Some Considerations in Structuring a Terminological Knowledge Base

Rita Marinelli
Istituto di Linguistica Computazionale C.N.R.
Via Moruzzi 1, 56124 Pisa Italy
Rita.Marinelli@ilc.cnr.i

Giovanni Spadoni
Sauro Spadoni s.r.l. Shipping Agency
Via delle Cateratte 90, 57100 Livorno Italy
g.spadoni@saurospadoni.it

Abstract

Exploiting the computational instruments of ItalWordNet (IWN), we built a terminological Database containing about 3000 lemmas. This allowed us to outgo the concept of "dictionary", and obtain data not only described (by the definition), but also codified (by relations), easily managed automatically and linked to the corresponding closest concepts in English through the Inter-Lingual Index (ILI). We started to design the terminological data base top level, identifying the most relevant and representative domain concepts.

The users demand has determined the need of managing the ever-increasing new technical terminology which includes also very different domains as the juridical or the economic one.

Up to now our database is connected, by means of the 'plug_in' relations, to the general ontology which IWN inherited from EuroWordNet.

Now we outline a new domain ontology design, for better defining the boundary of this research, setting the base of the terminological concepts and gaining more functional information. Before defining the ontology, a reflection is preliminary about the concept of 'term' and 'domain', the 'relevance' of each term, the knowledge potential of the terminological lexicon, together with the possibility of manipulating this knowledge with huge cognitive effects, specifying how to represent it as a concrete (suitable to be instantiated) data structure.

The set of characteristics recognized in our terminological Database and verified, lead us to qualify it a Knowledge Base System, that is a body of represented knowledge, based on a conceptualized view of the world, with axioms and inference rules productive of new knowledge generated from existing one.

Introduction

We were encouraged to perform this type of study by a precise request of specialized professional users asking for a terminological maritime dictionary written in Italian and referring to the English, prevailing in this field, therefore exploiting the availability of the computational instruments of ItalWordNet (IWN) able to handle this type of information., we have carried out the building of a terminological database (DB), which contains about 3000 lemmas, belonging to the maritime domain.

1 The Terminological Database

The terminological subset 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 Inter-Lingual Index (ILI) – of linking the specialized terms to the corresponding closest concepts in English.

We started to design the terminological data base top level, identifying the most relevant and representative domain concepts or basic concepts (BCs), which constitute the root nodes of the specialized database we are developing.

The set of BCs has been selected taking into account:

1. Terms belonging either to the generic lexicon or to the specialized one.
2. Terms that have a huge number of hyponyms.
3. Terms that are significant (only) in that knowledge field.

In this case, term and base concept are assimilated, that is the main concepts of the terminological database are 'terms'.

As a first step, for the beginning of our work on the maritime domain, it was important to get a comprehensive list of the most salient terms. So we started from one hand, with the definition of the most general concepts in the domain (using the above criteria) and the subsequent specialization of the concepts (top-down development process); on the other hand, we decided to define the most specific concepts, and then to group them under more general concepts (bottom-up development) (Marinelli et al., 2003).

This 'combination' approach may be considered the easiest, since the concepts 'in the middle' tend to be the more descriptive concepts in the domain (Rosch, 1978).

The exploiting IWN and its semantic relations available as a reliable instrument, allowed us to outgo the concept of "dictionary", and obtain data not only described (by the definition), but also codified (by relations): data structured only alfabetically in the dictionaries taken into account can become synsets, linked to each other by many types of semantic relations (*hyperonymy, hyponymy, holo/mero part*, etc.) which can also be easily/nimbly managed automatically.

There are three kind of semantic relations in the Database:

- **Internal relations**: the information is encoded in the form of lexical-semantic relations between pairs of

Petr Sojka, Key-Sun Choi, Christiane Fellbaum, Piek Vossen (Eds.): GWC 2006, Proceedings, pp. 217–224. © Masaryk University, 2005

synsets (synonym sets). Synonymy and hyponymy are the most important relations encoded; this linguistic model is very rich and contains many other lexical-semantic relations such as part of, cause, purpose, subevent, belong-to-class relations etc.. (n. of relations: **4581**).

- **Equivalence relations:** between the Italian synsets and the closest concepts (synonyms, near synonyms, etc.) in an Inter-Lingual Index (ILI), a separate language-independent module containing all WN1.5 synsets but not the relations among them. By this link the possibility to use IWN and the terminological DB for multilingual applications is ensured. (n. of relations: **2079**).
- **Plug-in relations:** allow to link the specialized wordnet to the generic one connecting a terminological subhierarchy (represented by its root node) to a node of the generic wordnet (n. of relations: **286**).

Up to now our database is connected, by means of the 'plug_in' relations, to the general ontology which IWN inherited from EuroWordNet (Marinelli et al., 2004).

Now we propose to outline a new domain ontology design and to show that the terminological semantic database can actually have all the features to be considered a Knowledge Base System (KBS).

We deem however that first it is necessary to do some considerations about some outstanding concepts which we have to face with.

2 The Concept of Term

Depending on our experience, we have ascertained that it is very difficult to evaluate which are the BCs, because it is not possibile to determine with absolute precision if "ship", for instance, is a term, and if it is, why it is a term: why among the most representative terms of the specialized lexicon, there are synsets belonging either to the terminological wordnet or to the generic one.

A reflection is preliminary about the definition of 'term' or 'terminological unit' and the features that make the same word considered in the generic DB to become a term in the specialized lexicon.

We refer to recent and significant theories of terminology (Cabré, 2003) and to some more cognitive aspects of linguistic theories to support our considerations.

The terminological units in specialized domains differ from the lexical units because of their cognitive and pragmatic conditions. A term and a word are different by their way of meaning.

The terminological or specialized value of a unit is activated when the communication context requires it, highlighted by a selection of precise semantic features corresponding to the specialized meaning of the unit in 'that' determined specific field (Cabré, 2003).

There is a strong relationship between the concept of 'term' and the concept of 'domain': "the existence of the concept 'domain' is required before the concepts 'terms' or 'terminology' can be consolidated". (Kaguera, 1998).

2.1 Relevance Salience Functionality

We think that relevance, salience and function have a fundamental role in governing selection. Salience and relevance are theoretical notions which are influential in accounting for how or why certain objects, concepts, properties or actions are highlighted or preferred in natural language processing (Pattabhiraman and Cercone, 1990), while the use or function and contextual factors interact in the interpretation of utterances.

2.1.1 Relevance

From the cognitive point of view the meaning potential of a term can be explained by the importance it has as input that satisfies our expectations of relevance.

The search for relevance is a basic feature of human cognition, which communicators may exploit, improving their knowledge on a certain topic.

According to relevance theory, an input is relevant to an individual when its processing in a context of available assumptions yields a positive cognitive effect. (Sperber & Wilson, 1995).

The notion of relevance to an individual, for a given assumption and an individual with access to a variety of contexts is a matter of choice. The aim of the individual is to choose the best possible combination of assumption and context; we claim that the choice is again governed by the search for maximal relevance (Wilson, 1998).

The most important type of cognitive effect is a contextual implication, a conclusion deducible from input and context together, but from neither input nor context alone.

2.1.2 Salience

While **relevance** is related to speaker-internal factors such as goals and motivation, **salience** is connected with context and speaker-external objects or properties: a strong and supportive context improves processes of knowledge comprehension. The higher the level of **salience** of an object, the higher is its level of activation in the speaker's mind. Salient meanings of the words are the meanings that stand out as most prominent in our minds and shape how we speak and how we think. For information to be salient - to be foremost in one's mind- it needs to undergo consolidation, that is to be stored or coded in the mental lexicon (Giora, 2003).

2.1.3 Functionality

From a functional point of view, particular aspects of a given context (such as the topics discussed, the language users and the medium of communication) define the meanings likely to be expressed and the language likely to be used to express those meanings, taking into account the way the linguistic dynamics can activate the meaning potential of the words.

The terms are a way to know; actually, linguistics, philosophy and the technical-scientific disciplines consider terminology as a 'conjunction' of units with an essential aim, and, therefore, with a functional value (Cabré, 2000). In the different applications a twofold function of the terminological units is activated: the specialized knowledge representation and its conveyance. The terms are used in the specialized communication, characterized by linguistic and prag-

matic factors: pragmatics study how the meaning potentials are completely specified and actually used by the speakers; terms are worth of a new more dynamic approach, considering the meaning not only as 'content', but as a way to change the state of information of the speakers (Chierchia, 1997).

The specialized communication admits different levels of specialization, various degrees of knowledge opacity, several indexes of cognitive and terminological density and distinct aims; and to take this into account means to consider the terms with all the meaning and knowledge potential they can have (Cabré, 2000).

3 Knowledge Base

The second consideration about the specialized lexicon structuring is concerning the nature and structure of Domain.

3.1 Domain

Domains may be more or less specific; domains may be more or less tangled (Poli, 2002). The maritime domain includes also many other fields of knowledge ranging from meteorology to astronomy, from law and maritime contracts to transport technology. The detailed structuring of a context of analysis with respect to its sub-domains a very complex task. Within our lexicon, in fact, 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 and maritime transport. They were included and encoded in our terminological database aiming at representing this complexity.

Now we want to better define the domain of interest drawing on the 'extensive' definition of terminology given as 'the set of all terminological units belonging to a specialized knowledge field' (Cabré, 2000), that can be represented in a more schematic and formal way by the symbolic language of the FOL.

We would give an inductive definition of 'term belonging to this specialized lexicon', that suggests how it is possible to collect the elements of the set considered, defining it through its genesis/developing.

We can use a function symbol and the First Order Logic formalism; FOL is often used for knowledge representation: it is considered as the formalized substitute of the natural language.

We define the predicative function f: "concerning the sea, the navigation, the transports" and the set of argument values for which this function is defined:

$$M = (\forall x.\, f\,(x))$$

where f is the " characteristic function" of the set Mbecause (every) its argument is an element of M.

So M can be considered the 'conceptual universe' (Lyons) or 'domain' specified by the set of argument values for which the function is defined.

3.2 Knowledge Base and conceptualization

The conceptual universe represents the domain and the domain can be defined 'intensionally' by the characteristic function or "extensionally" by the set of all elements that satisfy the given property.

To do this, the domain has to be structured by means of a systematic explicit (formal) specification of how to represent the objects, concepts and other entities that are assumed to exist in the/this area of interest and the relationships that hold among them.

For Artificial Intelligence (AI) systems, what "exists" is that which can be represented; so we can describe the ontological structure defining a set of representational **terms** and (formal) axioms and rules that constrain the interpretation and well-formed use of these terms, so that an inferential mechanism (possibly very simple) for knowledge managing can be elicited.

Every knowledge base or knowledge-based system is committed to some conceptualisation, explicitly or implicitly. Choosing a conceptualisation is the first stage of knowledge representation concerned with designing and using systems for storing knowledge - facts and rules about some subject or domain (Marinelli and Roventini, 2005).

We define a common "vocabulary" for researchers who need to share information in this domain, for professionals and not professionals as well to enable reuse of domain knowledge, to clarify and separate domain and operational knowledge. We describe our domain structure taking into account the need of managing the ever increasing new technical terminology which includes also very different domains as the juridical or the economic one. Our approach to information integration and ontology building is not to create a homogeneous, rigid system with a reduced freedom of interpretation, but admitting and navigating alternative interpretations, conceiving different context of use which has to be promptly highlighted for effective usefulness. To do this we require a comprehensive set of basic concepts, organized in such a way to admit the existence of different possible pathways among subdomains under a common conceptual framework. Our analysis and modelling processes should be guided by domain independent criteria and relations i.e. by an upper ontology.

IWN top ontology can be considered as an upper ontology, including the most general high level concepts, divided at the first level in three types of entities: the 1^{st} order entities that are distinguished in terms of four main ways of conceptualizing or classifying a concrete entity (Origin, form, composition, function); the 2^{nd} order entities, classified using two different classification schemes (the SituationType and the SituationComponent); the 3^{rd} order entities limited in number and so not further subdivided. A domain-independent (upper) ontology should characterize all the general notions (such as *cause, subevent, part, object, process, location, movement, person*, etc.) that are needed to talk about navigation, charting, goods species, transport techniques, etc.

Our domain structure is described defining a core set of terms representing the main two subdomains specified in the maritime terminology that are: the *technical/nautical* and the *transport* one, to be supported by specialized documentation

and studied by ontological engineers and domain experts in close collaboration. They are general enough to be the root nodes of the core ontology we are developing. The model of this structure is WN-like, as the database itself as well: the most important relations are the is-a relations and among the "horizontal" relations, only a subset is exploited (is means, for purpose, role, has instance, etc.). In facts, they seem to be the most appropriate to characterize the main conceptual schemas that people of the technical-nautical or maritime transports "world" actually use, that is activity plans, programs involving particular devices for cargo stowage, goods shipping, navigation managing, etc. (See Fig. 1 and Fig. 2)

While the top concepts are mostly domain dependant, the link with the Top Ontology of IWN remains exploiting again the plug-in relations: in such a way it is possible to guarantee either general and "basic/fondational" information, or detailed information directly connected with the specific domain. In particular it has to be noted that by the means of the plug-in relation connecting these BC to correspondent IWN concepts, our "tool" allows to extend the IWN top ontology to the maritime domain: through the semantic relations linking the synsets, every term "inherits" the top ontology definitions and becomes itself an integral part of the structure.

At the same time while codifying a term in the maritime database, the "tool" allows the reference to the BC of the terminological ontology embedding the term in the semantic network.

Upper and core ontologies provide the framework to integrate in a meaningful way different *views* on the same domain, such as those represented by the *queries* that can be done to an information system (Gangemi, 2005). (see Fig. 4 and Fig. 5).

For generality, we prefer to define an ontology rather loosely as a set of terms, associated with definitions in natural language and, if possible, using formal relations and constraints, about some domain of interest. Terminological Ontologies used for natural language applications tend to be more general (high-level, abstract), especially such language-related ontologies, while Domain Models used for domain-oriented applications are naturally more specific (Hovy, 2001).

A collection of knowledge represented using some knowledge representation language is known as a knowledge base.

We view a Domain Model as an ontology that specializes on a particular domain of interest, and fits to our terminological knowledge base representation.

In this case the semantic relations, inherited from the generic database IWN, are viewed as the knowledge representation language in the database; we can consider as axioms in this knowledge base the constrain/rule regulating the application of the semantic relations, e.g.: compelling a) to define the proper names instances of classes and not hyponyms, or b) to consider the *belongs_to_class* relation as the 'characteristic' code available only for proper names, or c) to apply the *antonymy* relation only between synsets belonging to the same grammatical category, etc.

We propose to find a set of rules and constrains to explicit in order to give and possibly grant an axiomatic structure for the conceptualization of the database.

The deduction and proof activity originates the knowledge that is implicitly contained in the initial knowledge appearing in the form of axioms.

3.3 Inference Rules

Knowledge differs from data or information in that new knowledge may be created from existing knowledge using logical **inference** i.e. the logical process by which new facts are derived from some known facts by the application of inference rules.

Inference is usually a multi-step process. Each step leading from premises to conclusion must be licensed by a rule of inference in the system (Pustejovsky, 2004).

A KB expressed in a predicative language can be asked in a forward or a backward way: in the first case, beginning from initial facts, applying repeatedly the inference rules one can obtain all that springs out; in the second case, beginning from the fact that we want to obtain, we try to test if it is deducible from the initial facts.

We can consider the inference rule allowing us to confirm the transitivity or the inheritance of the *hyperonymy* relation or of the *hyponymy* relation, i.e.:

Ancoraggio (anchorage) *has_hyperonym* manovra (manoeuvre)

Manovra (manoeuvre) *has_hyperonym* azione (action)

Ancoraggio (anchorage) *has_hyperonym* azione (action)

Barca (boat) *has_hyponym* barca a vela (sailing boat)

barca a vela (sailing boat) *has_hyponym* dinghy (dinghy)

barca (boat) *has_hyponym* dinghy (dinghy)

In this way the hyponyms of "barca a vela" (sailing boat) become also the hyponyms of "barca" (boat) and therefore this type of relation can be increased, expanded to a more numerous set of hyponyms.

We could propose also an inference rule that allows to confirm the transitivity of the part-of relation:

Nave (ship) *has_mero_part* scafo (hull)

Scafo(hull) *has_mero_part* fasciame (planking)

Nave (ship) *has_mero_part* fasciame (planking)

Applying this inference rule means to obtain new explicit and inferred *part_of* relations.

We could propose also an inference rule that allows us to inherit the *part-of* relation through an hyperonymy or through an hyponymy chain:

Albero di maestra (mainmast) *has_hyperonym* albero (mast)

Albero (mast) *has_mero_part* testa d'albero (masthead)

Albero di maestra (mainmast) *has_mero_part* testa d'albero (masthead)

Alberatura (masting) *has_mero_part* albero (mast)

Albero (mast) *has_hyponym* trinchetto (foremast)

Alberatura (masting) *has_mero_part* trinchetto (foremast)

Studying other types of relations such as the *cause* relation or the *has_subevent* relation, that are available in the

DB for verbs coding, we deem interesting to focus on the behaviour of the transitive property with the verbs approdare, attraccare, ormeggiarsi:

Approdare (to shore) *has_subevent* attraccare (to dock)

Attraccare (to dock) *has_subevent* ormeggiarsi (to moor)

Approdare (to shore) *has_subevent* ormeggiarsi (to moor)

Issare le vele (to hoist) *cause* prendere vento (to take the wind)

Prendere vento (to take the wind) *cause* aumentare velocità (to take speed)

Issare le vele (to hoist) *cause* aumentare velocità (to take speed)

It would be worth while highlighting the behaviour of the transitivity of these two relations, comparing what happens in the generic Italian Wordnet and in the terminological one.

We think that it is possible to speak about a "weak" transitivity, i.e. possible and a "necessary" transitivity, i.e. conceived as sequence of actions strictly tied up by a causality relationship.

Moreover also the possibility of applying the *xpos_near_synonym* relation implies inferences productive of knowledge. We can consider the knowledge potential that is implicit in every semantic relation of the database to confirm the inferential capabilities of the Knowledge Base System.

Many other examples of semantic relations in the terminological database could be taken under consideration to compare the generic and the specialized database, starting from some practical examples to focus in particular on the reference relationship, to investigate about our intuitions of the semantic commitments, based on a system of inference rules in anyway realized in our mind. We have to take into account such rules supposing that they are useful and necessary to deal with sets of objects with a certain structure and point out structural properties.

In choosing a Domain Model there are 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. An ontology is a model of reality of the world that is not fixed, but dynamic and the concepts in the ontology must reflect this reality (Friedman Noy and McGuinness, 2001) and its potential capacity.

Up to now we dealt with the terminological KBS, examining the most significant points in the field both of cognitive linguistics and of pragmatics: the concept of 'term', the definition of 'domain', the conceptualization of the maritime terminology, the figuring of axioms and inference rules in the system. We have also to refer about the KB management, by means of a tool developed for the treatment of the data and the semantic relations, now increased and upgraded. The program for extending and/or querying the terminological Knowledge Base allows also the building and the updating of the specific ontology. At the moment a few concepts are inserted, representing the two main subdomains specified in the maritime terminology. Hereafter (see Fig. 3.3 and Fig. 1) the set of concepts is shown regarding the technical/nautical and the transport domain, which, according to our experience, can be considered representative of these two sub-domains and useful to develop a specific domain ontology.

These terms could be considered the main concepts in the ontology and become the 'anchor' points in our domain hierarchy.

Hereafter an example is shown concerning the term "stato del mare" (sea condition) as it appears in the Ontology navigation tool (Fig. 3.3).

In the Figures 3.3 and 4 it is possible to see respectively the link with the IWN Top Ontology and the link with the specific ontology.

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 Domain Model development purpose.

Conclusion

The above characteristics, verified in our terminological Database, lead us to qualify it a Knowledge Base System (KBS), that is a body of represented knowledge, based on a conceptualized view of the world, with axioms and inference rules productive of new knowledge generated from existing one. In order to manipulate this knowledge we aim at specifying how the abstract conceptualisation is represented as a concrete data structure; we want to show/highlight that it is possible to build a 'deductive terminological database' from which one can infer much more information than from initial relations, always considering that ontology design is a creative process, trying to guarantee not completeness, but consistency (Gruber, 1993) and that we can assess its quality enlarging, testing and refining it, actually, using it (Friedman Noy and McGuinness, 2001).

References

Cabré Castelvì M. T., La terminologia: representacion y comunicacion, Institut Universitari de Linguistica Aplicada, Barcelona, 2000.

Cabré Castelvì M.T., Theories of terminology, Terminology 9:2 (2003), 163-199.

Chierchia G., Semantica, Bologna, Il Mulino, 1997.

Friedman Noy N., McGuinness D.L., "Ontology Development 101: A Guide to Creating Your First Ontology". Stanford Knowledge Systems Laboratory Technical Report, 2001.

Gangemi, A., Development of an Integrated Formal Ontology and an Ontology Service for Semantic Interoperability in the Fishery Domain, Draft project plan v.7, CNR – Institute of Cognitive Sciences and Technologies, Ontology and Conceptual Modeling Group, 2005.

Giora R., On our Mind: Salience, Context, and Figurative Language, New York: Oxford University Press, 2003.

Gruber, T.R. (1993). A Translation Approach to Portable Ontology Specification. Knowledge Acquisition 5: 199–220.

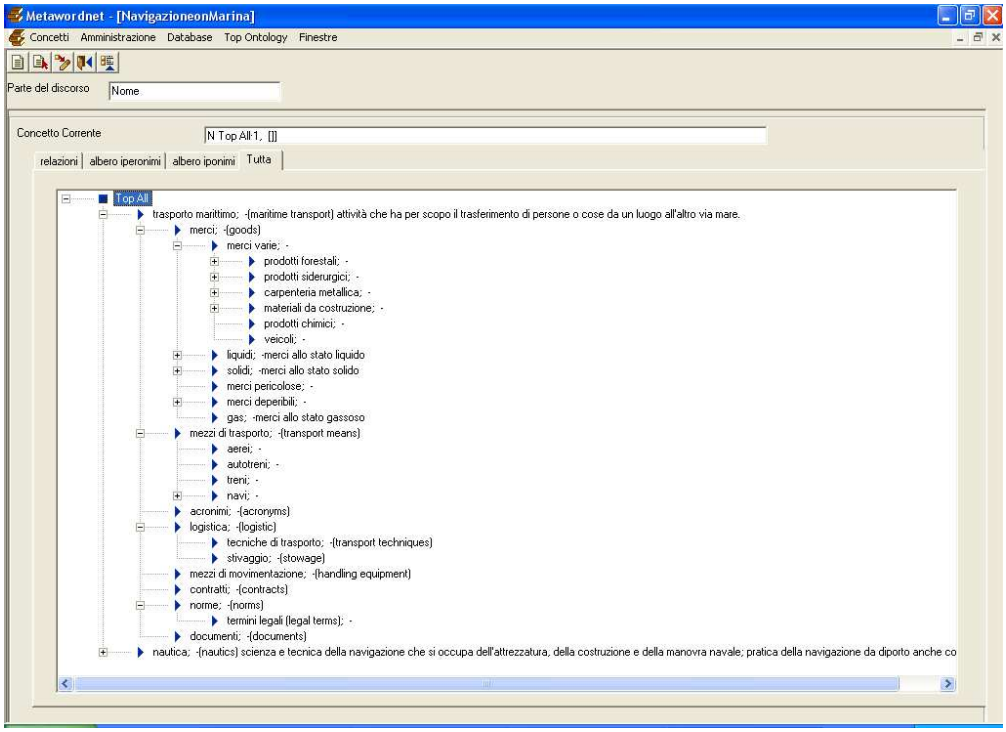

Figure 1:

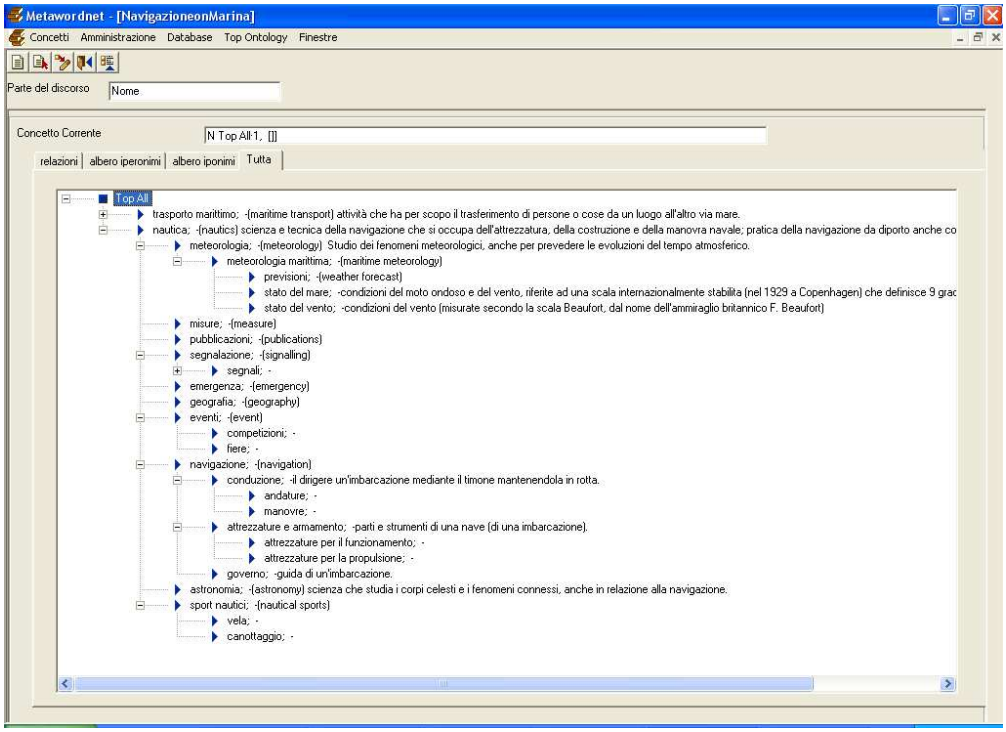

Figure 2:

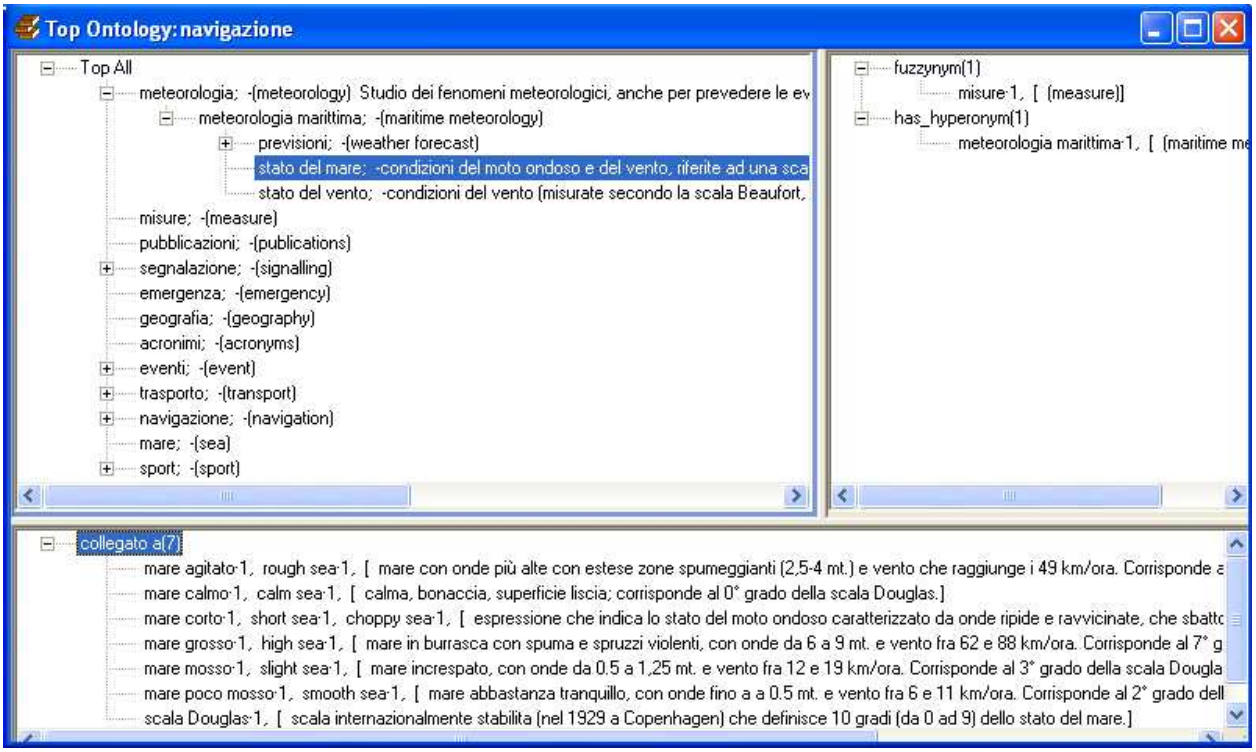

Figure 3:

Hovy E., Comparing Sets of Semantic Relations in Ontologies, in R. Green, C.A. Bean, and S.H. Myaeng (eds) Semantics of Relationships. Kluwer, 2001.

Kaguera K., The Study of Terminology from Application to Theory EAST ASIA FORUM ON TERMINOLOGY-EAFTerm 1998.

Marinelli R., Roventini A., Spadoni G. 2003. Linking a subset of maritime terminology to the Italian WordNet. Proceedings of the Third International Conference on Maritime Terminology. Lisbon.

Marinelli R., Roventini A., Enea A. 2004. Building a Maritime Domain Lexicon: a Few Considerations on the Database Strucure and the Semantic Coding. LREC 2004: Fourth International Conference on Language Resources and Evaluation, held in Memory of Antonio Zampolli. Lisbon, Portugal, 26th, 27th & 28 May 2004, Proceedings, Volume II, Paris, The European Language Resources Association (ELRA). 465–468.

Marinelli R., Roventini A. 2005. Some Considerations about the Italian Maritime Lexicon Structuring. In Atti del IX Simposio Internacional de Comunicación Social, Santiago de Cuba, 24–28 de Enero de 2005. 635–639.

Pattabhiraman T., Cercone N., Selection: Salience, Relevance and the Coupling between Domain-Level Tasks and Text Planning, Centre for Systems Science, Simon Fraser University, Burnaby, B.C., Canada, 1990.

Poli R., Ontological Metodology, in Human Computer Studies, (2002), 56, 639-664.

Pustejovsky J., CS112 Notes: First Order Logic: Part 2, 2004.

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., Marinelli R. 2004. Extending the Italian WordNet with the Specialized language of the Maritime Domain. Proceedings of the Second International WordNet Conference, GWC 2004. 193–198.

Wilson D., Relevance and Lexical Pragmatics. (Forthcoming.) To appear in Italian Journal of Linguistics/Rivista di Linguistica, Special Issue on Pragmatics and the Lexicon.

Wilson D., Discourse, Coherence and Relevance: a Reply to Rachel Giora, Journal of Pragmatics 29, 1998, 57–74.

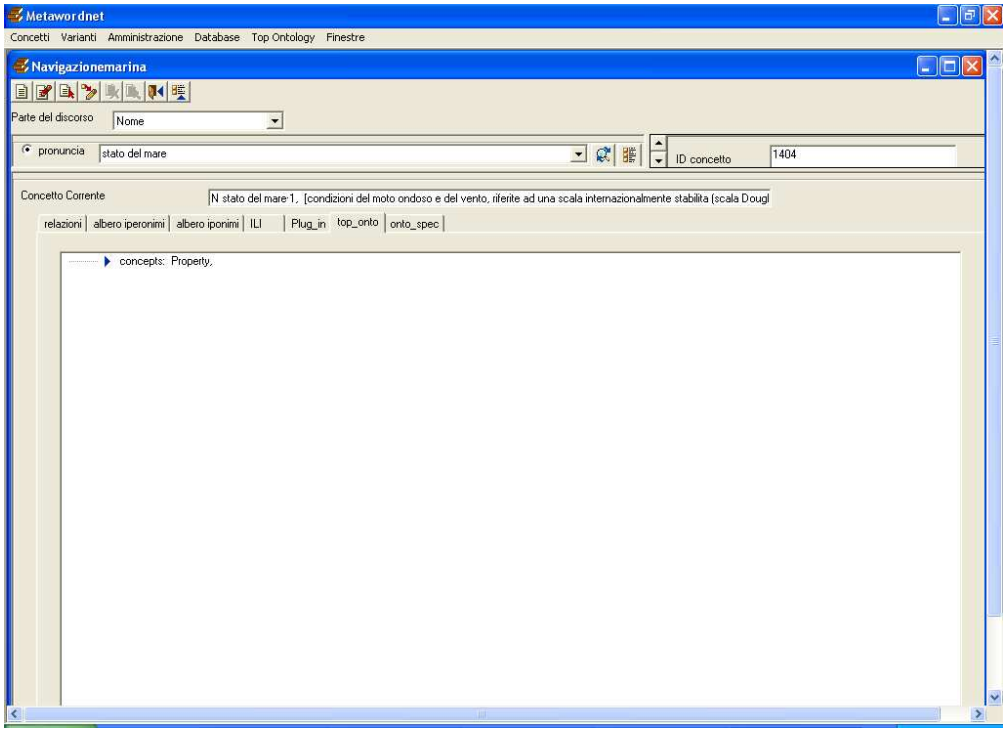

Figure 4:

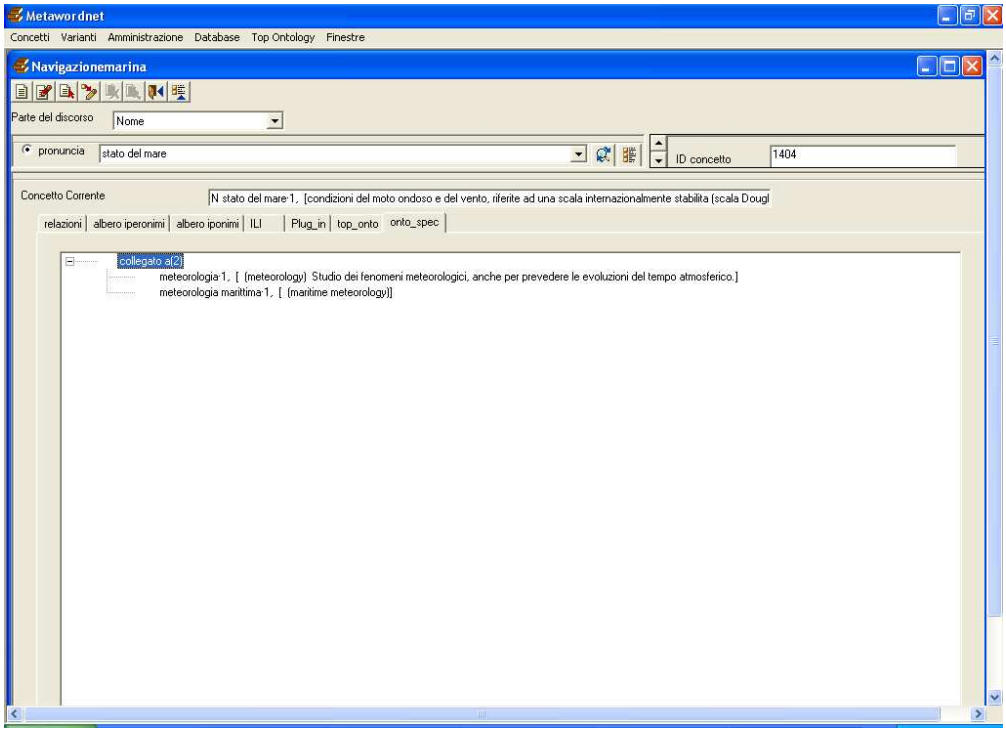

Figure 5: