Extending the Italian WordNet with the Specialized Language of the Maritime Domain

Adriana Roventini and Rita Marinelli

Istituto di Linguistica Computazionale, Consiglio Nazionale delle Ricerche,
via Moruzzi 1, Pisa, Italy
Email: adriana.roventini@ilc.cnr.it, rita.marinelli@ilc.cnr.it

Abstract. In this paper we describe the creation, we are carrying out of a specialized lexicon belonging to the maritime domain (including the technical and commercial/maritime transport domain) and the link of this lexicon to the generic one of the ItalWordNet lexical database. The main characteristics of the lexical semantic database and the specific features of the specialized language are described together with the coding performed according to the ItalWordNet semantic relations model and the approach adopted to connect the terminological database to the generic one. Some of the problems encountered and a few expected advantages are also considered.

1 Introduction

The growing amount of non-structured information, stored in natural language, requires the availability of computational instruments able to handle this type of information. In this context, the extension of the ItalWordNet (henceforth IWN) database with the navigation and the shipping terminology, constitutes an important enrichment, given the remarkable incidence of this lexical domain in many contexts of everyday and business life; in its turn, the specialized lexicon gains semantic information automatically manageable, as well as the link to WordNet 1.5.

The globalisation of trade, business and travel, alongside technology development are producing changes also within the maritime activity and the related terminology; consequently the techniques of communication, translation and diffusion of terms have also changed. Historical reasons and, most of all, 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 'monopole' of the English language in this sector of economy. Furthermore, the great importance of transports, together with the continuous technical progress, have determined the need – for the countries involved in the transportation network – to introduce reliable tools 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.

The Italian lexical-semantic database IWN (Roventini et al., 2002), contains encoded detailed information of a semantic and conceptual type according to a multidimensional model of meaning which is particularly useful for applications dealing with textual content. Within the IWN database, lexical information is represented in such a way as to be used by different computational systems in many types of applications. Therefore, we have

Petr Sojka, Karel Pala, Pavel Smrž, Christiane Fellbaum, Piek Vossen (Eds.): GWC 2004, Proceedings, pp. 193–198.

considered it useful to take advantage of the IWN linguistic model to build and structure the specialized language of navigation and maritime transport, aware that "Adopting the perspective of linguistics to account for terms, requires their description by means of the same models that we use for other lexical units. " (Cabré, 1998/99).

In the following sections we describe: the main features of the IWN database (Section 2), the construction of the terminological subset (Section 3), the foreseen advantages and improvements (Section 4).

2 The Italian WordNet

IWN is a lexical-semantic database developed within the framework of two different research projects: EuroWordNet (Vossen 1999) and SI-TAL (Integrated System for the Automatic Treatment of Language) a National Project devoted to the creation of large linguistic resources and software tools for the processing of written and spoken Italian. During the SI-TAL project the Italian WordNet was improved and extended by the insertion of adjectives, adverbs and a set of proper names belonging to both the geographic and human domains. Moreover, a terminological wordnet was added for the economic and financial domain, in such a way that it was possible to access both the generic lexicon in the database and the specialized one, or also both lexicons at the same time (Roventini et al., 2000, Magnini & Speranza 2001).

IWN inherited the EWN linguistic model (Alonge et al., 1998) which provides a rich set of semantic relations, and the first nucleus of data (verbs and nouns). The wordnet was structured in the same way as the Princeton WordNet (Miller et al., 1990, Fellbaum 1998) around the notion of synset (i.e. a set of synonymous word meanings), but many other semantic relations between the synsets were identified and extensively (e.g. the hyponymy or IS-A relation) or partially encoded; among these the cross-Part of Speech (PoS) relations between words referring to similar concepts and belonging to the same semantic order: for example the noun *ricerca* (research) and the verb *ricercare* (to research), which indicate the same situation or eventuality, are linked by a xpos_near_synonym relation.

IWN has also inherited from EWN the distinction between language-internal relations and equivalence relations and the Top Ontology. The language internal relations apply between synsets of the Italian wordnet, among which the hyperonymy/hyponymy relation is the most important relation encoded for nouns and verbs together with synonymy and xpos_near_synonym. This is due to the possibility it provides to identify classes of words for which one can draw generalizations and inferences. The equivalence relations between the IWN synsets and the Inter-Lingual Index (ILI)[1] are defined similarly to the internal relations. Thus, for instance, synonymy and eq_synonymy can be defined in a similar way, the only difference being that the latter holds between a synset in the Italian wordnet and a synset in the ILI. The Top Ontology (TO) is a hierarchy of language-independent concepts, reflecting fundamental semantic distinctions, built within EWN to provide a common framework for the most important concepts and partially modified in IWN to account for adjectives and adverbs. Via the ILI, all the concepts in the generic and specific wordnet are directly or indirectly linked to the TO.

[1] The ILI is a separate language independent module containing all WN1.5 synsets but not the relations among them.

3 Construction of the Terminological Wordnet

The maritime terminological lexicon has been structured according to the design principles of the generic wordnet, i.e. applying the same semantic relations model and exploiting the possibility – available in IWN through the ILI – of linking the specialized terms to the corresponding closest concepts in English and, consequently, to the EuroWordNet multilingual lexical database.

First of all, with the suggestions of a domain expert and consulting various sources[2] we started to design the terminological data base top level, identifying the most relevant and representative domain concepts or basic concepts (henceforth BCs). The choice of these BCs was carried out following various criteria, but in particular we selected the concepts that in both the generic database and the specialized dictionaries show a large number of hyponyms, and/or that are more frequently used in this particular domain of maritime navigation and transport (Marinelli et al., 2003).

A first nucleus of over 150 BCs was identified, such as *nave* (ship), *vela* (sail), *porto* (harbour) *ormeggio* (mooring), *carico* (cargo), *spedizione* (shipment), *navigazione* (navigation), *trasporto* (transport), *tariffa* (tariff), *nolo* (freight) and so on, which are sufficiently general and constitute the root nodes of the specialized database we are developing. Most of these BCs were exported from the generic database and then imported in the terminological one exploiting the export/import capabilities of the IWN management tool. It is possible, in fact, to import or export one or more concepts as XML files. As a next step all these BCs were linked to the generic wordnet by means of the plug_in relations (see the following paragraph). Other BCs were included "ex novo", because they were not present with their maritime senses in the generic database, but very frequently used and representative of this specific domain, e.g.: *classe* (class), *fanale* (light), *armare* (to equip), *agente marittimo* (shipping agent), *punto* (position), *destino* (destination).

Starting from this first nucleus the database has then been increased, by coding the hyponyms and codifying other important semantic relations.

Most BCs are the root of a terminological sub-hierarchy and their hyponyms are often constituted by the base concept term itself followed by an adjective or a prepositional phrase which narrows and at same time specifies the meaning, a typical new-words formation that is particularly frequent in specialized languages. For instance considering the BCs *carico* (cargo), *tariffa* (tariff), *nolo* (freight) the following compounds or multiwords were encoded: *carico completo* (full cargo), *carico di merci varie* (general cargo), *carico in coperta* (deck cargo), *carico parziale* (part load cargo), *tariffa doganale* (custom tariff), *tariffa di trasporto* (transport tariff), *tariffa forfettaria* (flat-rate tariff), *nolo anticipato* (freight prepaid), *nolo intero* (full freight), *nolo secondo il valore* (ad valorem freight), *nolo a destino* (freight payable at destination).

Terms belonging to all the different grammatical categories of nouns, verbs, adjectives, adverbs and a small set of proper names are being codified in the terminological data base

[2] Several information sources have been used to select the BC: the "Dizionario Globale dei termini marinareschi", edited by the *Capitaneria del Porto di Livorno*, online on the Web; the "Dizionario di marina", edited by Barberi Squarotti G., Gallinaro I, (2002); the "Glossario dello spedizioniere" (Annuario Federspedi 1988); the "Dizionario di termini marittimi mercatili", compiled by P. R. Brodie and translated by E. Vincenzini, Lloyd's of London Press, Legal Publishing and Conferences Division, 1988.

(until now 2000 lemmas), using the many types of IWN semantic relations. The BC *porto* (harbour), for instance, is linked to *luogo* (place), by means of the hyperonymy relation; it is also connected to *imbarco* (shipment) and *sbarco* (unloading) by a role_location relation, to *avamporto* (outer harbour) by a has_mero_location relation, to the adjective *portuale* (harbour) by the has_pertained relation, to a set of proper names by the has_instance relation.

Each term is connected with the ILI by an equivalence relation: when possible an eq_synonym or eq_near_synonym relation is used, otherwise an eq_has_hyperonym relation is coded, e.g. *porto* eq_synonym harbour, *carico parziale* eq_has_hyperonym cargo; by these links to the ILI, the terms are also connected to the TO.

When the English synonym of the term was not found in the ILI and the term was linked to its hyperonym, the English synonym of the term was recorded in a list by which the ILI should be updated and enlarged. A feasibility study is envisaged with this aim.

The English term or multiword (or its acronym) is often known and used much more than the Italian one in the maritime transport activity: for instance the abbreviation RO-RO (Roll On/Roll Off) usually indicates *nave traghetto per automezzi* (ferry for vehicles transport), the abbreviation FOB (Free On Board) is used to say *con le spese pagate fino a bordo*, (loading costs paid up to ship's broadside), CIF (Cost Insurance and Freight) to say *costi fino a bordo più assicurazione e nolo mare pagati* (loading costs, insurance and sea-freight prepaid). In these and in many similar cases, we included in the synset both the English term (or multiword or acronym) and the Italian one as variants.

3.1 The Link Structure

As said before, the BCs identified for this terminological lexicon constitute the top level and are the root nodes for the plug-in operation which allows linking between the generic and specialized wordnets.

The database management tool has the following main functions: i) a simultaneous parallel consultation of the two databases to facilitate insertion of the relations; ii) three types of plug_in relations can link synsets of the two different databases: the eq-plug-in relation, as equivalence synonymy relation, the hyp-plug-in relation, as equivalence hyperonymy or hyponymy relation; iii) an integrated research between the two databases in such a way that if the synset is found in both the databases and there is an eq-plug-in relation between the synsets, the synset belonging to the specific domain partially eclipses the generic one.

As a matter of fact, once defined, a 'plug-in' relation connects a terminological sub-hierarchy (represented by its root node) to a node of the generic wordnet, so that all downward (hyponymy and istances) and horizontal (such as part-of relations, role relations, cause relations, derivation, etc.) relations are taken from the terminological wordnet, while all upward (hyperonymy) relations are taken from the generic one.

If the lemma is retrieved in both databases and there is not a eq-plug-in relation between the synsets, the synset belonging to the specific domain does not eclipse the other one and the results of the research are presented all together.

4 Final Remarks

Our choice to perform this type of study was determined by the fact that nowadays maritime terminology is object of great interest in a marine nation like Italy; furthermore, maritime

terminology dictionaries are rare and sometimes it is very difficult to find the English translation of these terms or, on the contrary, the English terms prevail over the Italian synonyms, in particular as far as maritime transport is concerned.

The availability of definitions and translations of specific terms is a useful tool for work (export-import companies, maritime agencies, etc.), for school and for didactic activities of various types (nautical Institutes, professional training, etc.) and, in general, whenever a reference to terms of this specific domain is needed.

The sea transport field is managed by English terminology, but in everyday life a constantly updated translation is necessary, on many particular occasions. From a 'commercial' point of view, the English language prevails over all other languages: contracts, negotiations, chartering and operation documents of cargo ships (bills of lading, etc.) are in English, and so are a great number of reference books. From the point of view of 'usefulness', there are circumstances in which it is necessary to refer to a translation of technical terms that is correct, abreast and absolutely unambiguous. This is for example the case of legal actions, when a judge is faced with English terminology, the Italian translation is very often difficult or unknown, and, at the same time, he is forced to refer strictly to the Italian Navigation Code written in Italian.

In this context, we think it would be desirable to carry on with this work, increasing the number of terms and starting a cooperation with the concerned organizations[3] in order to enrich and refine this maritime navigation and transport lexicon and reach a definite version officially recognized and validated, which could be greatly useful in many future activities. Furthermore, we believe that the link between the specialized wordnet and WN1.5, through the IWN generic lexicon, is essential both to face globalization and to maintain our linguistic identity.

References

1. Alonge, A., Calzolari, N., Vosse, P., Bloksma, L., Castellon, I., Marti, T., Peters, W.: The Linguistic Design of the EuroWordNet Database, Special Issue on EuroWordNet, in: N. Ide, D. Greenstein, P. Vossen (eds.), "Computers and the Humanities", XXXII (1998), 2–3, 91–115.
2. Cabré, M. T., Do we need an autonomous theory of terms?, in: "Terminology", vol. 5, n. 1, (1998/1999), pp. 5–19.
3. Dizionario Globale dei termini marinareschi, edited by the "Capitaneria del Porto di Livorno", online `http://www.capitanerialivorno.portnet.it/Dizionario/`.
4. Dizionario di Marina medievale e moderno della Reale Accademia d'Italia, Roma, 1937.
5. Fellbaum, C. ed.: WordNet: An Electronic Lexical Database, MIT Press, Cambridge, MA, (1998).
6. Layton, C. W. T. Dictionary of nautical words and terms, Glasgow, 1958.
7. Magnini, B., Speranza, M. Integrating Generic and Specialized Wordnets, in: Proceedings of Recent Advances in Natural Language Processing, RANLP-2001, Tzigov Chark, Bulgaria, 2001, pp. 149–153.
8. Marinelli, R., Roventini, A., Spadoni, G.: Linking a subset of Maritime Terminology to the Italian WordNet in: Proceedings of the Third International Conference on Maritime Terminology, Lisbon, 2003.

[3] For example organizations such as Confitarma/Associazione Armatori Italiani, Federagenti/Federazione Agenti Marittimi, Federspedi/Federazione Spedizionieri, Assologistica/Associazione dei Terminal e Imprese portuali, Assoporti/Associazione delle Autorità Portuali Italiane.

9. Miller, G., Beckwith, R., Fellbaum, C., Gross, D., Miller, K. J. (1990) Introduction to WordNet: An On-Line Database, in: "International Journal of Lexicography", 3(4), pp. 235–244.
10. Roventini, A., Alonge, A., Bertagna, F., Calzolari, N., Magnini, B., Marinelli, R., Speranza, M., Zampolli, A.: ItalWordNet: a large semantic database for the Automatic Treatment of the Italian Language in: Proceedings of the First Global WordNet Conference, Central Institute of Indian Languages, Mysore, India, 2002, pp. 1–11.
11. Roventini, A., Alonge, A., Bertagna, F., Calzolari, N., Cancila, J., Girardi, C., Magnini, B., Marinelli, R., Speranza, M., Zampolli, A.: ItalWordNet: Building a Large Semantic Database for the Automatic Treatment of Italian, in: "Linguistica Computazionale", vol. XVI–XVII (2003), pp. 745–791, Giardini, Pisa.
12. Vossen, P. (ed.): EuroWordNet General Document, 1999. `http://www.hum.uva.nl/~EWN`.