Lexical Semantic Resources in a Terminological Network

Rita Marinelli* Adriana Roventini* Giovanni Spadoni Sebastiana Cucurullo***

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

**S. Spadoni s.r.l. Shipping Agency
Via delle Cateratte 90, 57122 Livorno, Italy
e-mail:g.spadoni@saurospadoni.it

Abstract

A research has been carried on and is still in progress aimed at the construction of three specialized lexicons organized as databases of relational type. The three databases contain terms belonging to the specialized knowledge fields of maritime terminology (technical-nautical and maritime transport domain), taxation law, and labour law with union labour rules, respectively.
The EuroWordNet/ItalWordNet model was firstly used to structure the terminological database of maritime domain. The methodology experimented for its construction was applied to construct the next databases. It consists in i) the management of corpora of specialized languages and ii) the use of generic databases to identify and extract a set of candidate terms to be codified in the terminological databases. The three specialized resources are described highlighting the various kinds of lexical semantic relations linking each term to the others within the single terminological database and to the generic resources WordNet and ItalWordNet. The construction of these specialized lexicons was carried on in the framework of different projects; but they can be seen as a first nucleus of an organized network of generic and specialized lexicons with the purpose of making the meaning of each term clearer from a cognitive point of view.

1. Introduction

A research has been carried on and is still in progress at the ILC aimed at the construction of three specialized lexicons organized as databases of a relational type.
The databases contain terms belonging to different knowledge fields: the first (Marinelli and Roventini, 2006) is a database of maritime terminology (technical-nautical and maritime transport domain) (MDB); in the other two databases terms belonging to the knowledge field of taxation law (TDB) and to the domain of labour law and union labour rules (LDB) are codified.
The database of maritime terminology (MDB) was built first, on the basis of the EuroWorNet (EWN) (Vossen, 1999) and ItalWordNet (IWN) model (Roventini et al., 2003), using lexical semantic relations to codify terms, within the framework of the Princeton WordNet (WN) philosophy (Miller et al., 1990); it includes about 4000 lemmas. The other two databases (1600 and 1500 lemmas respectively) were structured following similar criteria, in keeping with the methods already successfully experimented to create the MDB and also to enhance it with a subset of terms belonging to the scientific domain of Meteorology (Marinelli, 2008).
These methods which have been firstly used to build the MDB, consist of an organized set of research phases constituting a true methodology to create terminological databases. These phases are herewith described: i) the corpus approach; ii) the generic database approach. Then the whole terminological network is depicted, highlighting the different kinds of relations linking the terminological and the generic resources.

2. The corpus approach

Terminological corpora play a key role in ensuring that specialized dictionaries and terminological databases mirror the language used in particular domains of knowledge and work environments. Corpus based methods are fundamental for the extraction and structuring of terminological sets of data.

2.1 The terminological corpora

The databases of this project were built starting from a set of terms with high frequency in the three corpora of terminology, each created in the frame of the project.
An initial corpus of maritime terminology was constructed which contains texts from manuals, glossaries and various specialized sources (web sites, maritime law documents, specialized newspaper articles, etc.); it consists of nearly 240,000 occurrences. Two other corpora containing texts dealing with fiscal and syndicate-labour subjects (360,000 and 160,000 occurrences) were built according to the same methodology.
A textual database managing system (Picchi, 2003) was exploited which provides instruments for various types of study and research and, among others, to create both alphabetical and decreasing lists of frequencies and concordances.
It was possible to produce a list of decreasing frequencies from each corpus.

The most frequent terms were selected for analysis, to be added to the related terminological database. The corpus approach is useful to verify term usage and to assess and refine, if needed, the term definitions, since "the meaning is fundamentally guided by context" (Evans, 2006).
The co-occurrences were also examined as a useful reference for codifying the most frequent words (adjectives, nouns, etc.) that occur together with the word considered, and as a support to single out compounds or terminological sub-hierarchies constituted by the term itself followed by an adjective or a prepositional phrase, particularly frequent in specialized languages.
One of the features of the textual database managing system consists in identifying collocates of nouns in the terminological corpora also on the basis of their mutual information index (Church and Hanks, 1989).
Considering for instance, the concept *carico* (cargo), the following compounds or multiwords were identified and then encoded: *carico completo* (full cargo), *carico di merci varie* (general cargo), *carico in coperta* (deck cargo); taking into account the concept *imposta* (duty), *imposta sul reddito* (income tax), *imposta di successione* (death duty), *imposta sui beni mobili* (personal property tax) were encoded as its hyponyms.

3. The generic lexicon approach

A percentage of the most frequent terms in each corpus was found in the IWN database and checked analysing the semantic correspondence of senses.
These were exported in xml files, exploiting one of the tool functionalities, and imported in the terminological resources, constituting the first conceptual nucleus of the database to be populated. Other sets of terms, with various level of specificity, were taken from many kinds of qualified and well-known sources: web sites, miscellaneous documents, private archives, manuals; a top down or a bottom up methodology was followed to increase the lexical coverage of each database adding hyperonyms and hyponyms.
All of these operations were supervised by the domain expert.

4. The relational model

The model used to structure the MDB and the other two terminological lexicons is based on the concept of "synset" (set of synonyms), e.g.: {*bollo auto1* (road tax), *tassa automobilistica1*(automobile tax)} and on the vision of the synset as fully defined by the lexical semantic relations which connect it to other synsets and are a kind of representation of the user's mental lexicon (Miller, 2003). The lexical semantic relations can be described as:
1) relations which link the terms with other terms "within" the terminological database or "internal relations";
2) relations which connect the terms to the synsets of the generic databases WordNet (WN) and ItalWordNet (IWN).

4.1 Internal relations

While synonymous meanings are joined in a "synset", language-internal relations hold between pairs of synsets. WNs are based on the management of conceptual structures which are hierarchically organized by means of "vertical" relations -"has_hyperonym/ has_hyponym"- ensuring coherence and consistency.
In the EWN/IWN linguistic model a variety of lexical-semantic relations (horizontal relations) such as "part_of", "cause", "purpose", "sub_event", "belong_to_class", etc. are also used to represent the organization of lexical knowledge. 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 basic level is "the most natural, preferred level at which to conceptually carve up the world" (Murphy, 2004). Horizontal relations are exploited to indicate semantic relatedness between concepts that are neither synonyms nor hierarchically dependent; in Rosch's (1988) words, the use of the horizontal dimension for categorization implies the improvement of the distinctiveness and flexibility of categories:

bollo auto1 (road tax) has _hyperonym *tassa1* (tax) (*corrispettivo che un privato deve ad un ente pubblico per la fornitura di un bene o di un servizio*)

somma1 (amount) fuzzynym *pagare* (to pay) (*dare una somma di denaro dovuta*)

lavoro1 (work) involved_location *fabbrica1* (plants, works) (*stabilimento in cui si svolge una produzione*)

4.2 Term connection to the generic lexicons

The database design principles provide, besides the lexical semantic relations which connect each term to other terms within the specialized lexicon, also semantic relations that link synsets of the terminological database to the generic lexicons WordNet and ItalWordNet.

4.2.1 Equivalence relations

According to the conceptual architecture of the model, equivalence relations link the terminological synsets to their closest equivalent concepts of the Princeton WN through the Inter Lingual Index (ILI)[1]. This happens in terms of
a) synonymy (or near synonymy) relationship:

polizza di carico eq_has_synonym *bill_of_lading* (a receipt from the carrier for the goods being shipped)
agenzia di collocamento eq_near_synonym *employment agency* (company that finds work for applicants)

b) inclusion of the synsets in a taxonomic chain:

[1] The ILI is an unstructured fund of synsets (mainly taken from WorNet1.5), the so-called ILI-records (Vossen, 1998).

lettera di assunzione eq_has_hyperonym *document*

c) "part", "role", "means", etc. relations, helping define the semantic field of each term more precisely:

lavoratore (worker) eq_role *work, do_work, (be employed)*
imbarcarsi (to go on board) eq_antonym *to_go_ashore.*

By these links to the ILI, the terms are also connected to the Top Ontology (TO), that is a set of concepts with high level of abstraction, hierarchically organized and language independent:
navigazione (navigation) → Agentive, Dynamic, Purpose,
tassa (tax) → Dynamic, Possession, Quantity,
lavoro (job) → Social, Static.
When the closest concept of the term was not found in the English WN, the term was linked to its hyperonym and the English synonym of the term was recorded in a list by which the ILI had to be updated and enlarged.
An example is shown hereafter:

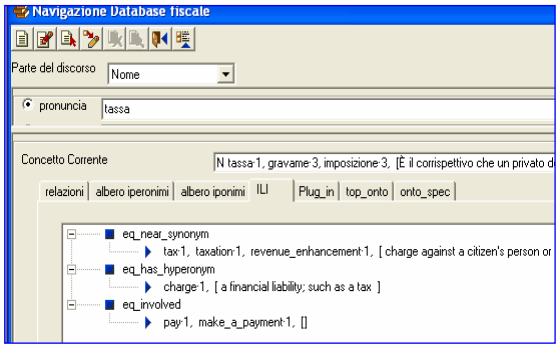

Figure 1. *Tassa* (tax) equivalence relations

All these kinds of semantic relations are managed by a tool which allows visualization and updating of each database singularly. The database management system has been realized in Visual Basic and the data are filed in a SQL database.

4.2.2. Plug-in relations

Plug-in relations connect the term of the specialized lexicons to synsets of the generic IWN, in terms of hyper/hyponymy or equivalence relationship:

imposta (duty) has-hyperonym plugin *somma* (amount of money)
nave (ship) eq-plug-in *nave* (ship).

The computer interface which allows browsing and updating of each database singularly also admits an "integrated" consultation of the database. In fact, when a plug-in relation is codified, the term being sought, e.g.: "*merce*" (goods), is linked to a synset of the generic IWN which is a kind of hookup point from which all the downward and the horizontal relations of the terminological database flesh out while all the upward (vertical/hierarchical) relations of the generic IWN are shown:

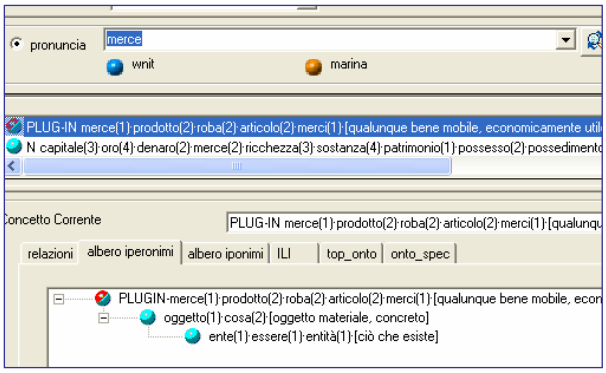

Figure 2. Upward relations

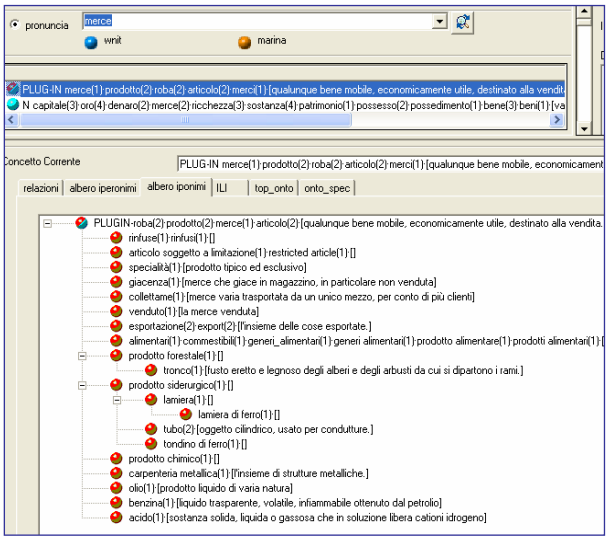

Figure 3. Downward relations

By means of the plug-in relations the link to the concepts of the Domain Ontology (DO) and to the IWN TO is permitted and visualized in the integrated consultation: each term is connected to one or more domain dependent concepts constituting the "core" set of concepts of the domain modelling; at the same time, the plug-in relations described above bridge the term to the TO which IWN inherited from EWN.

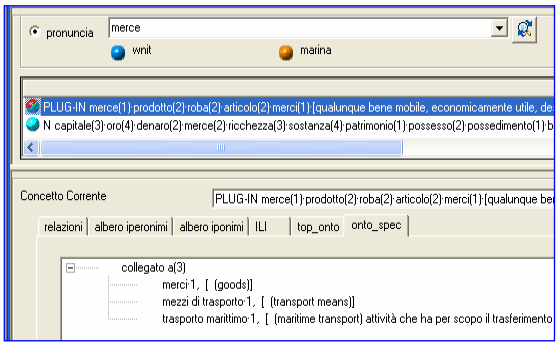

Figure 4. The link to the Domain Ontology

Figure 5. The link to the IWN TO

Thus, using Miller and Fellbaum's words (2007), the mapping of higher-level concepts to a lexical resource extends the concepts covered by the ontology down into the "leaves" of the hierarchies.

In such a way the knowledge of a term is assured from both a specialized point of view, directly connected with the specific domain of interest and a general, foundation perspective (Marinelli and Spadoni, 2007).

5. Future work and conclusion

The relevance of these specialized knowledge fields and the presence of this kind of terms in our everyday life leads us to carry on this project i) increasing the lexical coverage of each database, ii) starting a cooperation with the concerned organizations[2] in order to enrich and refine these lexicons and reach a definite version officially recognized and validated, iii) exploiting the web consultation: the maritime database is already available in the Internet, and our future research will consist in making available all the three terminological resources with the realization of a navigation "dashboard" which could be greatly useful in many future (commercial, business, didactic, etc.) activities, but also could give the possibility to combine and compare information from multiple independently-created resources.

Each domain specific synset in the terminological wordnets will have at least one equivalence relation with a synset or record in the ILI, in a perspective of a multi-lingual information retrieval.

Furthermore, we believe that the link between the specialized wordnets and the generic lexicons could be relevant also to maintain our linguistic identity, allowing to obtain both clear definitions and unambiguous translations of specific terms.

These terminological resources were created to answer the needs of various kind of communities, professionals and non-professionals alike, in the frame of different projects; they can be seen as a first nucleus of an organized network of generic and specialized lexicons in such a way as to make clearer, from a cognitive point of view, the meaning of each term which is conceived as a point of access in the terminological knowledge network

[2] For example, for the MDB, organizations such as Federagenti / Federazione Agenti Marittimi, Assoporti / Associazione delle Autorità Portuali Italiane.

6. References

Church K., Hanks P. (1989). Word association norms, mutual information and lexicography. In *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics,* Vancouver, British Columbia, pp. 76-83.

Evans V., Green M. (2006). *Cognitive Linguistics. An Introduction.* Edinburgh: EUP.

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, pp. 511-515.

Marinelli R. (2008). Enhancing a Terminological Database with Terms from a Scientific Domain. In *Proceedings of the Third Baltic Conference on Human Language Technologies*, October 4-5 2007, Kaunas, Lithuania, pp.165-172.

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.* 3 (4), pp. 235-244.

Miller G. A. (2003). The cognitive revolution: a historical perspective. *Trends in Cognitive Sciences.* 7 (3) 2003, pp. 141-144.

Miller G. A., Fellbaum C. (2007). WordNet then and now. In *Language Resources & Evaluation.* 41, pp. 209-214.

Murphy G. L. (2004). *The big book of concepts*. Cambridge MA, MIT Press.

Picchi E. (2003). PiSystem: sistemi integrati per l'analisi testuale. *Linguistica Computazionale*, Special Issue, XVIII-XIX, II. Pisa, Giardini, pp. 597-627.

Rosch E. (1978). Principles of Categorization. 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.

Roventini, A., Alonge, A., Bertagna, F., Calzolari, N., Cancila, J., Girardi, C., Magnini, B., Marinelli, R., Speranza, M., Zampolli, A. (2003). ItalWordNet: Building a Large Semantic Database for the Automatic Treatment of Italian. In *Linguistica Computazionale*, XVI-XVII. Pisa, Giardini, pp. 745-791.

Vossen P., L. Bloksma, H. Rodriquez, S. Climent, N. Calzolari, A. Roventini, F. Bertagna, A. Alonge, W. Peters. (1998). The EuroWordNet Base Concepts and Top Ontology. *EuroWordNet (LE-4003) Deliverable D017D034D036*, University of Amsterdam.

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