

Automatic Annotation of Discourse and Semantic Relations supplemented by Terminology Extraction for Domain Ontology Building and Information Retrieval

¹ Florence Le Priol, ¹Brahim Djioua, ²Daniela Garcia

¹ LaLIC, Université Paris-Sorbonne
28 rue Serpente - 75006 Paris – France
[flepriol, bdjioua]@paris4.sorbonne.fr

² EDF R&D
1 avenue du Général de Gaulle – 92141 Clamart Cedex – France
daniela.garcia@edf.fr

Abstract

In this article, we develop a framework for the building of domain ontologies and a semantic index based on two technologies: terminology extraction with LEXTER (© EDF R&D) and discourse and semantic annotation with EXCOM. We have selected two specific points of view for this study: causality and part-whole notions. In the first part of this paper, we explain the contributions of a terminology and the discursive and semantic relations for domain ontology building. In the second part we propose a semantic based index for information retrieval.

Introduction

In the traditional meaning, a term is considered as the linguistic label of a concept. The classical doctrine of terminology relies on an unifying view of knowledge, which assumes that knowledge is organized into domains, each domain being equivalent to a network of concepts.

Recently, several linguists in terminology have focused their attention on the notion of rich contexts which are involved in the detection of terms and properties of terms (Condamin & Rebeyrolles 98). Furthermore, several other linguists (Descles & al. 91) use textual contexts to identify textual segments with discourse and semantic relations between terms.

Domain Ontology Building and Information Retrieval are both essential elements for the Semantic Web. Ontology building using classical methods is expensive because a model must be created for each new domain. Information Retrieval is based on terms indexation without taking into account neither the semantics of the relations nor the context.

We propose in this paper a framework based on LEXTER, a terminology extraction tool, EXCOM, a discourse and semantic annotation engine with contextual exploration and MOCXE, an indexation engine of textual segments. This framework allows on the one hand domain ontology building, by using information contained in texts, and on

the other hand, information retrieval.

Framework Description

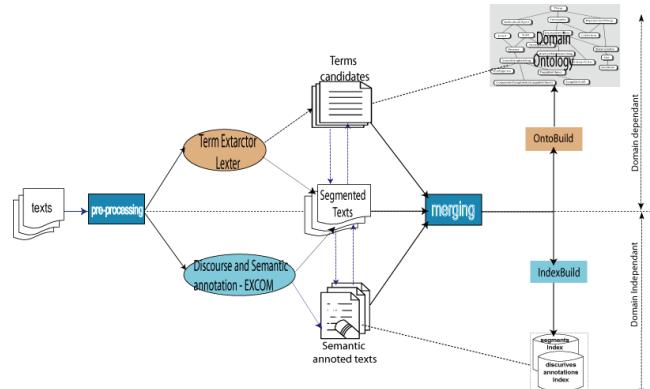


Fig. 1: framework

As shown in this architecture, texts are at the same time processed by LEXTER for terms that are candidates for the ontology building and by EXCOM for discourse and semantic annotation. In our paradigm, discourse means the linguistic discourse organization of a text. The two models mixed here deal with a structured text (title, section, para, sentence). Each structure is identified by a unique ID. The two models relate the candidate terms and the discourse and semantic annotations to these physical segments.

EXCOM uses linguistic rules for identifying segments by using different strategies defined through rules. Some of these rules use lists of simple patterns coded as regular expressions, others need to identify structures like titles, sections, paragraphs and sentences for extraction purposes. The most relevant rules for EXCOM are those called Contextual Exploration (CE) rules. A CE rule is a complex algorithm based on a prime textual mark (called indicator), and secondary contextual clues intended to confirm or invalidate the semantic value carried by the indicator.

The next sections describe the two models for the terminology extraction and the discourse and semantic

annotation.

Terminology

The terminology extraction is based on LEXTER. The acquisition and the exploitation of a structured lexicon are carried out automatically by the LEXTER system, developed at EDF R&D in the framework of a PhD research project (Bourigault 94). LEXTER was designed to extract noun phrases (NPs) from a corpus of texts (in French).

The candidate terms are NPs that are not extracted by direct pattern matching, but they are isolated by spotting their syntactic boundaries, like, for instance, verbs. Extracted NPs are filtered and then automatically organized in a structured network of head-expansion relations. This phase is based on the hypothesis that eligible NPs must exhibit the syntactic pattern of candidate terms, as established by a terminology theory. For example: definite article + noun + preposition + noun is an observed candidate term pattern.

LEXTER accounts for morphological variants and head-modifier relations of nouns and NPs, that are grouped into families. It also supplies simple distributional figures, such as frequency of a candidate term in the corpus or candidate term head-modifier productivity within the structured network..

LEXTER also stores the whole corpus divided into paragraphs, along with a pointer to the location of each candidate term in the text. This feature was initially designed to supply the terminologist with a linguistic context for validation.

EDF R&D and Temis (www.temis.com) decided to collaborate and build a new generation of terminology extraction tools, based on the LEXTER prototype and the Temis extraction technology namely Insight Discovery Extractor (IDE). The name of the extraction solution is EXTER. Having worked on a project for a car manufacturer on a corpus of 3 million words in French, the EXTER solution provided an extremely good level of quality and relevancy of the proposed terms, according to the users with output 50 000 term candidates instead of the expected 300 000. As a result, not only is the quality of the proposed term higher than the previous Temis solution, but also the validation and the cleaning has been divided by a factor of 6. The current version supports French and English, with extensions in German, Spanish and Italian.

Discourse and Semantic Annotation

We already presented the EXCOM system (Djioua & al. 06) for discourse and semantic annotation. The hypothesis is to reproduce “what makes naturally a human reader” who underlines certain segments related to a particular point of view which focuses his attention. There are several points of view for text mining, corresponding to various focusing on more specific research information.

For this paper, we are interested in two particular points of view which are the causality and the part-whole notion. The methodology used by EXCOM, called Contextual Exploration (Desclés & al. 91, Desclés 06), describes the discursive organization of texts exclusively using linguistic knowledge present in the textual context. Linguistic knowledge is structured in the form of lists of linguistic marks and declarative rules for the Contextual Exploration for each linguistic mark. The constitution of this linguistic knowledge is independent of a particular domain. Domain knowledge describes the concepts and sub-concepts of a subject domain with their relationships. The contextual knowledge concerns communicative knowledge as a discursive organization, which deals with the preferences and needs of those who use information in the texts. Linguistic rules define different strategies for identifying and semantically annotating textual segments. Some of these rules use lists of simple patterns coded as regular expressions, others need to identify structures like titles, sections, paragraphs and sentences for extraction purposes.

Relations

Semantic relators, whether static, kinematic, or dynamic, are based on the cognitive representations built either by the perceptions of space, stability, and temporal change, or by action. The first categorization takes place around the opposition static / kinematic-dynamic (Desclés 90). In this paper, we take as an example the part-whole and causality relations.

Part-Whole

In French, static relations (identification, localization, part-whole, attribution, inclusion...) are narrowly attached to the linguistic expression of primitives "est" and "a" (Desclés 87). Static relations are binary relations which make it possible to describe states (static situations) in the expert domain. The semantics of each static relation corresponds to intrinsic properties: functional type (standard semantics of relation arguments); algebraic properties (reflexivity, symmetry, transitivity, ...); properties of fitting (combination) with other relations in the same context (i.e. in a given static situation).

The Part-Whole relation allows the decomposition of the object to its components. This relation is transitive and reflexive but non symmetrical.

It is expressed by "*est une partie de*" or in statements like

La France fait partie de la Communauté Européenne (France belongs to the European Community)

Le fluor entre dans la composition des dents (Fluorine uses the composition of teeth bones).

In the first example, the linguistic mark is "fait partie de" (belongs to) and there is no contextual clue (Ringfnc04 CE rule).

The second example illustrate the Ringrdvrs10 CE rule: "entre dans" is an element of a list of marks (indicateur);

“composition” (element of the list “lingaux7”) and “des”, “element of the list “lingprep2”) are clues (indice) and are in the right context.

```
<regle nom_regle="Ringrdvrs10"
tache="Relation_Static"
point_de_vue="partie_tout" type="EC">
<conditions>
<indicateur espace_de_recherche="phrase"
type="liste" valeur="entrerdans" />
<indice contexte="droit"
espace_de_recherche="identique" type="liste"
valeur="lingaux7" />
<indice contexte="droit"
espace_de_recherche="identique" type="liste"
valeur="lingprep2" />
</conditions>
<actions>
<annotation type="ajout_attribut"
espace="identique" annotation="partie_de" />
</actions>
</regle>
```

There are 29 rules that organize the linguistic knowledge of part-whole relation.

Causality

There is no particular lexicon or a particular grammatical category in French that expresses the relation of causality. Our approach is based on the identification of certain privileged linguistic units (causality marks), in particular certain verbs, which direct the semantic interpretation of the analyzed sentence towards a causal semantic value. We studied the causal relations present in technical texts among which we could distinguish two separate families of relations, each one having its own characteristics:

- Formal causal relations

Le niveau de la puissance produite par l'usine varie en fonction de l'hydraulicité
(the level of the power produced by the factory varies according to the hydraulicity)

- Efficient causal relations

Cette opération facilite l'implantation des futurs ouvrages
(This operation facilitates the establishment of the future works).

In the second case, the cause produces an effect which is distinct from it and the relation has an orientation in time. The indicating verbs of the concept of causality, that translate the efficient relations, can in their turn

- specify the nature of the produced effect , for example a disturbance:

La mise en parallèle de différents types d'ouvrages entrave le fonctionnement du réseau
(The parallelization of various types of works blocks the operation of the network)

- specify the nature of the causal action, for example the participation of a cause in the production of the effect:

La longueur maximale du départ le plus long ramené à moins de 90 km contribue à améliorer la qualité de service (The maximum length of the longest departure brought back to less than 90 km contributes to improve the quality of service).

The system exploits the semantics of 300 French verbs organized in 25 classes which specify the semantics of their causal value.

OntoBuild

(Gruber 93) defines an ontology as the specification of a conceptualization of knowledge domain.

We prefer another definition, more operationally formulated as follows: an ontology is a semantic network that contains a number of concepts necessary for the complete description of a domain. These concepts are subject to hierachic or semantic relations.

We propose to use textual segments annotated as discursive and semantic relations and extracted terms in order to build ontologies:

- i. The text is automatically annotated by EXCOM with discursive and semantic relations,
- ii. Terms are automatically extracted by LEXTER, from the annotated textual segments,
- iii. The triplet argument - relation - argument is built by using each annotated segment and its extracted terms.

Let us take as example the “Guide de Planification” (EDF R&D) and annotate it with causality and part-whole relations.

Textual segments are annotated as a causality relation or as a part-whole relation.

```
<?xml version="1.0" encoding="UTF-8" ?>
- <!-- Segmentation générée par SegaTex V0.8
le [Mon Nov 13 10:26:38 2006] LaLIC Copyright
-->
<article lang="fr">
<title ID="1">Guide de Planification</title>
<author>EDF R&D</author>
<section ID="1">
...
<para ID="497">
<phrase ID="650"
annotation="lalic.excom.partie_tout.partie_de
">En amont de la planification des réseaux
régionaux le calcul économique permet de
choisir les matériels et en particulier les
éléments suivants : sections des lignes,
niveaux de tension, puissances nominales des
nouveaux types de transformateurs.
</phrase>
</para>
...
<para ID="1778">
```

```

...
<phrase ID="2328"
annotation="lallic.excom.causalité.creation">
Dans ce cas des dispositions particulières
sont prises pour la mise à la terre, mais
elles entraînent un surcoût non
négligeable.</phrase>
</para>
...
</section>
</article>

```

From these textual segments, terms are extracted.

```

<?xml version="1.0" encoding="ISO-8859-1"
standalone="no"?>
<!DOCTYPE TERMINO SYSTEM "wtXML.dtd">
<TERMINO>
...
<TERME id="T11361" libelle="matériels de
réseaux" freq="1" cat="SN" statut="non
validé" >
<HYPERLIEN id="H55322" refctxt="GdPlanif-1-
2-4-3_p3-000000CE">
</HYPERLIEN>
<PROP refprop="lemme">matériel de
réseau</PROP>
<PROP refprop="prod">0</PROP>
<PROP refprop="freqisol">1</PROP>
<PROP refprop="frequnonisol">0</PROP>
<PROP refprop="effectvar">2</PROP>
<PROP refprop="representvar">non</PROP>
</TERME>
...
<LIEN id="L20480" refsref="T7244"
refcib="T7243" reflien="T"/>
<LIEN id="L12462" refsref="T16709"
refcib="T10029" reflien="E"/>
<LIEN id="L20481" refsref="T7246"
refcib="T7245" reflien="T"/>
</TERMINO>

```

This knowledge is structured in a graph (Fig. 2).

Fig. 2: Part of the graph of “Guide de Planification” (EDF R&D)

The causality and part-whole relations appear with labeled edges. The not labeled edges come from the terminology extracted by LEXTER.

The generalization of the graph makes it possible to obtain the domain ontology.

IndexBuild

The general process of indexing use a multilevel structure composed by textual segments (such as titles, sentences, paragraphs, sections, ...) and discourse and semantic annotations (such as causality and part-whole relations). On the other hand terms and their terminology, extracted from the same texts, are associated to the physical structure. This organization explicits the relationship between the initial document, the constitutive textual segments (sentence, paragraph and section title), the discursive and semantic annotations, and the extracted terms that compose the textual segments . A document is considered, in this model, as a set of textual segments identified automatically by EXCOM together with its discursive and semantic organizations expressed by the author. Each textual segment is associated with several important pieces of information, namely:

- i. a set of semantic annotations (discursive marks such as *causality* and *part-whole notion*). Each textual segment identifies as relevant for a document can be associated to a set of discursive and semantic annotations according to the semantic points of view used in the annotation engine EXCOM. Then, the same textual segment can be chosen by the search engine MOCXE as an answer to a query about *causality* and *part-whole notion*.
- ii. the document's URI that ensures its unique identification on the Internet.
- iii. the document's title in order to have a better readability of the answer.
- iv. the terminology associated to the terms founded in the full-text content for a relevant answer to users. The procedure used for the indexing of textual segments is the same as the methodology used by (Salton & al. 75).

The indexing engine MOCXE generates an index of textual documents by using the output of the annotation engine EXCOM. The index gives the possibility for information retrieval according to discourse and semantic relations (part-whole and causality). The queries are formulated by using semantic relations that are defined inside a point of view for text mining (test are yet to be performed for the causality relation); the answers are given in the form of “annotated textual segments”. The first realization for the relation of connection between named-entities has been created in November 2005 and presented for France Telecom R&D (www.excom.fr). For indexing process we use techniques already available in open-source software for the search engine Lucene/Nutch architectures (www.apache.org).

The majority of existing automatic indexing methods select index terms from the text of the document. The index terms selected in the way constitute single words or multi-word phrases extracted from titles and the full text of each documents is assumed to represent its contents. Presently, search engines operating on the Internet index the documents on the basis of this principle. However, not every word from a text constitutes an appropriate index term and on the other hand, different index terms associated to a text do not have equal contributions to the contents of this text. A number of techniques could help in identifying and weighting the content terms. The objective of the MOCXE system is to build up a multiple index composed of textual segments (sentences, paragraphs, section titles...), semantic discursive annotations (relations between concepts, causality...) and, as an experiment, a terminology.

In order to illustrate the way in which a term – with its terminology - can be used in an information retrieval system, let's consider queries about “causalities” related to the French term “surcoût” (overcharge).

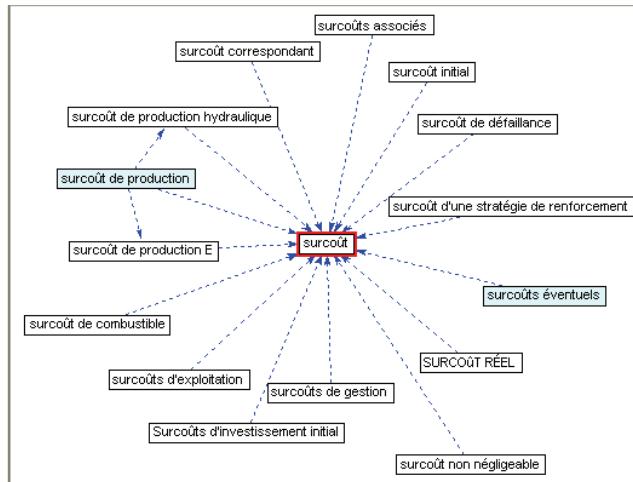


Fig. 3: Terminology of “surcoût” (overcharge) (visualization in Worldtrek – © EDF R&D).

This terminology given by LEXTER organize the terms (Fig. 3) as a specification of terms like “surcoût initial” (initial overcharge), “surcoûts de gestion” (overcharge of management) and “surcoût de production hydraulique” (overcharge of hydraulic production).

Basically, terms can be regarded as a particular type of lexical data for this application. Contrary to general language lexicons, term bases here contain multi-word related units.

We proceed with a terminology indexation similar to that used in dependency-based approaches to NLP, such as the SPIRIT system (Andreevsky & al. 77).

This approach uses the Tesnière's structural dependency model in which words are included in a lattice of dependency relations. These relations are associated to the textual segments indexed with the discursive and semantic annotations. By the technique that establishes a

relationship between segments and semantic annotations (causality and part-whole) and through the terminological relationship between “surcoût” (overcharge) and “surcoût de production hydraulique” (overcharge of hydraulic production), the utterance “*Dans certains postes anciens, l'installation de transformateur de 170 MVA entraîne un surcoût parfois très élevé qui peut éventuellement conduire à changer de stratégie.*” is semantically (through the discourse relation of causality) related to the utterance “*Dans le cas des usines de lac, il existe également un surcoût de production hydraulique résultant d'un placement imposé dans le but de lever ou de diminuer les contraintes.*”

This figure shows a screenshot of the MOCXE search interface. The search bar at the top contains the query "Trouver ‘surcoût’ pour la relation discursive de causalité". Below the search bar, there are several filter and search options:

- MOCXE : un moteur de recherche d'informations sémantiques**
- Recherche d'informations sémantiques** (MOCXE 16.1) Copyright - LalICC
- Catégorie**: Thématique Recherche Causalité Définition
- Caractéristique**: Qualitative Analytique Fonctionnelle Synthétique surcoût
- Termes à rechercher**: surcoût
- chercher**

The results pane shows a list of search results, each with a snippet of text and a link to the full document. One result is highlighted in yellow:

[1] Guide de planification
Dans le cas des usines de lac, il existe également un surcoût de production hydraulique résultant d'un placement imposé dans le but de lever ou de diminuer les contraintes [annotation]...
Dans ce cas des usines, l'installation de transformateurs pour la mise à la terre, dans elles entraîne un **surcoût non négligeable** [annotation]...
Donc dans ces usines, l'installation d'un transformateur de 170 MVA entraîne un **surcoût** parfois très élevé qui peut éventuellement conduire à changer de stratégie [annotation]...
Le modèle PECO (voir schema 3) permet, d'évaluer pour un point horaire donné, le surcoût de production du réseau et le coût de l'énergie non distribuée [annotation]...
Dans certains postes anciens, l'installation de transformateurs pour la mise à la terre, dans elles entraîne un **surcoût non négligeable** [annotation]...
Le **surcoût d'une stratégie de renforcement** 225 kV ou 110 kV est inférieur de 5 M€ au coût de la stratégie de création d'une source 400 kV, en tenant compte de l'augmentation du nombre de composants (cellules, lignes,...) des axes
Cette technique présente l'inconvénient d'engendrer un **surcoût initial**, par contre elle améliore considérablement la création de la deuxième liaison que ne demande pas de procédures administratives et qui peut donc être réalisée dans un délai très court (5 J) à bas coût de compensation [annotation]...
Si le **surcoût d'une stratégie de renforcement** 225 kV ou 110 kV est inférieur à 5 M€ au coût de la stratégie de création d'une source 400 kV [annotation]...
Cet allégement résulte essentiellement : - de la dérivation éventuelle de la sécurité du système de transport que peut apporter l'augmentation du nombre de composants du réseau et de l'augmentation de la valeur des transferts ; - des économies de gestion éventuelles du réseau au 400 kV [annotation]...
En ce qui concerne la haute tension, si l'enroulement est techniquement adapté, les **surcoûts associés** initient notamment la baisse de la valeur des transferts, des économies de gestion éventuelles du réseau au 400 kV [annotation]...
L'optimisation des stratégies dépend de la valeur d'un certain nombre de paramètres tels que le taux de croissance des charges, le niveau de puissance pouvant être fourni en secours par les postes Ht voisins et finis les **surcoûts** éventuels d'assainissement de certains appareils dépassant de certains seuils [annotation]...
La fréquence du dépassement des **surcoûts** de gestion éventuels peut être évaluée par le modèle ANSEC et les **surcoûts** de gestion éventuels sont obtenus par les modèles MECCO et MEDDA (voir article "Les modèles utilisés par la DPT" - chapitre 5) [annotation]...]

Laboratoire LalICC - Indexation sémantique | Crédits | MOCXE 16.1, Copyright ©2005, Laboratoire LalICC

Fig. 4: example of a query

Example of a query (Fig. 4): “Find the causalities related to the term ‘surcoût’”. The system treats two incoming related informations:

1. the discourse relation of causality (find causalities) – the system answers by finding in the structured index documents which contain textual segments annotated by the discourse information causality;
2. the terminology of the term “surcoût” - the system answers by selecting from the textual segments, the ones that contain the term “surcoût” and all the terms organized by the terminology as shown below.

Tests

In this paper, we presented examples from the “Guide de Planification” (EDF R&D), annotated with the causality and part-whole relations. This corpus contains the tasks planning of EDF (French electricity company).

The corpus characteristics are as follows :

| Pages | Paragraphs | Sentences |
|-------|------------|-----------|
| 262 | 5479 | 7245 |

Processing by EXCOM and LEXTER gives the following results :

| Causality relations | Part-Whole relations | Candidate terms |
|---------------------|----------------------|-----------------|
| 434 | 362 | 20469 |

Processing by MOCXE with the query “*Find the causalities related to the term 'surcoût'*” gives 10 textual segments.

Conclusion

In this paper, we have presented a framework for the building of domain ontologies and a semantic index used to information retrieval based on two technologies: terminology extraction with LEXTER and discourse and semantic annotation with EXCOM. We have selected two specific points of view for this study: causality and part-whole notions and presented examples from the “Guide de Planification” (EDF R&D).

This prototype is operational.

We must now include the other points of view (inclusion, membership, localization, identification...) and evaluate the results on large volumes.

Acknowledgment

We thank Henry Boccon-Gibod and Vincent Godefroy (EDF R&D) to have allowed us to use, in this data processing sequence, various tools which they developed.

References

Abbas Y. Picard M.-L 2000, With WORLDTREK Family, create, Update and Browse your Terminological World, in *2nd International Conference Resources & Evaluation*

Andreewsky, A., Debili, F., Fluhr, C. 1977. “Computational learning of semantic lexical realtions for the generation and automatic analysis of content”. Proceedings, IFIP Congress, Toronto

Bourgault Didier 1994, *LEXTER un Logiciel d'Extraction de TERminologie. Application à l'extraction des connaissances à partir de textes.*, Ph. D., EHESS, Paris.

Chantrier C. 2005, EXTER: A Breakthrough Solution for Efficient Terminology Extraction, TEMIS SA, France, in *Translating and the Computer 27*, New Connaught Rooms, London

Condamin A., Rebeyrolles J. 1998, CTKB : A Corpus-based Approach for Terminological Knowledge Base ». In *Proceedings of the first workshop on computational terminology (COMPUTERM'98)*, Workshop of Coling'98. Montréal.

Desclés Jean-Pierre 1987, Réseaux sémantiques : la nature logique et linguistique des relateurs, in *Langages, Sémantiques et Intelligence Artificielle*, 87:55-78

Desclés J-P., Jouis C., Oh H-G., Reppert D. 1991, Exploration contextuelle et sémantique : un système expert qui trouve les valeurs sémantiques des temps de l'indicatif dans un texte., in *Knowledge modeling and expertise transfert*, Eds D. Hérin-Aime, R. Dieng, J-P. Regourd, J-P. Angoujard, 371-400. Calif : IOS Press

Desclés Jean-Pierre 2006, Contextual Exploration Process for Discourse and Automatic Annotation, *FLAIRS 2006*, Floride, 11-13 Mai

Djioua B., Garcia Flores J., Blais A. Desclés J-P., Guibert G. Jackiewicz A., Le Priol F., Nait-Baha L., Sauzay B., 2006, EXCOM : an automatic annotation engine for semantic information, *FLAIRS 2006*, Floride, 11-13 Mai

Garcia Daniela 1997. "COATIS, an NLP system to locate expressions of actions connected by causality links". In *Proceedings of the 10th European Workshop on Knowledge Acquisition, Modeling, and Management (EKAW'97)*. In *Springer's Lecture Notes in Artificial Intelligence*. Sant-Feliu-de-Guixols, Catalogne, 15-18 octobre, 1997.

Garcia Daniela 1998, *Analyse automatique des textes pour l'organisation causal des actions, système COATIS*, Ph. D., Univ. Paris-Sorbonne

Gruber Thomas R., 1993, Towards principles for the design of ontologies used for knowledge sharing, in *Formal ontology in conceptual analysis and knowledge representation*, Khewer Academic Publishers

Le Priol Florence 2000, Extraction et capitalisation automatiques de connaissances à partir de documents textuels. Seek-Java : identification et interprétation de relations entre concepts, Ph. D., Univ. Paris-Sorbonne

Le Priol F., Blais A., Desclés J-P., Djioua B., Garcia-Flores J., Guibert G., Jackiewicz A., Nait-Baha L. and Sauzay B., 2006, Automatic annotation of localization and identification relations in platform EXCOM, *FLAIRS 2006*, Floride, 11-13 Mai

Salton, G., Yang, C.S., and Wong, A. 1975. A vector-space model for automatic indexing – Communications of the ACM, 18(11):613-620.