Building a Maritime Domain Lexicon: a Few Considerations on the Database Structure and the Semantic Coding

Rita Marinelli Adriana Roventini Alessandro Enea

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 – Alessandro.Enea@ilc.cnr.it

Abstract

In this paper we refer about the building we are carrying out of a specialized lexicon belonging to the maritime domain, together with the coding performed according to the ItalWordNet semantic relations model. The main characteristics of the lexical semantic database and the specific features of the specialized language are taken into consideration and, in particular, we concentrate our attention on the following items: i) the main characteristics of this lexicon, its different levels of specificity and its distribution within different sub domains; ii) the verb class semantic coding; iii) the suitable concepts to outline a specific maritime domain ontology.

1. Introduction

The Italian lexical-semantic database ItalWordNet (IWN) (Roventini et al., 2003), contains detailed encoded information of a semantic and conceptual type according to a multidimensional model of meaning (Alonge et al., 1998) 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, in this last year, we have considered it useful to take advantage of the IWN linguistic model to build and structure the specialized lexicon of navigation and maritime transport (Marinelli et al., 2003). The lack of researches in this field for the Italian language, makes it useful, in our opinion, the contribution that this instrument can provide for work and didactic activities (and in general whenever a reference to terms of this specific domain is needed), as reference for a proper technical terms use and information and for a translation abreast and unambiguous.

The maritime terminological lexicon has been structured according to the design principles of the generic wordnet, i.e. applying the same semantic relations and exploiting the possibility - available in IWN - of linking the specialized terms to the corresponding closest concepts in WordNet 1.5. Terms belonging to all the different grammatical categories of noun, verb, adjective and adverb (plus a small set of proper names) are being codified.

In the following sections we illustrate: the specialized lexicon building; the lexicon composition from a quantitative point of view; the semantic coding taking as an example the verb class; the more proper concepts to outline a domain ontology and the foreseen future work.

2. The Building of the Specialized Lexicon

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, in particular we selected the concepts that in both the generic database and the specialized dictionaries show a large number of hyponyms and/or are more frequently used in the particular domain of maritime navigation and transport. Several sources have been used to select the BCs and for any synset a definition was introduced after cross-checking, revising and summarizing those contained in the sources, also under a domain expert's supervision.

A first nucleus of BCs was identified: for the most part nouns such as *nave* (ship), *vela* (sail), *porto* (harbour), *carico* (cargo), *nolo* (freight), *ancora* (anchor), *ormeggio* (mooring), *albero* (mast), but also a few verbs such as *navigare* (to navigate, to sail), *manovrare* (to manouvre), *stivare* (to stow), 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.

Other BCs are not present in the generic database with their domain specific senses. Some of them are *armatore* (ship owner), *nolo* (freight), *classe* (class), *fanale* (light), *punto* (position), *destino* (destination), *agente marittimo* (shipping agent), *spedizioniere* (freight forwarder). Starting from this first nucleus the database has been then increased, by coding the hyponyms and other important semantic relations.

All these BCs were linked to the generic wordnet by means of the *plug_in* relations which allow links between the generic and the specialized wordnet by connecting a terminological sub-hierarchy (represented by its root node) to a node of the generic wordnet.

Two types of *plug_in* relations were codified: a) the *eq-plug-in* relation, which is a synonymy relation between synsets of the two databases, b) the *has_hyperonym (hyponym)_plug_in* relation, which is a hyperonymy /hyponymy relation between synsets of the two databases. When a specialized domain BC was not present in IWN, it was included in the generic database only if considered a quite general concept, otherwise we used the *has_hyperonym_plug_in* relation to link it to a superordinate concept.

Not only a parallel consultation of the two databases is allowed to facilitate the insertion of the relations, but also an integrated research is possible, in such a way that if a synset is found in both databases 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.

3. The Lexicon Composition

This lexicon is very complex because it involves many other fields of knowledge ranging from geography and meteorology to cartography, from astronomy and law to maritime contracts and transport technology. Furthermore events such as sailing races or publications such as ephemeredes are also involved in this domain. For this reason, within our lexicon we find different levels of specificity depending both on the hierarchical structure of taxonomies and on the many lexical items coming from various disciplines strictly connected with maritime navigation that we included and encoded in our terminological database aiming at representing this complexity. Many terms, belonging to the lexical domains of geography and meteorology, denoting natural phenomena regarding tide, wind, sea motion, clouds, atmospheric conditions or coast conformation, have great importance for maritime navigation and are present in the terminological dictionaries that we used as reference books. So also in our database we encoded both terms and more general "domain involved concepts", e.g. for geography 122 synsets were included and coded.

We have coded, up to now, the most important lemmas belonging to both the maritime commercial transport and the sailing navigation. In the following tables a few data about the lexicon composition, the types of equivalent relations to the Inter-Lingual Index (ILI)[1] and the most used semantic relations are shown. In the first table we see that nouns are the most represented but we find also a noticeable set of verbs. In the table 2 the different types of equivalence relations are shown and we see that about 50% of synsets have a synonym or near_synonym relation with WN1.5. When the English synonym of a term was not found in the ILI, the term was linked to its hyperonym by an *eq_has_hyperonym* relation and the English synonym of the term was recorded in a list by which the ILI should be eventually updated and enlarged. In a few cases, when the English term is well known and used in alternative to the Italian one, we included in the synset both the English and the Italian term as variants.

In the third table the most used semantic relations and the number of the *plug_in* relations are shown.

Synsets	1736
Lemmas	2256
Word Senses	2386
Nouns	1803
Verbs	258
Adjectives	43
Adverbs	23
Proper Names	249

Tab.1 Quantitative data

Eq_has_hyperonym	655
Eq_synonym	496
Eq_near_synonym	358
Eq_belongs_to_class	204
Eq_involved	76
Eq_has_holonym	63

Tab. 2 Equivalence_relations

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

Has_hyperonym	1048
Has_hyponym	1048
Belongs_to_class	211
Has_instance	211
Fuzzynym	156
Has_holo_part	117
Has_mero_part	117
Xpos_near_synonym	102
Involved	67
Role	67
Role_patient	61
Involved_patient	61
Antonym	56
Role_instrument	43
Involved_instrument	43
Has_subevent	40
Is_subevent_of	40
Has_holo_location	35
Has_mero_location	35
Role_location	24
Involved_location	24
Plug_in relations	228

Tab. 3 The most used internal semantic relations and the *plug_in* relations

4. The Semantic Coding of the Verbs

The verb class, in this domain, constitutes an interesting subset containing a high percentage of terms belonging exclusively to the maritime lexicon. The subset is formed by 258 word senses distributed in 187 synsets. These verbs, for the most part, represent actions and movements and can be roughly grouped as follows: verbs indicating general actions which precede or make possible the navigation such as: *armare* (to equip), *carteggiare* (to chart), *varare* (to launch); verbs which are general enough to be referred to all types of navigation such as: *salpare* (to weigh anchor), *fare scalo* (to call at, to out in), *approdare* (to dock); verbs which refer to sailing as *filare* (to ease up), *lascare* (to loosen), *orzare* (to luff), (this group is the most numerous and homogeneous because they all are hyponyms of *manovrare* (to manoeuvre) and are sub events (i.e. show temporal inclusion) of *navigare a vela* (to sail); verbs which exclusively refer to maritime transports as *rizzare* (to lash), *sollevare* (to lift), *zavorrare* (to ballast).

All these verbs benefit from the IWN semantic model, which allows high level of granularity in codifying the many relations holding among them.

Considering, for example, the verbs strictly related to the navigation, which are hyponyms of *manovrare condurre* (to steer, to direct), we find a very homogeneous group of 52 verbs, denoting many types of actions, or more properly manouvres, for handling and wielding sails, ropes, hulls and other navigation instruments.

Most of these verbs are telic: in fact, they denote actions directed toward precise concrete purposes.

In particular, they may be transformative telic, denoting a rapid change of state as *virare* (to tack), or resultative telic, denoting the reaching of a concrete outcome or effect as *stivare* (to stow). In many cases this telic feature is expressed by a specific object completing and determining the verb meaning toward a precise target,

e.g.: *gettare l'ancora* (to drop the anchor), *lascare una cima* (to loose a rope), etc. Also continuative verbs are represented in this subset, denoting processes that may be interrupted by the subject e.g.: *navigare*, (to navigate), *costeggiare* (to coast), *planare* (to plane). Another distinguishing feature of a few verbs belonging to this domain is the particular behavior they show in their diathesis or the relation between the subject and the action expressed by the verb itself: e.g. *sbarcare* used with the sense *scendere a terra* (to go ashore) or with the sense *porre a terra* (to unload), showing a transitive/intransitive alternation.

In this subset the semantic coding allowed to point out the many pairs showing semantic opposition, temporal inclusion and cause/effect relation. Furthermore the various agents or patients, instruments and locations involved in these operations were explicited by means of the semantic *role* relations.

In the following picture (Fig.1) we see the coding of the verb *allascare, lascare* (loosen, make loose, make looser) which has an *antonym* relation with its opposite *cazzare, tesare, bordare* (to tauten, to firm, to make taut), an *invoved_instrument* relation with *scotta* (mainsheet), a *sub_event* relation with *navigare a vela* (to sail), a *causes* relation with the adjective *lasco* (loose), an *hyperonymy* relation with the verb *manovrare* (to manoeuvre, to operate, to maneuver).

In the second picture (Fig.2), the coding of the verb *stivare* (to stow) is shown: an *xpos_near_synonym* relation with *stivaggio* (stowage) is used, an *hyperonymy* relation with *porre, situare, mettere* (to put), an *involved_agent* relation with *stivatore* (stevedor), an *involved_patient* relation with *carico* (cargo), an *involved_location* relation with *stiva* (hold), an *is_purpose_of* relation with *avvolgere* (to roll up).

All these verbs show a similar rich encoding which highlights the many relations holding among them.

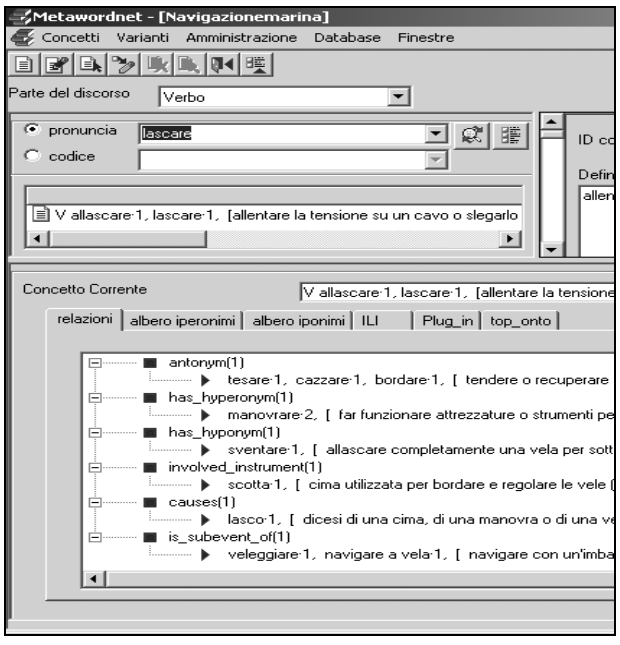

Fig. 1 The verb *lascare* (to loosen)

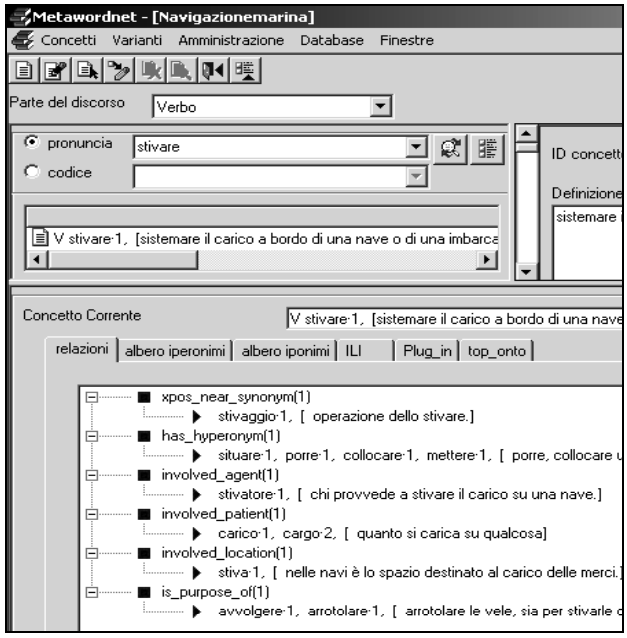

Fig.2 The verb *stivare* (to stow)

5. Outlining a Domain Ontology

An ontology is a set of concepts and relationships that reflect the overall conceptual model of a specific knowledge domain, the explicit formal specifications of the terms in the domain and the relations among them (Gruber, 1993). Up to now our terminological database is connected, by means of the *plug_in* relations, to the general ontology which IWN inherited from EuroWordNet (Vossen, 1999). But here we try to single out a group of concepts which could be the starting point in outlining a new domain ontology design, based on declarative knowledge representation. In our database there is an explicit description of the concepts belonging to a specific domain, of the properties of each concept with the various features and the relations holding among the concepts.

Among several viable alternatives, we have to determine which one would work better for the planned task, or would be more intuitive, more extensible, and more maintainable. We also need to remember that an ontology is a model of reality of the world and the concepts in the ontology must reflect this reality (Noy and McGuinness, 2001). In this maritime domain, for example, we could classify the concept *'ship'* into military, passengers, or cargo ships, considering the different uses. Alternatively, from a different point of view, we could divide the concept of ship into sailing or propeller ships.

5.1 Developing a Concepts Hierarchy

When beginning our work on the maritime domain it was important to get a comprehensive list of terms without worrying about possible overlap of the concepts they represent. So we started with the combination of a top-down development process, through the definition of the most general concepts in the domain and the subsequent specialization of the concepts, and a bottom-up development, defining the most specific concepts, and

then grouping them under more general concepts. The 'combination' approach is often the easiest way of developing an ontology, since the concepts 'in the middle' tend to be the more descriptive concepts in the domain (Rosch, 1978).

Therefore we started with a few top-level concepts such as s*hip*, and a few specific concepts, such as *cargo ship* or *passenger ship.* We then related them to a middle-level concept, such as *merchant ship.*

Then we divided/distributed the many types (hyponyms) of *cargo ship* or *passenger ship*, by that structuring a number of middle-level concepts and their hyponyms of various levels.

In defining a class hierarchy, we have to consider that the ontology should not contain all the possible information about the domain. Ontology design is a creative process, which tries to guarantee not completeness, but consistency. (Gruber, 1993). We can assess its quality enlarging, testing and refining it, actually, using it (Noy and McGuinness, 2001).

As far as our top level concepts are concerned, the problem arises of establishing the criteria of classification: for example *ship* has a huge number of hyponyms, but, as we said above, they can be classified from different points of view: on the basis of the type of propulsion (oars, sails, propeller), of the use for which they were built (transport of goods or passengers, competitions, war operations, etc.), of the place where they move (river, lake, sea). As most knowledge-representation systems allow, multiple inheritance in the concepts hierarchy is represented: a concept can be a subconcept of more than one concept. For example, if in the domain specific ontology we defined the two separate classes of sailing vessels and of military ships, the *Vespucci* would inherit both the concepts being at the same time a sailing vessel and a military ship.

Hereafter two sets of concepts are shown regarding respectively the technical/nautical and the transport domain, which, according to our experience, can be considered representative of these two sub-domains and useful to define a specific domain ontology.

TECHNICAL/NAUTICAL	**TRANSPORT**
• **Charting**	• **Transport means**
• **Direction**	• **Handling equipment**
⇒ **Points of sailing**	• **Logistic**
⇒ **Manoeuvres**	⇒ **Transp. techniques**
• **Steering**	⇒ **Stowage**
• **Equipment**	• **Goods**
• **Events**	• **Maritime contract**
⇒ **Sailing races**	⇒ **Documentation**
• **Forecasts**	➢ **Documents**

Tab. 4 The most representative domain concepts

These terms could be considered the main concepts in the ontology and become the 'anchor' points in our domain hierarchy**.** As pointed by Gruber (1993), there is no single correct ontology-design methodology.

The concepts that we present here are the first ones that we propose as useful in our own ontology-development purpose.

6. Final Remarks

Up to now the many IWN semantic relations are exploited, and our experience confirms that they fit well the terms of this specialized lexicon. We think that in this maritime domain more than in other fields the use of semantic relations can help us in giving more information about each term and make our conceptual connections quicker and more agile in comparison with simple taxonomies.

Nevertheless, it would be worth to create 'ad hoc' relations aiming at making this terminological database consistent and totally 'autonomous'. We would implement the domain ontology by linking the high nodes of our lexicon (i.e. the basic terms we plugged to the generic IWN) to the anchor concepts we showed above. Furthermore we would reach a good coverage of this lexical domain introducing other terms as well as other proper names and acronyms which are very important specially in the transports sector.

References

- 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, in: N. Ide, D. Greenstein, P. Vossen (eds.), «Computers and the Humanities», XXXII (1998), 2-3, 91-115.
- Dizionario Globale dei termini marinareschi, edited by the "Capitaneria del Porto di Livorno", http://www.capitanerialivorno.portnet.it/Dizionario/
- Dizionario di Marina medievale e moderno della Reale Accademia d'Italia, Roma, 1937.
- Fellbaum, C. ed.: WordNet: An Electronic Lexical Database, MIT Press, Cambridge, MA, (1998).
- Gruber, T.R. (1993). A Translation Approach to Portable Ontology Specification. Knowledge Acquisition 5: 199-220.
- Marinelli, R., Roventini, A., Spadoni, G.: Linking a subset Maritime Terminology to the Italian WordNet in: Proceedings of the Third International Conference on Maritime Terminology, Lisbon, 2003.
- Noy, N.F., McGuinness, D.L., "Ontology Development 101: A Guide to Creating Your First Ontology". Stanford Knowledge Systems Laboratory Technical Report, 2001.
- Rosch, E. (1978). Principles of Categorization. *Cognition and Categorization*. R. E. and B. B. Lloyd, editors. Hillside, NJ, Lawrence Erlbaum Publishers**:** 27-48.
- 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.
- Vossen, P. (ed.): EuroWordNet General Document, 1999.http://www.hum.uva.nl/~EWN.