to this domain.
The lack of researches in this field for the Italian language, together with the frequency of maritime terms in spoken and written texts and, in general, in everyday life has determined the need to manage the growing new technical maritime terminology
In the frame of the development of the ports in the last thirty years, the "port" is a focal point in the network which involves not only commercial, but also industrial, economic, financial, and logistic activities. The introduction of industrial techniques and logistic procedures, originated and developed in Anglo-Saxon countries, in the field of transport have led to a kind of "monopoly" of the English language in this sector of economy. Therefore there is a growing need to become familiar with the terminology which, up to now, was reserved to a limited number of experts, and to have reliable tools available to manage the ever-increasing new English technical terminology, in an attempt to avoid the far too easy attitude to simply introduce new English terms as neologisms in the national languages.
Our aim is to provide a useful instrument for work, didactic activities and in general whenever a reference to terms of this specific domain is needed, concerning a proper technical term use and information and an abreast and unambiguous translation.
This paper reports on the main phases of a project for the enhancement of the maritime database, recently undertaken and still in progress, with a set of terms belonging to the scientific domain of Meteorology.

2. Meteorology Relevance

The maritime domain structure is described considering the complexity of this domain, defining a core set of terms representing the many knowledge fields that are included in the maritime domain.
Each concept of the database is connected to both the Top Ontology of IWN and to the specific ontology, namely a core set of concepts, belonging to maritime terminology, which are the basis of our domain modelling (Marinelli et al., 2006). They represent the two main sub-domains specified in maritime terminology: the technical/nautical (nautics) and the maritime transport (transport) domains and the many knowledge fields that are included in the maritime domain (Marinelli and Spadoni, 2006). Among these knowledge fields, Meteorology has a particular relevance.
Marine meteorology is a discipline which uses meteorological knowledge to understand and foresee the phenomena related to the sea which is taken to be all of the salt water covering the Earth's surface.
Weather reports and forecasts, especially referring to wind and wave strength and direction, are necessary for

shipping and navigation operations. Furthermore, considering the maritime transport field, weather forecasts' accuracy makes it possible to plan the most "economical" and safest routes, in order to maintain the scheduled "transit time" between ports, to evaluate the necessity to shelter when rough sea conditions can damage the cargo, to program cargo operations minimizing idle time and consequent costs, due to bad weather conditions.

The weather component plays a significant role in maritime contracts as, e.g., the Expected Time of Arrival (ETA) for a ship into a port is always computed "Weather Permitting" (WP) and the calculation of the "lay time", the maximum time that the maritime contract assigns to perform the cargo operations, is always based on a fixed number of "Weather Working Days" (WWD) (Marinelli and Spadoni, 2007).

Very often, in the shipping field, the commercial operators too will deal with meteorological terms to justify the reason for a vessel delay or to describe the "time lost" for meteorological causes in the "statement of facts", which is the report issued by the shipping agent, detailing the operations performed in the port and respective times.

Meteorology is included as a fundamental discipline in the curriculum of schools and various types of didactic activities connected to the maritime domain (nautical Institutes, professional training, University courses, etc.).

The maritime terminology database contains above 3,500 lemmas, corresponding to 2,500 synsets. A first subset of concepts, over 300, strictly related to maritime navigation and pertaining to sea status, weather forecasts, atmospheric phenomena, etc., is already present in the database.

We thought it was worthwhile to increase this part of the terminology by adding a set of new concepts: a set of nearly 1,000 terms have recently been chosen to be analyzed. The terms were selected using different approaches, that is starting from corpora and from databases, using both resources as source and reference.

3. Corpus Approach

An initial corpus of meteorology was constructed which contains texts from manuals and various specialized sources (web sites, weather forecast reports, specialized newspaper articles, etc.), some of which (e.g.: *Appunti di meteorologia marina* (Notes on marine meteorology)) were suggested or provided by a CoMMA–Med Laboratory of the Institute of Biometeorology (C.N.R.) (Brugnoni et al., 2006).

A second corpus was built starting from the PAROLE corpus.

The Italian PAROLE Corpus (Marinelli et al., 2003) consists of 20 million word tokens, including texts collected until 1996. One of the main goals of the LE PAROLE project was to ensure the creation of a comparable set of large Written Language Resources (WLRs) for all the European languages.

The set of newspapers was taken into account. The corpus managing tool provides instruments for various type of study and research, and, among others, to create a subset of the PAROLE corpus. In fact, it is possible to select texts on the basis of the Topic they deal with, represented in the heading of each PAROLE text.

The Topic "meteorologia" (meteorology) was focused on when selecting texts from the newspapers "La Stampa" and "Corriere della sera"; the Topic "clima" (climate) was considered to select texts from the newspaper "Repubblica".

The first corpus, consists of nearly 140,000 occurrences and contains texts dealing with meteorology from a more specialized perspective; the second corpus is composed of over 70,000 word forms including less technical texts, which report various events such as water-floods, crashes, etc., and situations such as healthcare, preventive treatments, etc., related to meteorology. None of these texts has a very high degree of specialization, but together, they represent how much the concept of "meteorology" is involved in our everyday life..

It was possible to produce a list of frequencies from each corpus. The most frequent terms were selected for analysis, then to be added to the terminological database. The corpus approach is useful to verify the term usage, to confirm the term meanings and to assess and refine, if needed, the term definitions.

The co-occurrences were also examined. In fact, it is possible to examine a word "α" (e.g.: *pioggia* (rain)) and to give the dimension of the context: e.g.: 1, which is a context composed of the word before or the word after "α". The result obtained from the corpus managing tool will be the list of all the words close to "α", ordered on the basis of: i) mutual information index (Church and Hanks, 1990); ii) the frequency of the words in the corpus; iii) the number of co-occurrences of the words with "α" in the corpus.

The co-occurrence list is a useful reference for codifying the most frequent words (adjectives, nouns, etc.) that occur together with the word considered: e.g.:

pioggia (rain) → *incessante* (endless), *battente* (pouring), *intermittente* (intermittent).

C:\DBT2000\DBTDATA\DBTPR2.MET PIOGGIA

1)	3	3	9.175	1.000	mista
2)	2	2	9.175	1.000	incessante
3)	7	8	8.982	1.000	battente
4)	2	3	8.590	1.000	intermittente
5)	2	5	7.853	1.000	violenta
6)	3	30	5.853	1.000	continua
7)	5	132	4.452	1.000	dalla
8)	3	88	4.300	1.000	qualche
9)	2	62	4.221	1.000	lungo
10)	4	181	3.675	1.000	alla
11)	19	1192	3.204	1.000	la
12)	30	2157	3.008	1.000	e
13)	2	146	2.985	1.000	ancora

Table 1: *Pioggia* (rain) co-occurrences

4. Database Approach

Other terms were selected using a database approach. In fact, we used both the WordNet database and the IWN database as source and reference.

4.1 Synsets from WN 3.0

The most recent version of the WordNet database (WN 3.0) was taken into account: the synset "meteorology", was examined, this time starting from English.
The various semantic relations (vertical and horizontal) exploited were worked out and the synsets involved were compared and evaluated; the "sister" terms and the "derivationally related forms" (in WN terms) were studied and considered as reference for the English translation of the Italian terms and guideline to align and correlate the presence of sets of terms in the two databases.
The synsets "meteorology" and the direct hyponyms are shown hereafter:

(n) **meteorology**, weather forecasting (predicting what the weather will be)
(n) **meteorology** (the earth science dealing with phenomena of the atmosphere (especially weather))

direct hyponym / *full hyponym:*
(n) aerology (meteorology of the total extent of the atmosphere; especially the upper layers)
(n) climatology (meteorology of climates and their phenomena)
(n) nephrology (the branch of meteorology that studies clouds and cloud formation)

The forms derivationally related from "meteorology":

derivationally related form:
◦ **(adj) meteorologic** [Related to: meteorology] (of or pertaining to atmospheric phenomena, especially weather and weather conditions) *"meteorological factors"; "meteorological chart"; "meteoric (or meteorological) phenomena"*
◦ **(adj) meteorological** [Related to: meteorology] (of or pertaining to atmospheric phenomena, especially weather and weather conditions) *"meteorological factors"; "meteorological chart"; "meteoric (or meteorological) phenomena"*
◦ **(n) meteorologist** [Related to: meteorology] (a specialist who studies processes in the earth's atmosphere that cause weather conditions).

4.2 Synsets from IWN

The generic database IWN was examined and a set of concepts belonging to Meteorology were identified and focused on. These synsets were exported and then imported into the terminological database, as "xml" files, exploiting one of the functionalities of the database managing tool.
The terms belonging to Meteorology were classified and codified by means of the semantic relations used to codify all the other terms of the maritime domain, that is, the semantic relations available in the EWN/IWN model:

► **Internal relations**: which encode information in the form of lexical-semantic relations between pairs of synsets. Since WNs were built by working on taxonomies, hierarchical links are instantiated in a rather consistent way - as in IWN - and expressed by two relations (vertical relations) '*has_hyperonym*' and '*has_hyponym*' that allow a crosschecking of data.
The linguistic model is very rich and contains many other lexical-semantic relations (horizontal relations) such as *part_of, cause, purpose, sub_event, belong_to_class,* etc. The use of vertical (*hyperonymy/hyponymy*) relations leads up the definition of the most basic level of categorization namely "the most inclusive (abstract) level at which the categories can mirror the structure of attributes perceived in the world" (Rosch, 1988). The use of the horizontal dimension for categorization implies the improvement of the distinctiveness and flexibility of categories.
► **Equivalence relations:** link the Italian synsets with the closest concepts (synonyms, near synonyms, hyperonyms, etc.) of the Inter Lingual Index (ILI), an unstructured version of WN 1.5.
When possible an *eq_synonym* or *eq_near_synonym* relation is used, otherwise an *eq_has_hyperonym* relation is coded, e.g.:

vento *eq_synonym* wind
vento in poppa *eq_has_hyperonym* wind

by these links to the ILI, the terms are also connected to the Top Ontology (TO).
► **Plug-in relations:** allow the linking of a synset of the specialized wordnet to the generic one (IWN), connecting a terminological sub-hierarchy (represented by its root node) to a node of the generic wordnet.
By means of the *plug in* relations the tool we are using to manage the terminological database and the specific ontology also allows an "integrated" consultation of the database; it shows that if a synset is found in both databases (and plugged), the synset belonging to the specific domain partially "obscures" the generic one: downward and horizontal relations (*part_of* relations, *role* relations, *causes* relations, *derivation*, etc.) are taken from the terminological wordnet, while upward (*hyperonymy*) relations are taken from the generic one (see Fig. 1 and Fig. 2 hereafter).

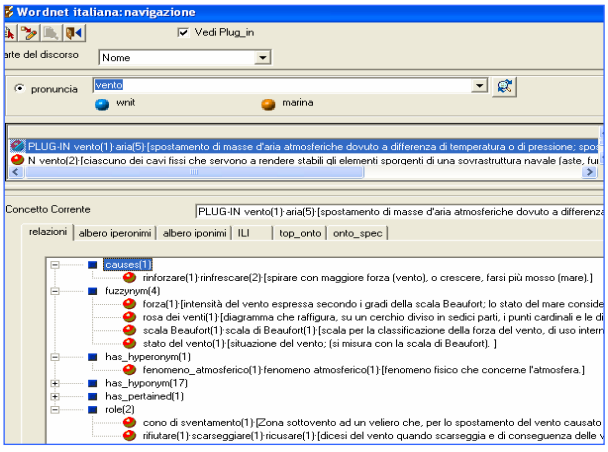

Figure 1: *Vento* (wind) downward relations

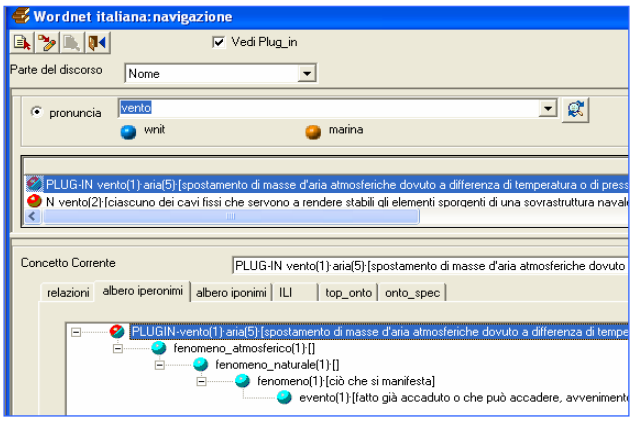

Figure 2: *Vento* (wind) upward relations

4.3 New Relations

A large number of terms, in the maritime domain and, in particular, in the Meteorology knowledge field, have to be coded as "events". We can refer to the definition of "event" given in WN 3.0, namely: "something that happens at a given place and time", and in IWN: "fatto già accaduto o che può accadere, avvenimento di una certa importanza" (something that has already occurred or that may happen, a fact of a certain importance). As Hobbs and Pustejovsky (2005) say, speaking about automatic recognition of temporal and event expressions in natural language text, there has recently been a renewed interest in temporal and event-based reasoning in language and text and "in recognition of events and their 'temporal anchorings'". The temporal aspects of the properties of the entities considered to be encoded in the semantic database are particularly relevant from the perspective of events representation. They are necessary to understand the event perceived and to connect it with the other concepts belonging to our knowledge of the world.

To codify the space dimension of an event a set of relations is provided by the IWN model. These relations are available for encoding space/location relationships, either if the concept of "space" is involved as the location in which something happens, or when the space is conceived as a path along which a movement occurs. It can be further specified as movement "from" a place (source direction) and/or movement "to" a place (target direction). But this set of relations can be exploited taking into account a number of constrains (e.g. neither of these relations can be employed with proper names) that limit or prevent us from using them in many cases.

Therefore, we thought it useful to integrate the set of the semantic relations provided by the IWN model introducing new relations more suitably tailored to specific needs (*event_location, event_time*):

→ Uragano (hurricane) Ivan *event_location* Alabama
→ Uragano (hurricane) Ivan *event_time* September 2004

From this point of view, a study is in progress to create specification relations suitable for representing the space-temporal dimension in a more exhaustive manner; they will appear together with the other semantic relations (see Figure 3).

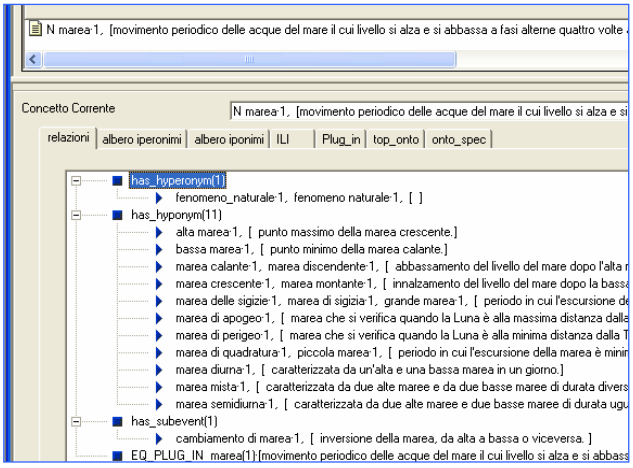

Figure 3: *Marea* (tide) semantic relations

5. Associative Relations

The relational structure of the EWN/IWN model lends itself well to encode and embed the terms in the semantic network; it gives them an ontological categorization using the concepts belonging to the upper ontology (TO) of IWN and the concepts belonging to the specific ontology, as it is shown in the Figures below.

Figure 4: Marine meteorology - *Stato del vento* (Wind status)

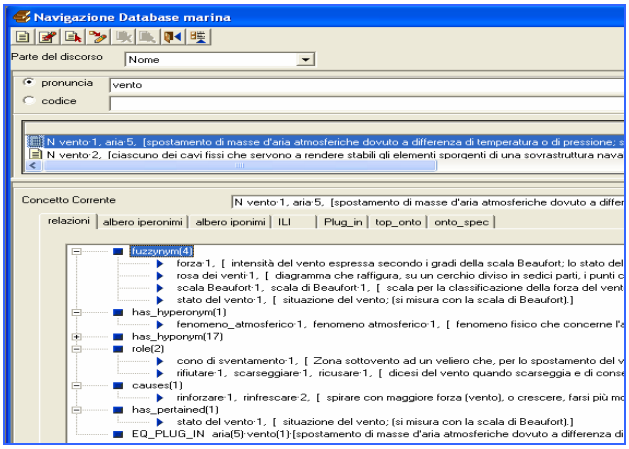

Figure 5: *Vento* (wind) Internal and a Plug_in relation

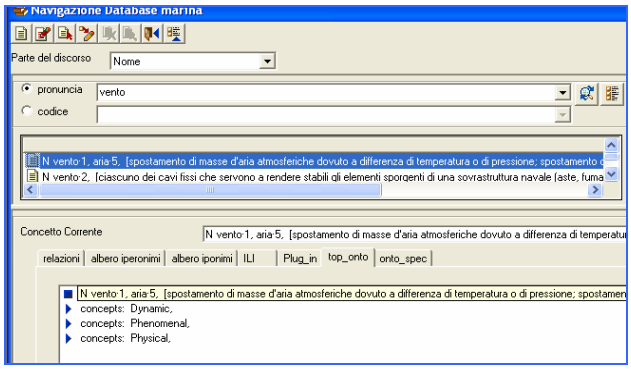

Figure 6: *Vento* (wind) Top Ontology

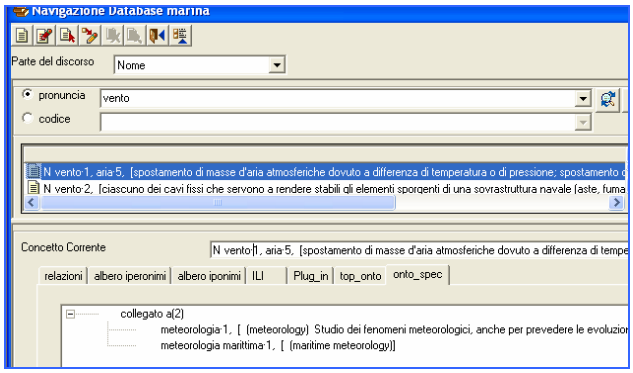

Figure 7: *Vento* (wind) Specific ontology

While carrying out the semantic coding and building the domain ontology following our model, the lack of a particular kind of relation became evident, such that we can call "associative", suitable for represent important connections that are intuitively obvious.

It is different from all those vertical and horizontal relations already present in IWN and it is useful to codify the relationship between "intuitively related pairs" representing that the first concept brings to mind the second (Boyd-Graber et al., 2006).

We could say that it is similar to associative relations used in Thesauri[1]; they are useful and necessary to retrieve information and to represent relationships of thematic pertinence such as, given a first term, to suggest a second one that is strongly associated with it and that is neither a synonym nor hierarchically related.

The associative relation is defined as "pertaining to terms pairs which are not members of an equivalence set, nor can be organized as a hierarchy …. yet, are mentally associated" (UNI ISO 2788).

This kind of relation is useful to highlight very strict relationships, when meaning overlapping or interchange must be shown.

[1] Thesaurus: The vocabulary of a controlled indexing language, (i.e. a controlled set of terms selected from natural language, and used to represent, in summary form, the subjects of documents), formally organized so that the a priori relationships between concepts are made explicit.

The associative relation usually connects terms belonging to different logic categories e.g.: *"meteorologia"* (meteorology) – *"previsioni del tempo"* (weather forecasts), which belong respectively to *"discipline"* (disciplines) and *"azioni ed eventi"* (actions and events), and permits linking between different hierarchical structures.

Moreover, in ISO 2788, there are also cases in which two terms, e.g.: *"navi"* (ships) and *"barche"* (boats), have the same category *"oggetti"* (objects), are siblings, with similar overlapping meanings and with the same hyperonym *"veicoli"* (vehicles).

Since there is no rigid rule for its application, there is a chance that it might be used in an inappropriate and subjective way.

The IWN model provides a particular relation, the *fuzzynym* relation, which has been employed until now: it is a "wild card" relation used when no other relation seems to fit, to indicate a general, not well defined connection between two concepts. Instead, relational terms have to be codified by means of a suitable associative relation, namely the "*associated_with*" relation, when an association relationship is established to indicate that a term has similarities with other concepts, that other information of interest may be classified under a different, but related, set of terms, e.g.:

Vento (wind) *associated_with* rosa dei venti (wind rose)
Vento (wind) *associated_with* scala Beaufort (Beaufort scale)
Rosa dei venti (wind rose) *associated_with* punto cardinale (cardinal compass point)
Nube (cloud) *associated_with* precipitazione (precipitation)
Meteorologia (meteorology) *associated_with* bollettino meteorologico (weather forecasts)
Meteorologia (meteorology) *associated_with* meteorologo (meteorologist), as it is shown in the Figure 8 hereafter:

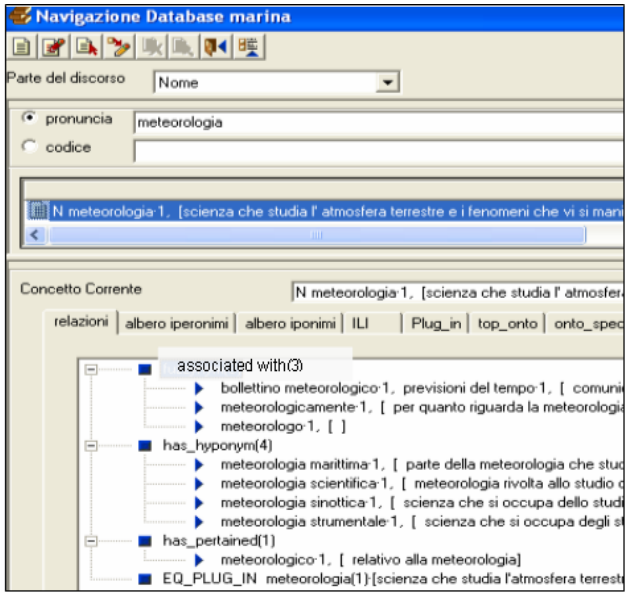

Figure 8: *Meteorologia* and the *associated_with* relation

In this way, we look at the aim that is guiding more or less rigidly, and more or less specifically, "domain" choices.
Such a relation could be used productively in a scientific domain, because it attests and grants the connections with the "macro-domain" of maritime terminology that has to be increased. After structuring the terms following the "classic", properly logical, classification scheme, a relation may be useful which allows to the meteorological terminology, in this case, to "dialogue" with so huge and multidisciplinary a domain as the maritime domain. Starting from the definition of a term belonging to meteorology (if possible, from the point of view of maritime meteorology) some semantic connections can be created with all the rest of the terminological database.
Pragmatically and conceptually, these semantic connections allow the term to maintain its own degree of specialization and, at the same time, to be integrated with the semantic and computational structure in which it is embedded.
An association relationship is established to indicate that a term has similarities with other concepts, that other information of interest may be classified under a different, but related, set of terms.
In this way each term is inserted into a relational network that helps clarify its semantic content, respecting the high specificity of this scientific domain.

6. Conclusions

The main phases of our research have been described consisting in the enhancement of the maritime terminological database by means of a set of terms belonging to meteorology. The criteria employed to build corpora of specialized texts were detailed; these corpora were used as the source for term selection and information extraction. We described the use of semantic databases as the source (IWN) for exporting synsets to be coded in the terminological resource and as reference (WN 3.0) for emphasizing the similarities between or among synsets, though not losing sight of the differences.
It was necessary to examine and weigh the set of semantic relations that are useful for codifying the new terms belonging to the discipline of meteorology and the need to introduce, in addition to the semantic relations provided by the IWN model, new relations which are designed to satisfy specific scientific or pragmatic requirements..
In fact, firstly, the coding of some terms by means of relations representing the space – temporal perspective was introduced; then we proposed the associative relationship as another relation to be used productively in a scientific domain, because it indicates and gives evidence of the connections with the huge and multidisciplinary "macro-domain" that has to be increased.

7. References

Boyd-Graber J., Fellbaum C., Osherson D., Schapire R. (2006). Adding Dense, weighted Connections to WORDNET. In: Sojka P., Choi K.-S., Fellbaum Ch., Vossen P. (Eds.): *Proceedings of the Third International WordNet Conference, Seogwipo, Korea, 2006* Brno, Masaryk University, pp. 29-35.

Brugnoni G., Doronzo B., De Sario G., Petralli S., Scartazza A.., Taddei S., Gozzini B., Pellegrino L., Vaccari F. P. (2006). Appunti di meteorologia marina. Firenze, Edizioni Regione Toscana.

Church K., Hanks P. (1990). Word association norms, mutual information and lexicography. In: *Computational Linguistics*, Volume 16 , Issue 1, (March 1990). MIT Press Cambridge, MA, USA, pp. 22-29.

Hobbs J., Pustejovsky J. (2005). Annotating and reasoning about time and events. In: Mani. I., Pustejovsky J., Gaizauskas R. (Eds.), *The Language of Time*, Oxford University Press, pp. 301-316.

ISO 2788 – 1986. 2005. *Documentation - Guidelines for the establishment and development of monolingual thesauri.* International Organization for Standardization, printed in Switzerland.

Marinelli R., Biagini L., Bindi R., Goggi S., Monachini M., Orsolini P., Picchi E., Rossi S., Calzolari N., Zampolli A. (2003). The Italian *PAROLE* corpus: an overview. In: A. Zampolli, N. Calzolari, L. Cignoni, (eds.) Computational Linguistics in Pisa, *Linguistica Computazionale*, Special Issue, XVI-XVII, Pisa-Roma, IEPI, pp. 401-421.

Marinelli R., Spadoni G. (2006). Some considerations in structuring a terminological knowledge base**.** In: Sojka P., Choi K.-S., Fellbaum Ch., Vossen P. (eds.): *Proceedings of the Third International WordNet Conference, Seogwipo, Korea, 2006* Brno, Masaryk University, pp. 217-224.

Marinelli R., Roventini A., Spadoni G. (2006). Using core ontology for domain lexicon structuring. In: *Proceedings of LREC 2006: 5th International Conference on Language Resources and Evaluation. Genoa, Italy.* Paris, ELRA, pp. 207-212.

Marinelli R., Roventini A. (2006). The Italian Maritime Lexicon and the ItalWordNet Semantic Database. In Eloína Miyares Bermúdez and Leonel Ruiz Miyares (Eds.), *Linguistics in the Twenty First Century;* Cambridge Scholar Press, pp. 173-182.

Marinelli R., Spadoni G. (2007). Modeling a Maritime Domain Ontology. In: *Proceedings of the Tenth International Symposium on Social Communication*. Centre for Applied Linguistics. Santiago de Cuba, January 22-26, 2007.

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

Rosch E. Principles of Categorization. (1978). In *Readings in Cognitive Science, a Perspective from Psychology and Artificial Intelligence*, A. Collins & E. E. Smith, Morgan Kaufmann Publishers, San Mateo, California, 1988, pp. 312-322.

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