Saying is not modelling

Christophe Roche

NLPCS 2007 (Natural Language Processing and Cognitive Science); pp 47 – 56. ICEIS 2007, Funchal, Portugal, June 2007

Saying is not modelling

Christophe Roche

Condillac-Listic, University of Savoie, Campus Scientifique
73376 Le Bourget du Lac cedex, France
roche@univ-savoie.fr

Abstract: In this article we claim that the conceptual modelling built from text is rarely an ontology. Such a conceptualization is corpus-dependent and does not offer the main properties we expect from ontology, e.g. reusability and soundness. Furthermore, ontology extracted from text in general does not match ontology defined by expert using a formal language. Such a result is not surprising since ontology is an extra-linguistic conceptualization whereas knowledge extracted from text is the concern of textual linguistics. Incompleteness of text and using rhetorical figures, like synecdoche, deeply modify the perception of the conceptualization we may have. It means that ontological knowledge, which is necessary for text understanding, is not in general embedded into documents. The article will end on some remarks about formal languages. If they allow to define "a specification of a conceptualization" they nevertheless raise their own issues mainly due to their epistemological neutrality. Ontology design remains an epistemological issue.

1 Introduction

Whatever their domain: information systems, databases, natural language processing, knowledge based systems, etc. applications are more and more ontology-oriented [1],[2],[3]. Such a success is mainly due to what ontology[1] promises; it means a way of capturing and representing a shared understanding of a domain that can be understood and used by humans as well as by software. Then, one of the main problems to be solved is to build domain ontology.

Ontology building, as knowledge base building, is a difficult and time-consuming task. It is the reason why a lot of works are currently done on ontology acquisition (e.g. Ontology learning ECAI workshops, KCAP workshops). Since we can consider that technical and scientific documents convey some domain knowledge, ontology building can rely on knowledge acquisition from texts [4].

Ontology building from text corresponds to a "knowledge reverse engineering from text" process described by the following figure.

[1] There is today an agreement on the definition of ontology. We can resume the most of definitions by saying that: "an ontology is a shared description of concepts and relationships of a domain expressed in a computer readable language".

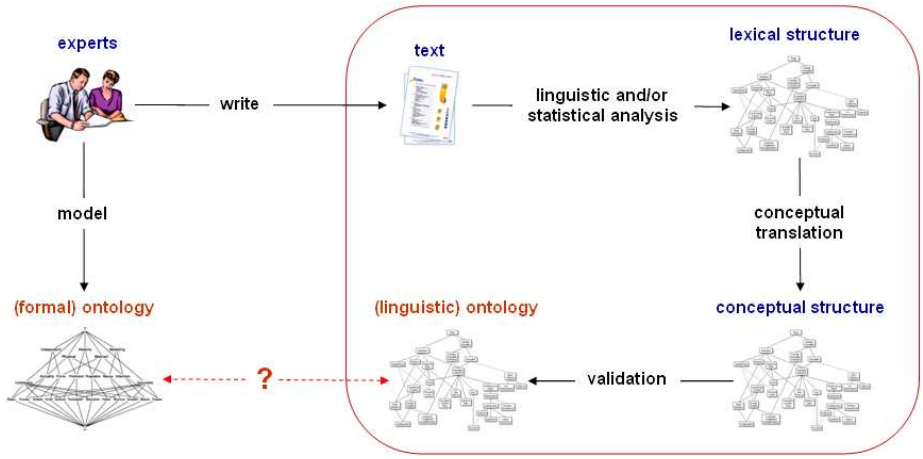

Fig. 1. Ontology building from text

This process raises several questions. The main one is "Do ontology built from text and ontology defined by experts in a formal language match?"[2] In other words what are the consequences for ontology building of using a given language, either natural or formal? What is information lost – and what is information introduced – when experts write text in natural language? On the other hand, do – and how – formal languages influence conceptualization? This article will try to give some answers to these questions.

2. Ontology Reverse Engineering from Text

2.1 Industrial Context

In order to illustrate our talk, let us take an industrial application carried out by Ontologos corp.[3] for EDF Research & Development (Electricity of France) [5], [6]. The main goal of this application was the re-appropriation of the ontology describing the concepts in the field of control and instrumentation. The problem is all the more difficult in that this knowledge is not directly accessible in usable form but is spread out through various bodies of knowledge, and more especially into documents. In particular we worked on a corpus[4] about relay. The ontology of relay we have defined

[2] The difference between these ontologies appears through the expressions "linguistic ontology" and "formal ontology".

[3] http://www.ontologos-corp.com

[4] A corpus is a collection of texts which have been selected according to some criteria [7] and for a given objective. In the framework of our application, the criteria were mainly: "produced by a same community of practice", "about a same topic", "belonging to the same type of text (descriptive)", "under the same form".

is currently used for different applications including a content management system for document classification and information retrieval.

2.2 Lexicon and Lexical Structure

The goal of the first stage is to build a lexical structure. It means to build a network of words linked by linguistic relationships; where words – in general nouns or noun phrases – denote concepts[5] and linguistic relationships are mainly hyponymy, synonymy and meronymy relationships. Thus the first step is to extract candidate terms for concept's names. Extracting candidate terms and linguistic relationships from corpus by automatic text analysis is today an active research domain [4], [8], [9]. Statistical methods based on Harris's distributional hypothesis, i.e. collocation analysis of terms [10], as well as linguistic methods based for example on regular expressions can be used. Regular expressions like "adjective noun" and "noun noun" patterns allow the extraction of expressions like "electromagnetic relay", "threshold relay", "on/off relay", "voltage relay", "undervoltage relay", 'overvoltage relay", etc. from the relay corpus. The result, which must be validated by experts, is a lexicon of words of usage considered as many as possible concept's names.

This lexicon is structured according to linguistic relationships like hypernymy (versus hyponymy), synonymy, meronymy and so on. Here too, these linguistic relationships can be 'automatically' extracted from the corpus using both a syntactic analysis – "a voltage relay *is* a kind of relay" – and the lexical structure of noun phrases. For example, linguistic expressions made up of several words with the same ending (i.e. ending with the same words, for example with the same noun) give interesting information about the structure of the lexicon. The following linguistic expressions "voltage relay", "threshold relay", "electromagnetic relay" can be considered as many as hyponyms of "relay" (let us recall that in this article, linguistic expressions are given between quotation marks).

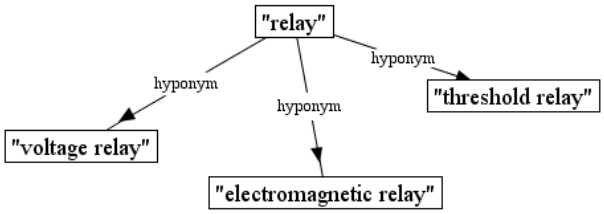

Fig. 2. A part of the lexical structure.

[5] In this article, names (nouns and noun phrases) as well as linguistic expressions will be written between quotation marks: "relay", "voltage relay", "A voltage relay is a kind of relay." etc.; while concepts will be written between the lower and upper symbols: the name "relay" denotes the concept <relay>.

2.3 Conceptual Structure and Ontology

The second stage is to deduce the conceptual structure from the lexical one. If we assert that a term[6] denotes a concept and the hyponymy relationship is a linguistic translation of the subsumption relationship, then the noun phrase "voltage relay" denotes the concept <voltage relay> which is a sub concept of the concept <relay>. The result is a conceptual structural which matches the lexical one.

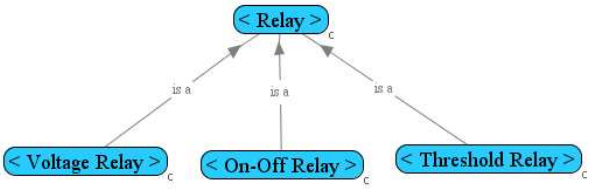

Fig. 3. The conceptual structure defined from the lexical structure.

The last, but not least, stage is the validation of the conceptual structure. During this stage, the conceptual structure is completed as necessary, concept's names are normalized and words of usage are associated with concepts. The result is the ontology of domain. In our example, since writing technical documents relies on the terminology[7] of the domain, the previous conceptual structure has been labeled as a valid ontology of the domain.

3. Does It Really Work?

The process of ontology building from text is quite clear and well defined, and a lot of very interesting work is currently being done on these different stages. But is it so simple (even if statistical and linguistic methods can be very complex)? And does such ontology really satisfy our expectations?

3.1 Application-Oriented Validation

One generally says that an ontology is defined for a given goal [1]. Let us take the example of an ontology-oriented content management system. As a lot of ontologies built from text, the ontology of relay is both used for classifying documents[8] and information retrieval [12]. Thus, search for information about a concept, for example <threshold relay>, must return all information about this kind of relay and about its more specialised concepts (using if necessary the 'is a' inheritance relationship for inferences). But in our case, no information about <voltage relay> is returned when it should do so. For experts, all information about <voltage relay> concerns <threshold

[6] A term can be considered as a lexicalised concept like in Wordnet [11] where a concept is defined as a *synset*, i.e. a set of synonymous words.
[7] A terminology can be viewed as a constraint language based on a normalized vocabulary.
[8] A document is classified on every concept the content of document refers to.

relay>. Although the conceptualization is not wrong since it has been validated by the experts (a <voltage relay> is really a kind of <relay>), it is not completely correct. Where is the problem?

3.2 Ontology from Text and Ontology from Expert

Ontology built from text (also called "linguistic ontology") is corpus-depend, which is not surprising. It means that even when different communities of practice share a same reality (for example between users and suppliers about relay), it is not possible to define a sharable and reusable ontology from text as far as these communities use their own language (Language for Special Purpose). It is the reason why some say that "these ontologies are domain- and task-specific: concept definitions result from the selection of a single interpretation context that reflects the application requirements; they are intended to reflect one of the ways knowledge can be perceived through the use of language in documents" [8].

In order to better understand the problem we were faced with, experts were asked to define directly their domain conceptualization in a formal language, independently of the words of usage[9]. The result is a formal ontology quite different from text-oriented ontology. An ontology-dedicated language based on the specific-difference theory was used by the experts. In such a theory a concept is defined from a previously defined concept by indicating its specific difference. The concept's names are arbitrary and can be normalized. Thus, for experts the concept <voltage relay> is not a kind of <relay> at the *same level* as the concept <on-off relay> or the <threshold relay> one. It is a kind of <threshold relay> whose threshold value is voltage. The final ontology, described below, does not match the ontology built from text. But it can be shared and reused between the different communities of practice.

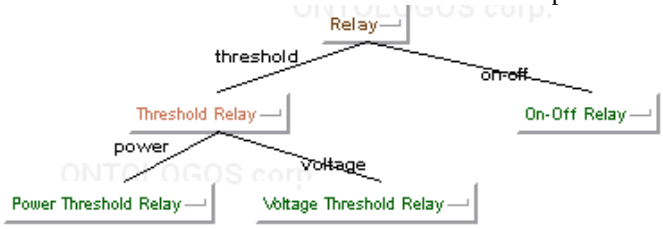

Fig. 4. A part of the formal ontology.

4. Saying is not modelling

Ontology acquisition from text relies on a set of strong hypotheses. The first of these is to say that experts can translate ontological knowledge within a corpus which more or less defines a closed world containing all the necessary information. The second hypothesis states that the reverse process is possible, based on the fact that a noun, or a noun phrase, denotes a concept and the hyponymy relationship is a linguistic

[9] By 'words of usage' we mean words which are used for writing texts.

translation of the subsumption one. The last hypothesis is to say that the conceptual structure matches the lexical structure and the former can be deduced from the latter. Validation by experts allows to bring up the conceptual structure as a domain ontology. But ontology built from text in general does not satisfy our expectations in terms of sharing, reusability, consensus and soundness. The problems we encounter are mainly due to the fact that these hypotheses are too strong and not always true, even if we use constrained languages for writing technical document. In reality, ontology building from text is the concern of textual linguistics. One of the latter's principles is the incompleteness of text. This implies that understanding text, and then understanding the meaning of terms, requires extra-linguistic knowledge which by definition is not included in the corpus. The 'closed world' hypothesis is not *stricto sensu* applicable. Furthermore text is written for a given goal according to a given writer's intention. All this information is not included in corpus: "the meaning (in general) of a sign needs to be explained in terms of what users of the sign do (or should) mean" [13].

Let us go back to the lexical structure (fig. 2) and the formal ontology (fig. 4). The noun phrase "voltage relay" really denotes a concept in the formal ontology. In fact it is a shortcut, a word of usage of the terminology, for the more *complete* expression "threshold relay whose threshold value is voltage". "Voltage relay" denotes the concept <voltage threshold relay> of the formal ontology whose name can be normalized as "voltage threshold relay". This example illustrates the fact that using rhetorical figures, like metonymy[10] or ellipse, and in particular synecdoche[11] is a very ordinary practice in writing technical documents. *Such rhetorical figures assume that both writers and readers share a same and pre-existent conceptualization of the world which is necessary for understanding meaning of term and text. This pre-existent knowledge – which is the domain ontology – is not included in texts and then can not be extracted from them.* Only some relationships are explicit; and if these relationships can always be justified in their context, they can be undesirable from the domain conceptualization point of view. This is the case for the conceptual structure built from the lexical one (fig. 3) where a <voltage relay> concept has been defined and linked by a 'is a' relationship with the <relay> concept when no relationship has been extracted with the <threshold relay> concept. These rhetorical figures refer to implicit knowledge, either concepts or relationships, which is not possible to know, except for experts. Let us precise that experts have validated the ontology built from text (fig. 3) for the same reasons they use synecdoche figures, in reference to the implicit knowledge of their domain. Natural language is not a suitable language for specifying conceptualization: it is not its aim; as well as the main goal of writing text is not to define ontology. Knowledge acquisition from text corresponds to a semasiologic approach where we first find terms (mainly nouns) and then define them in a given context. Such an approach is concerned with linguistics and more precisely with lexical semantics [14] whose main objective is word meaning[12]. It is then

[10] A metonymy is a figure of speech in which one word is substituted for another with which it is closely associated.

[11] A synecdoche is a figure of speech in which a part is used for the whole.

[12] Linguistics is mainly interested in the relationships between *signifier* and *signified* when ontology is mainly interested in the relationships between concept and object.

difficult to reuse and share such a contextual knowledge. Nevertheless, a lot of useful information can be extracted from text; especially if one considers than an ontology is also a vocabulary of terms[13] with their definition: "An [explicit] ontology may take a variety of forms, but necessarily it will include a vocabulary of terms and some specification of their meaning (i.e. definitions)." [15]. The main result of knowledge acquisition from text is a network of words of usage linked by linguistic relationships. There is no concept in text, and the lexical structure does not match with the domain conceptualization.

5. Formal languages

In theory, domain ontology represents task-independent, and then sharable and reusable, knowledge of a domain. A formal approach should allow to reach such a goal. As a matter of fact, natural language, even constrained in its syntax and semantics, cannot be used for concept definitions. We need a formal language for the definition of a conceptualization – such a "specification of a conceptualisation" is called an ontology [16] –. This is a useful means to avoid the issues raised by natural language and to reach agreement: if you accept the hypothetical and deductive approach of the formal system, you are obliged to accept its constructions, i.e. the domain conceptualization.

Nevertheless, the famous hypothesis of Sapir and Whorf [17], [18], concerning the interdependence of thought and language, is also applicable to formal languages. This means that the choice of the formal language for the definition of concepts is important. The way an ontology is built and the way a concept is defined directly depends on the formal language which is used; and the results will not be the same. Today formal languages are mainly logic-oriented. The concepts are represented as unary predicates when their attributes, or slots, are represented as binary predicates, also called roles. Description logic [19] is a good example of logic appropriated to knowledge representation. On the other hand, frame representation languages [20], in spite of the criticism of [21], are semi-formal and more human-readable languages. They allow to define concepts as a set of slots and organize them according to an inheritance and hierarchical relationship. OWL, the web ontology language [22], combines the advantages of these two approaches. It is a dedicated language for building ontology based on the W3C philosophy and description logic while providing a human readable formalism with the Protégé environment. The final ontology depends on the formal language which is used. So, the frame-oriented ontology of our relay example might be very similar to the conceptual model extracted from the lexical structure if we stay too close to text. On the other hand, the logic-oriented ontology of the relay, if it is different in its expression, does not solve our problem if we forget to explicitly express the relationship between <voltage relay> and <threshold relay>. In fact, knowledge representation requires a formal but also an epistemological oriented language. It means a language which can help the knowledge engineer to capture the nature of knowledge: for example a set is not a

[13] It is important to bear in mind that words of usage of LSP (Language for Special Purpose) and terms of terminology are not necessary the same.

concept, even if a concept can be interpreted as the set of its subsumed objects. As a matter of fact, logic is a neutral (or flat) language which cannot represent the different kinds of knowledge: a unary predicate can represent either a concept or a property while binary predicates can represent either attributes (internal relationships) or relationships between concepts (external relationships). Some interesting work has been done in order to introduce epistemological principles in logic, for example the 'ontological rigidity' constraint [23], [24]. But such principles do not really define guidelines for ontology building in the sense that they do not help the knowledge engineer to identify and structure concepts, they only constrained value of well formed formula.

6. The OK model

The OK (for Ontological Knowledge) language is an ontology-oriented language which relies on epistemological and formal principles [25]. It is based on the specific-difference theory. This theory considers a conceptualization as a system of concepts organized according to their differences more than factoring attributes shared by objects: a concept is defined from a previously existing one by adding a specific difference. The difference is then the main principle of the ontology building process on which identifying and structuring concepts rely on. Let us also remark that the agreement problem is reduced to the agreement on differences.

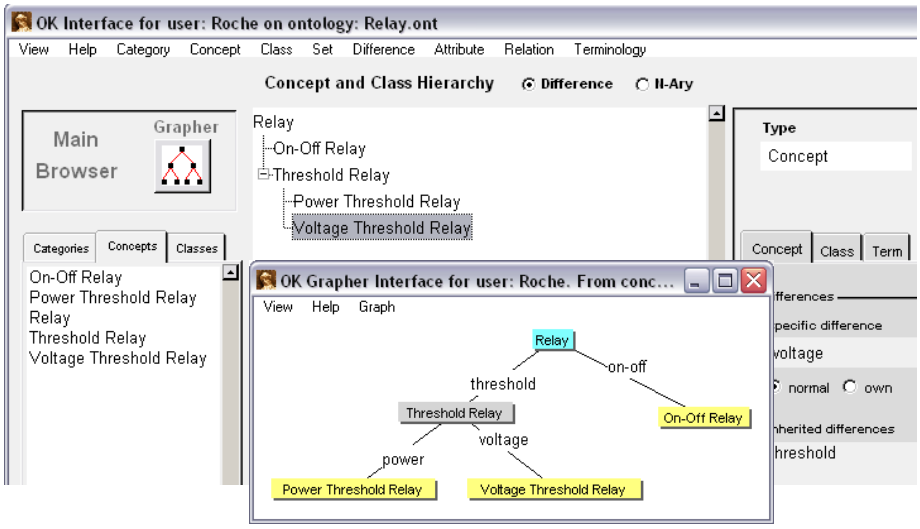

Fig. 4 – The OCW[14] environment for ontologies by specific differentiation.

Coming back to our example of 'relay', this approach reveals that the difference between the concept denoted by the linguistic expression "voltage relay" and the one denoted by "threshold relay" is the threshold value. There is no a <voltage relay>

[14] "Ontology Craft Workbench". Ontology written in OK can be translated in OWL [26].

concept (fig. 3) but a <voltage threshold relay> concept (fig. 6) whose linguistic expression "voltage relay" is one of its possible words of usage. Furthermore, in technical domains a lot of experts agree with the classical (Aristotelian) definition of a concept: a <voltage threshold relay> is a <threshold relay> whose threshold value is voltage, and then on the resulting conceptual structure.

7. Conclusion

Since we can consider that technical and scientific documents convey some domain knowledge, ontology building can rely on knowledge acquisition from texts. But such a conceptualization is corpus-dependent and does not offer the main properties we expect from ontology, e.g. reusability and soundness. Furthermore, ontology extracted from text in general does not match ontology defined by expert using a formal language.

The knowledge extracted from text is a linguistic knowledge. The lexical structure is a network of words of usage linked by linguistic relationships like hyponymy, synonymy, etc. This lexical structure is a linguistic "picture" of the domain ontology. A picture built in a given and particular context, for a given goal which reflects a particular linguistic use of the domain conceptualization. As a matter of fact, texts fall within language in action. Using rhetorical figures, such as metonymy and ellipse, is a very ordinary practice even for writing technical documents. Such figures of speech assume that both writers and readers share a same and pre-existent conceptualization of the world which is necessary for understanding meaning of words of usage and texts. This pre-existent knowledge – which is the domain ontology – is not included in texts and then can not be extracted from them. A conceptual model "directly" built from a lexical structure will probably not be reusable, because too corpus-dependent, and then probably not correct if we consider an ontology as a non-contingent knowledge. The lexical structure and the domain ontology do not match.

At last, if formal languages allow to define "a specification of a conceptualization" they nevertheless raise their own issues mainly due to their epistemological neutrality. Ontology design remains an epistemological issue which requires epistemological-oriented languages.

References

1. Staab, S., Studer, R.: Handbook on Ontologies. Steffen Staab (Editor), Rudi Studer (Editor), Springer 2004
2. Gomez-Perez, A., Corcho. O., Fernandez-Lopez, M.: Ontological Engineering : with examples from the areas of Knowledge Management, e-Commerce and the Semantic Web. Asuncion Gomez-Perez, Oscar Corcho, Mariano Fernandez-Lopez, Springer 2004
3. Roche, C.: Ontology: a Survey. 8th Symposium on Automated Systems Based on Human Skill and Knowledge IFAC, September 22-24 2003, Göteborg, Sweden
4. Buitelaar, P., Cimiano P., Magnini B.: Ontology Learning from Text: Methods, Evaluation and Applications (Frontiers in Artificial Intelligence and Applications, Vol. 123). P. Buitelaar (Editor) Ios Press Publication (July 1, 2005)

5. Dourgnon-Hanoune, A., Salaün, P., Roche, C.: Ontology for long-term knowledge. XIX IEA/AIE, Annecy 27-30 June 2006
6. Dourgnon-Hanoune, A., Mercier-Laurent, E., Roche, C.: How to value and transmit nuclear industry long term knowledge. ICEIS 2005, 7th International Conference on Enterprise Information Systems, Miami, 24-28 May 2005
7. Sinclair, J.: Corpus and Text - Basic Principles. In: Developing Linguistic Corpora: a Guide to Good Practice. Ed. M. Wynne. Oxford: Oxbow Books: 1-16. Available online from http://ahds.ac.uk/linguistic-corpora/ [Accessed 2007-04-11]
8. Aussenac-Gilles, N., Sörgel, D.: Text analysis for ontology and terminology engineering. Applied Ontology. n°1. pp. 35-46
9. Daille, B.: Recent Trends in Computational Terminology. Special issue of Terminology 10:1 (2004). Edited by Béatrice Daille, Kyo Kageura, Hiroshi Nakagawa and Lee-Feng Chien, Benjamins publishing company
10. Harris, Z.: Mathematical Structures of Language. 1968, reprint 1979. R.E. Krieger Publishing Company, Inc.
11. http://wordnet.princeton.edu/
12. Kiryakov, A., Popov, B., Terziev, I., Manov, D., Ognyanoff, D.: Semantic Annotation, Indexing, and Retrieval.Elsevier's Journal of Web Sematics, Vol. 2, Issue (1), 2005
13. Grice, H.P.: Meaning. Philosophical Review n°66. pp 377-88, 1957
14. Cruse, D.A.: Lexical Semantics. Cambridge University Press 1986
15. Ushold, M., Gruninger, M.: Ontologies: Principles, Methods and Applications. Knowledge Engineering Review, Vol. 11, n° 2, June 1996. Also available from AIAI as AIAI-TR-191
16. Gruber, T.: A Translation Approach to Portable Ontology Specifications. Knowledge Systems Laboratory September 1992 - Technical Report KSL 92-71 Revised April 1993. Appeared in Knowledge Acquisition, 5(2):199-220, 199
17. Sapir, E.: Language. An Introduction to the study of speech. Docer Publications, 2004. Originally published by Harcourt, Brace and Company, 1921)
18. Whorf, B.L.: Language, Thought and Reality. The MIT Press, 1956
19. Baader, F., Calvanese, D., McGuiness, D., Nardi, D., Patel-Schneider, P.: The Description Logic Handbook. Franz Baader, Diego Calvanese, Deborah McGuinness, Daniele Nardi, Peter Patel-Schneider, editors. Cambridge University Press, 2003
20. Wright, J. M., Fox, M.S., Adam, D.: 84 "SRL/1.5 Users Manual." Technical report;Robotics Institute, Carnegie-Mellon University 1984.
21. Woods 75. What's in a Link: Foundations for Semantic Networks. Representation and Understanding: Studies in Cognitive Science, 35-82, edited by D.G. Bobrow and A.M. Collins, New York: Academic Press, 1975.
22. OWL Web Ontology Language: http://www.w3.org/TR/owl-features/
23. Guarino, N., Carrara, M., Giaretta, P.: An Ontology of Meta-Level Categories. of Knowledge Representation and Reasoning: Proceedings of the Fourth International Conference (KR94), Morgan Kaufmann, San Mateo, CA.
24. Kaplan, A.: Towards a consistent logical framework for ontological analysis. FOIS'01. October 17-19, 2001, Ogunquit, USA.
25. Roche, C.: The 'Specific-Difference' Principle: a Methodology for Building Consensual and Coherent Ontologies. IC-AI'2001: Las Vegas, USA, June 25-28 2001
26. Spies, M., Roche, C.: Aristotelian ontologies and OWL modelling. Third International Workshop on Philosophy and Informatics. Saarbrücken, Germany - May 3-4, 2006