Automatic annotation of localization and identification relations in platform EXCOM

Le Priol F.[1], Blais A. [1], Desclés JP. [1], Djioua B. [1], Garcia-Flores J. [1], Guibert G. [2], Jackiewicz A. [1], Nait-Baha L. [1], Sauzay B. [2]

[1] LaLICC, UMR8139 Université Paris-Sorbonne/CNRS
28 rue Serpente, 75006 Paris, France
[flepriol, ablais, jpdescle, bdjioua, jgflores, ajackiewi, lnaitbah]@paris4.sorbonne.fr
[2] France Télécom, Division Recherche & Développement
2 avenue Pierre Marzin, 22307 Lannion cedex, France
[gaell.guibert, benoit.sauzay]@francetelecom.com

Abstract

Semantic annotation of localization and identification relations falls under an immense project of automatic annotation of relations embodied in the platform EXCOM. While localization and identification relations have been defined by Applicative and Cognitive Grammar, they are described here from the perspective of language processing, based on contextual exploration method, with the goal to detect these relations automatically in the text, to annotate it correspondingly and provide several representations of the annotations in textual or graphic form.

Introduction

Semantic annotation of localization and identification relations falls under an immense project of automatic annotation of relations in corpus: platform EXCOM ("EXploration COntextuelle Multilingue") (Djioua & al. 2006). From a search engine based on contextual exploration, one can, starting from EXCOM, annotate relations according to various points of view: location relation of localization, identification, whole-and-part...; relation of the type ' connection ', i.e. ' which is with which' or in other words, relations of collaboration, meeting, proximity...; filtering of information like the quotations, definitions; summary of texts... Applicative and Cognitive Grammar holds as a principle that a static relation (identification, classification, localization, whole-and-part, size, attribute,...) exists when two entities are situated, relative to each other, in an open temporal interval. These entities are connected by relators, in particular the copula "*être*" (Desclés 1987).

We describe static relations from the perspective of language processing, deploying the method of contextual exploration in order to detect these relations automatically and to annotate the texts in graphic or textual form.

Contextual exploration

Annotation of semantic relations in EXCOM platform is based on contextual exploration.

Contextual exploration method (Desclés & al. 91, Desclés 97) makes it possible to process the text with respect to the contextual indexes internal to the text, and can result in raising semantic indeterminations or making certain decisions in sense construction. It leads to a data-processing implementation: contextual exploration systems. These systems call upon knowledge exclusively linguistic and present in the texts.

Linguistic knowledge is structured in form of lists and is capitalized in a knowledge base. There are two kinds of lists: indicators lists on the one hand, contextual index lists on the other hand. Indicators are specific to a given task (such as for example: to recognize an aspectual value, to filter an important sentence of the text, to extract a semantic relation...). Each indicator is seen as associating a set of heuristic rules of contextual exploration; the application of a rule, called by an indicator, amounts seeking explicitly, in the indicator context, the linguistic indexes complementary to the indicator, in order to be able to solve the task.

Our hypothesis is as follows: texts contain specific linguistic units that are relevant indicators to solve a particular task. However, the identification of these indicators is not sufficient. Analysis of a linguistic unit identified in a context calls necessarily upon other complementary linguistic indices that must be Co-present in the context; these indexes contribute directly to task resolution.

Consequently, for a given task, it consists of:

1. Identifying the relevant linguistic units explicitly
2. Finding a procedure that would explore the context by seeking certain relevant linguistic

indexes Co-present in the context in order to progress in task resolution.

Contextual exploration does not require knowledge of the treated domain. Indeed, the constitution of indicators lists and contextual indexes is independent of the domain; we do not need representations of knowledge preliminary to text analysis. The richness of a contextual exploration system thus depends on the richness of the lists of indicators, contextual indexes and general information of contextual exploration rules.

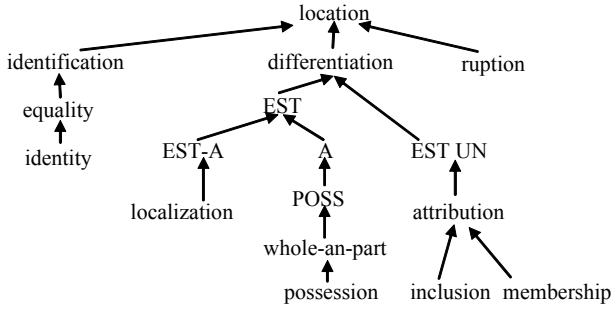
Semantic relations in Applicative and Cognitive Grammar

Semantic relators, whether static, kinematic, or dynamic, are based partly on the cognitive representations built either by perceptions of space, stability, and temporal change, or by action. The first categorization takes place around the opposition static / kinematic-dynamic (Abraham 95, Desclés & al. 98).

Cognitif level	
Dynamic relations	FAIRE, CONTR, TELEO…
Kinématic relations	MOUVT, CHANG, CONSV…
Statiques relations	identification, inclusion, membership, localization…

(Le Priol 2000)

In French, static relations 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 function of the archirelator of location is to establish a bond of location between a located entity and a locating entity; it is a general statement of relation. Identification, localization, whole-and-part, attribution, and ruption are its specifications.

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 the other relations in the same context (i.e. in a given static situation).

Localization annotation

The localization relator expresses "X is localised with respect to Y"("*Paris est en France*"). X (localised) is localised with respect to Y (the locator). Localization relation is directed from the localised one towards the locator: *X* → *localisation* → *Y*. Localization relations are of type1: FxFLH where X is of type J or type L, according to the context of the localised.

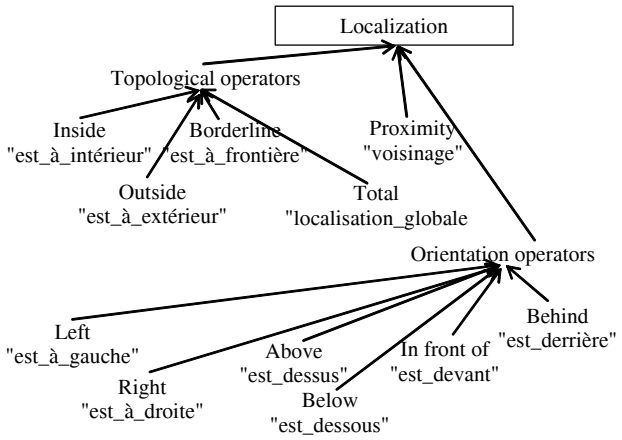
Topological operators allow the following relations to be specified:

- Relation "**est_à_intérieur**" (IN) is transitive, antisymmetric and non reflexive;
- Relation "**est_à_extérieur**" (*EX*) is irreflexive,
- Relation "**est_à_la_frontière**" (FR) is incompatible with outside and inside and more precise than closure
- Relation "**localisation_globale**" (FE) is incompatible with outside and redundant with borderline and inside.

Relation "**est_au_voisinage**" (VG) is not described by the traditional topological operators but its behavior approximates them.

Orientation operators allow specifying the following relations:

- Relation "**est_à_gauche**" is characterized by a side direction and a perpentacularity with the general orientation;
- Relation "**est_à_droite**" is characterized by the same elements as relation "est-à-gauche";
- Relation "**est_au_dessus**" is characterized by an orientation generated by gravity;
- Relation "**est_au_dessous**" is characterized by the same elements as relation "est-au-dessus";
- Relation "**est_devant**" is described by the operator ORI that assigns an orientation to the entity to which operator LOR applies;
- Relation "**est_derrière**" is described by symmetry with "est_devant".



In the case of semantic relations of localization, we chose, in the majority of the cases, the preposition as indicator because they are, at the cognitive level, more relevant to locate a relation of localization than the copula "is" or verbs expressing localization. Preposition (except for the verbs and the past participles that do not need any) will determine the annotation to allocate to the relation satisfying all the conditions of the rule.
Let us take the example of the group of rule which locates following constructions:

- a1 *être* + PREP a2 ("Luc *est* **dans** la cuisine")
- a1 (*être* +) ParticipePasseLoc + PREP a2 ("Luc est *allongé* **dans** le lit" ou "Luc, *allongé* **dans** le lit,…")
- a1 IverbeLoc1 + PREP a2 ("Luc *campe* **dans** la nature")

The orientation of arguments goes from the left towards the right.
The rule is declined in the same way for all the types of localization. The only changes will occur in the indicator class and also in the annotation allocated correspondingly.
The rule which, for this case, will annotate interior localization (est_à_intérieur) is presented below (Provôt 2005).

Rule name	RlocIN1a
Task	Relation_statique
Indicator	&IntroLieuxIN
Body	# Research space E1 := créer_espace(Gauche(I)) E2 := créer_espace(Gauche(x1)) E3 := créer_espace(Droite(I)) # Indices lists L1 := &IetrePresent L2 := &IetreImparfait L3 := &ParticipePasseLoc L4 := &IverbeLoc1 L5 := &Se L6 := &Iavoir L7 := &Capostrophe L8 := &IetreSaufPresentImparfait L9 := &ContreExLocIN
Conditions	Il_existe_un_indice x1 appartenant_à_E1 tel_que classe_de x1 appartient_à L1 ou L2 ou L3 ou L4 Il_n'existe_pas_d'indice x2 appartenant_à_E2 tel_que classe_de x2 appartient_à L5 ou L6 ou L7 ou L8 Il_n'existe_pas_d'indice x6 appartenant_à_E3 tel_que classe_de x6 appartient_à L9
Actions	attribuerEtSem(PhraseParent_de I, "est_à_intérieur")

In this case, one must find an index which is in a form of the verb "to be" at present (L1) or imperfect (L2), or a past participle (L3) or a verb (L4). This index must precede preposition (indicator).
L5 list comprises the pronouns or double pronouns (to avoid confusing the pronoun with the subject in the case of "nous" and "vous") {me, te, se, s', nous nous, vous vous} and is used for restriction on pronominal form. Pronoun should not be on the left of localization verb, verb "*être*" or past participle (indices x1).
L6 list comprises the conjugated forms of the verb "*avoir*" (restriction is of course useless if index x1 is a verb of &IverbeLoc1, but it does not harm in this case).
L7 list comprises the form "*c'*", in order not to locate constructions of highlighting (which are treated by another rule).
L8 list comprises the conjugated forms of "*être*" other than present and imperfect (composed past, preterit, pluperfect, future, conditional and subjunctive) when they are not in front of an eventual past participle.
L9 list comprises the lexemes of counterexamples of localization for the relation of interior.

For localization annotation, we have 192 rules and 65 classes of markers gathering 1001 indicators or indexes.

Let us launch the annotation of localization relations on an article of three pages drawn from Le Monde of November 20, 2002, with the headline "*Recep Erdogan, en visite à Athènes, affirme la nécessité d'une solution négociée sur Chypre*". User specifies that he wishes more particularly information concerning the terms: *garde d'honneur, aéroport d'Athènes, Erdogan; Europe, Madrid*.

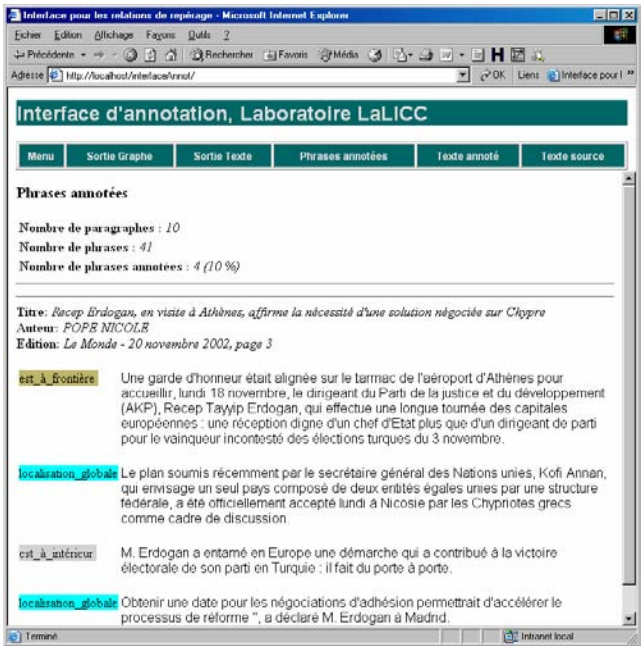
In the interface suggested, we choose the text which was segmented beforehand in sections, paragraphs and sentences by an in-house developed segmentor. We specify that we wish to annotate localization relations and we indicate certain number of terms which interest us particularly.

Clicking on the button "Valider" confirms the selections and initiates the annotation process by starting up the annotation engine of EXCOM platform. The results can be retrieved in various formats:

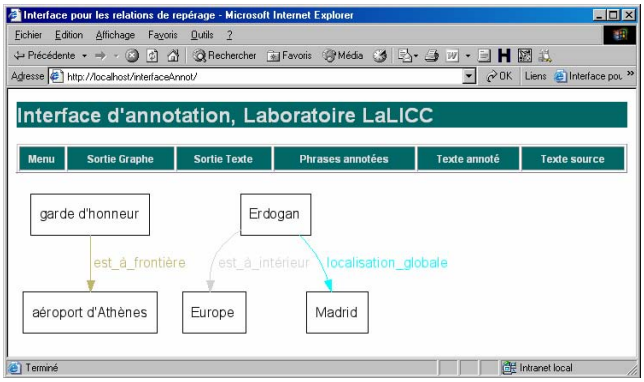
- The source text without any annotation (button "Texte source");
- Annotated text (button "Texte annoté") which has the entirety of the text in which the annotated sentences appear marked;
- Annotated sentences (button "Phrases annotées"), that presents only the sentences annotated with indication of the annotation;
- triplets argument-annotation-argument (button "Sortie Texte") which presents, when it is possible, by using the parameters entered by user at launching time, sentences annotated in the form

of a triplet giving only arguments of the semantic relation;

- The graph of semantic relations (button "Sortie Graphe") which presents, when it is possible, by using the criteria entered by the user at launching time, sentences annotated in the form of a directed graph.



Annotated sentences extracted the text



Graphs using the terms of the user

Navigation between the various forms of display is available to the user that permits, starting from the graph found in the annotated text or starting from an annotated sentence, to see the graph or to return to the original text.

Identification annotation

Identification relation is expressed as "X is identified with Y", i.e.: entity Y is used as identifier for entity X. It is used in statements such as: "*Paris est la capitale de la France*". Identification is defined by equality and identity, indeed, A=A and A=B do not have the same value of knowledge.

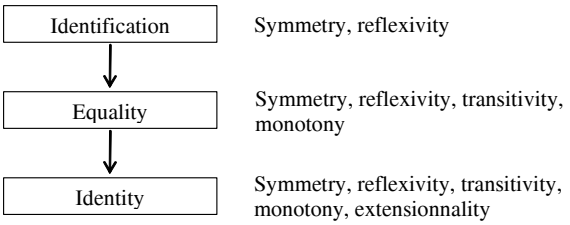
The concept of "egality" deployed within the meaning of identity (i.e. A=B) is interpreted by "A is the same thing as B" or "A and B coincides" raises a problem.

To say that "*l'étoile du matin est Vénus*" expresses the identity (or the equality) of two expressions appears too rudimentary analysis because this is a matter of identification and not of an identity: an object designating "*l'étoile du matin*" is identified as an object designating "*Vénus*" , this identification is performed inside of the language, independently of external reality or any knowledge stored in a collective or individual memory. Signs A and B in A=B on the one hand and the two occurrences of A in A=A on the other hand, return certainly to the same referents, but in A=B, it consists of an operation which identifies, by means of a stating, in A=A, it consists of an entry of charge of a general property of the identity concretized on a particular object.

In A=B, the referential value of A is identified with that of B (hence the identity of the referents results from this immediately), this discursive identification remains provisional, it determines a pragmatic environment (context). In A=A, identity is general and independent of the context created by the enonciator (Desclés 87).

The value of identification relator is characterized solely by the properties of symmetry and reflexivity. This relation is thus not directed (symmetry) as are the other semantic relations of static types (localization, whole-and-part, membership...). All identification values, i.e. all kinds of identification, of resemblance, of equivalence, of equality and identity, amongst other identifications of extensional equality or intentional equality (identity), are binary relators that have both properties of symmetry and reflexivity.

When given, in addition to symmetry and reflexivity, monotony (on the left and on the right) also called laws of Leibniz, as well as transitivity, there is equality. While adding, moreover, the axiom of extensionnality ("If for each X, the concept F applied to X gives the same result as the concept G applied to X, then the concepts F and G are extensionnality equivalent"), there is identity.



When a statement has value of identification, two distinct instances are brought to unification, they relate to the same object. There is ultimately only one object, but this unicity is the result of a construction often with, like attribute, nominalized adjectives by the article (i.e. "*être le meilleur*, *le plus beau*, *le premier*…").

Tendency to the equality of statute of the subject and the predicate appears in certain possibilities of reversibility of the relation: "*La capitale de la France est Paris*".

Nevertheless, however strong it may be, identification between the subject term and the attribute term is rarely connected with an identity: subject term remains normally most concrete, and the attribute the most abstract, the most qualitative.
The four paraphrases below differentiate from the point of view of their value and their employment (Le Goffic 93):

- *"Paris est la capitale de la France*": natural statement, very attributive identification
- *"Paris, c'est la capitale de la France*": natural alternative with marked value of identification
- *"La capitale de la France est Paris*": artificial statement, except in answer to a question
- *"La capitale de la France, c'est Paris*": natural statement of identification proceeding of the role towards theterm which fills it.

In corpus, one will find primarily identification and equality relations but very rarely identity relations. At the time of the annotation, we will thus not distinguish the relations. We will remain on the highest level: identification.
Copula "*être*" is the principal expression of the relation of identification. The prototype example is "*Paris est la capitale de la France*".
However, copula is polysemic and can express well other values of location. It will thus be necessary to restrict the rules with certain conditions to avoid a noise which could be important.
Following constructions of the type will thus be detected:

- Concept1 (entité nommée) *est* Concept2 / Concept1 (entité nommée), *c'est* Concept2 ;
- Concept1 *est* Concept2 (entité nommée) / Concept1, *c'est* Concept2 (entité nommée) ;
- Concept1 *est le… de*.

We defined rules which detect lexical turnings expressing a identification relation:

- Concept1 *est le même que* Concept2 ;
- Concept1 *n'est autre que* Concept2 .

Two lists of names expressing an identification, &InomIDE1 = {*égalité, identification…*} et &InomIDE2 = {*définition, notation…*}, are used as indicators to trigger the rules:

- Concept1 *est en égalité avec* Concept2 ;
- Il y a *égalité entre* Concept1 *et* Concept2 ;
- Concept1 *a la même définition que* Concept2 …

An adjective or a past participle expressing identification can be the indicator, the copula "to be" not being a necessary index:

- Concept1 *(est) égal à* Concept2 ;
- Concept1 *(est) défini comme le* Concept2.

Certain verbs can express a relation of identification, for example:

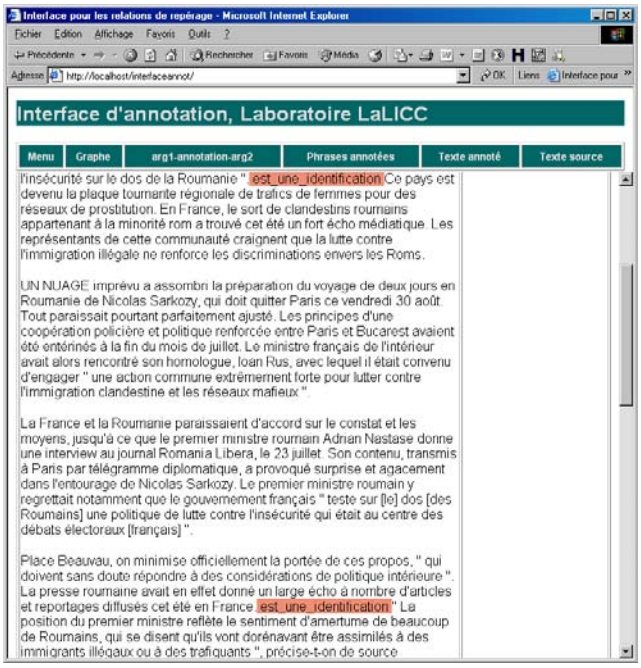
- Concept1 se *définit comme le* Concept2 ;
- Concept1 *signifie* Concept2 ;
- Concept1 *se traduit "*Concept2" …

In the rules of identification, the forms of "*être*", "*avoir*" and the verbs will have to be limited to the present, with imperfect and infinitive, in order to make sure that the relation expresses a static state.
For example, the rule below (Provôt 2005) expresses an identification relation between two concepts, where an occurrence of the verb "*être*" is the indicator. It is said that the first concept is the same ("*est le même*") one as the second concept.

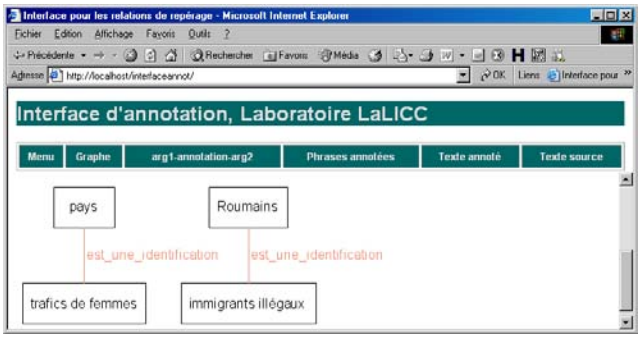
Rule name	Rident02
Task	Relation_statique
Indicator	&Iest
Body	# Research space E1 := créer_espace(Gauche(I)) E2 := créer_espace(Droite(I)) E3 := créer_espace(Droite(x1)) E4 := créer_espace(Droite(x2)) E5 := créer_espace(Droite(x3)) # Indices lists L1 := &articleDef L2 := &meme L3 := &que
Conditions	Il_existe_un_indice x1 appartenant_à_E2 tel_que classe_de x1 appartient_à L1 Il_existe_un_indice x2 appartenant_à_E3 tel_que classe_de x2 appartient_à L2 Il_existe_un_indice x3 appartenant_à_E4 tel_que classe_de x3 appartient_à L3
Actions	attribuerEtSem(PhraseParent_de I, "est_identifié")

For identification annotation, we have 17 rules and 27 classes of markers gathering 186 indicators or indexes.

Let us launch annotation of identification relations on an article of two pages originating from the Web, under the heading "*Visite en Roumanie de Nicolas Sarkozy sur fond de trafics humains*".



Complete text where the annotated sentences are marked



Graphs using the terms of the user

Conclusion

Automatic annotation of semantic relations with platform EXCOM is operational today as it is shown in the results presented above. The contextual exploration engine of the platform EXCOM annotates the texts according to the criteria selected by the user, for example localization or identification relation and, in a simple interface, makes the results available in a graphic form or textual form, either by presenting all the text with the annotated sentences marked by underlining, or by offering an extraction of the annotated sentences. It also permits navigation between all various available forms of annotation representations.

References

Abraham Maryvonne 1995, Analyse sémantico-cognitive des verbes de mouvement et d'activité, Contribution méthodologiques à la constitution d'un dictionnaire informatique des verbes, Ph. D., EHESS

Crispino Gustavo 2003, Une plate-forme informatique de l'Exploration Contextuelle : modélisation, architecture et réalisation (ContextO). Application au filtrage sémantique de textes, Ph. D., Univ. Paris-Sorbonne,

Cruse David 1986, *Lexical Semantics*, Cambridge University Press

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

Desclés Jean-Pierre 1990, *Langages applicatifs, Langues naturelles et Cognition*, Hermès, Paris

Desclés Jean-Pierre 1991, Architectures, représentations cognitives et langage naturel, in *Les sciences cognitives en débat*, Editions du CNRS, Paris

Desclés Jean-Pierre 1997, Système d'exploration contextuelle, in *Co-texte et calcul du sens*, 215-232. Calif : eds Guimier, Presses Univ. Caen

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 J-P., Flageul V., Kerenbosh C., Meunier J-M., Richard J-F. 1998, Sémantique cognitive de l'action, I. Contexte théorique, in *Langages*, 132:28-47

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

Flageul Valérie 1997 Description sémantico-cognitive des prépositions spatiales du français, Ph. D., Univ. Paris-Sorbonne

Hamon Thierry 2000, Variation sémantique en corpus spécialisé : acquisition de relations de synonymie à partir de ressources lexicales, Ph. D., Univ. Paris-Nord

Jouis Christophe 1993, Contribution à la conceptualisation et à la modélisation des connaissances à partir d'une analyse linguistique de textes, réalisation d'un prototype : le système SEEK, Ph. D., EHESS Paris

Le Goffic Pierre 1993, *Grammaire de la phrase française*, Hachette

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 Florence 2005, La relation d'identification, Technical Report 2005/01, LaLICC, Univ. Paris-Sorbonne

Le Priol Florence 2005, La relation de localisation, Technical Report 2005/02, LaLICC, Univ. Paris-Sorbonne

Provôt Agnès 2005, Étude en linguistique expérimentale des relations statiques d'identification et de localisation en vue d'une implémentation de règles d'exploration contextuelle pour la plate-forme EXCOM, DEA MIASH, Univ. Paris-Sorbonne