World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:2, No:1, 2008

Semi-Automatic Analyzer to Detect Authorial Intentions in Scientific Documents

Kanso Hassan, Elhore Ali, Soule-dupuy Chantal, and Tazi Said

***Abstract*—**Information Retrieval has the objective of studying models and the realization of systems allowing a user to find the relevant documents adapted to his need of information. The information search is a problem which remains difficult because the difficulty in the representing and to treat the natural languages such as polysemia. Intentional Structures promise to be a new paradigm to extend the existing documents structures and to enhance the different phases of documents process such as creation, editing, search and retrieval. The intention recognition of the author's of texts can reduce the largeness of this problem. In this article, we present intentions recognition system is based on a semi-automatic method of extraction the intentional information starting from a corpus of text. This system is also able to update the ontology of intentions for the enrichment of the knowledge base containing all possible intentions of a domain. This approach uses the construction of a semi-formal ontology which considered as the conceptualization of the intentional information contained in a text. An experiments on scientific publications in the field of computer science was considered to validate this approach.

***Keywords*—**Information research, text analyzes, intentional structure, segmentation, ontology, natural language processing.

I. Introduction

THE domain of automatic natural language processing is in the heart of the problematic of extraction and information search. It seems obvious that future progress will pass by better comprehension of the language. The current state of research is far from this comprehension, and many difficulties arise at all the levels of the writing analysis. The problems can be morphological, syntactic, semantic or pragmatic. To carry out the various tasks of classification, search and filtering for documents, it becomes essential to represent the texts in pragmatic way to be comprehensible by the machine. In order to do that many fields can be interested with this problem such as Information Retrieval, Artificial Intelligence or Natural Language Processing. Information retrieval to which we are going to refer to mainly follows two aspects: on the one hand the representation of the contents of documents, and on the other hand the information access.

In this article, we tackle the problem of representing a specific structure that can be extracted from documents, it is called Intentional Structure. Several types of structures can be identified and used to describe information and to facilitate research and the restitution. The most fluently approached structures in documentary information, according to the type of concerned documents, cover with supplementary aspects: physical structure (related to the lay out), logical structure (generally hierarchical organization of the various elements composing a document), semantic structure (semantic decomposition of a document), rhetorical structure (is a descriptive and functional theory of the textual organization based on the recognition of rhetorical relations between units of text). The exploitation of logical and physical structures has an already proven interest with the aim of facilitating fragmentation, storage and restitution of documents. However, the documents structures based on rhetoric, semantics and in particular the communication intention are neither yet sufficiently studied, nor exploited in the documentary systems.

The assumption of this work is inspired by the intentionality theory [14]. Indeed, the study of this theory conducts us to propose a model which will make it possible to integrate the concept of the communication intention of the authors of documents in the process of information research.

By basing on the theory of the intentionality, we developed an analyzer whose objective is to find the communication intentions of the authors from documents. This analyzer uses techniques from natural deduction, close to those used by an expert domain to perform the recognition. Its specificity is that it is able to find the author intentions, to refine its strategy of analysis of a new text and to produce automatically an ontology of the intentions. The research and the identification of the intentions are based on a segmentation of texts, then the analysis of each segment to extract the intentional verbs and their associated concepts. The used techniques of segmentations and the methods of extraction and analysis of the intentional verbs are described in this paper.

This article is organized as follows: Section 2 is dedicated to present the notion of intention. Section 3 presents the formalization of intentions. Section 4 presents our analyzer that we propose for the intentions recognition of the textual corpus. Section 5 presents the intention research in the analyzer. The last section presents ontology of intention extraction.

II. The Notion of Intention

The intention corresponds to a mental state of any actor who executes an action. We are interested exclusively in the intentional actions, i.e. the actions which are premeditated.

Three types of questions are posed on this level:

--What is an action?

--What is an intentional action?

--What is to explain an action?

The causal theory answers to the first question by stating that what distinguishes the behaviours or events which constitute actions is to have a certain type of mental cause or

to utilize a certain type of psychological causal process. It answers to the second question by stressing that an action under a given description is only intentional if certain relations between this description of the action and the contents of its mental antecedents are satisfied. Finally, it answers to the third question by saying that an explanation of an action is an explanation by reasons but such reason has truly value of explanation in the susceptible intentional states to rationalize the action cause the actions which they rationalize [11] [5].

Several works attempt to account for the relations between an action undertaken by a human being and the mental state which guides this action. Searle remains a main reference on the matter [14]. He distinguishes between two types of intention: intentions in the course of action and the 'pre-formulated' intentions

--Pre-formulated intentions represent a condition of satisfaction of the intention.

--Intentions in the course of action are those which represent these intentions.

whereas the intention in the course of action accompanies the action during its execution. This distinction makes it possible to treat only intentional actions, and not the "micro actions", or the movements which are inevitably non intentional.

III. The Formalization of Intentions

We propose a new concept which enables us to treat a document in terms of the intentions of its authors. This analysis was used as a specification to build the algorithms of the suggested system. Our objective is to have a representation of the intention by the relations between its components. By definition we represent an intention as:

Where: I (A, G*, M*, R*)

Where:
'I' represents the intention carried out by an action A;
A is the action that expresses what the author of the intention wants to do;
G represents the goal to achieve by performing the action;
M represents the means to express how the action is accomplished;
R represents the reason to express why the author chooses this action and for which reasons, *indicate that the number of the acts composing the intention can be 0 or N[1] [2] [3] [10] [17].

In the following section we present our system analyzer: its functional principle, its life cycle, and different stages of intention research.

In the following section we present our system analyzer: its functional principle, its architecture, and different stages of intention research.

IV. Analyzer

The analyzer is dedicated in the information research starting from textual corpus mono-domain. It is based on a algorithm which facilitates the intentions research and its principle of operation closer to an domain expert.

The system uses a knowledge base which represents how an author can explain his intentions. In a preliminary phase, the enrichment of the knowledge base is done manually through the set of intentions recognition of textual corpus by human domain expert.

The second phase consists of the introduction of textual corpus. The analyzer calls some tools of linguistic analysis such as: **Treetagger**, for extracting the verbs for each segment, **Wordnet** to find the synonyms of the verbs which belong to the same segment in order to minimize the set of verbs. The third tool used by the analyzer in this phase is the knowledge base that's contains the intentions in order to find out the intentional verbs of the segment of textual corpus.

This analyzer allows also to adopt a method of counting the intentional verbs of each segment to find the occurrences of each verb in order to announce the intention of each segment. It has also the possibility of generating intentions ontology of a corpus containing all possible intentions.

Fig. 1 depicts the architecture of the analyzer system with its various stages keys.

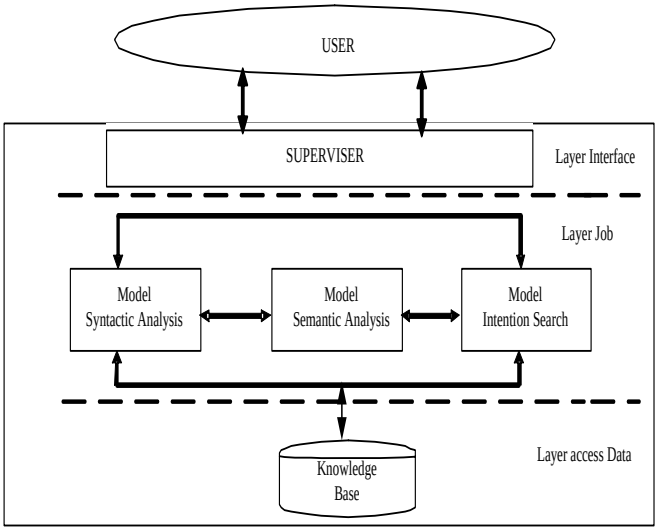

Fig. 1 Analyzer architecture

More precisely, the analyzer includes/understands three modules and a source of information. We cited briefly here.

The *Syntactic Analysis* based on the Treetagger system which makes it possible to make a syntactic analysis on each logical element of the document. This module generates textual files which contain lists of verbs.

The *Semantic Analysis* is based on Wordnet, it makes a semantic analysis in order to find the synonyms of verbs in the annotated corpus in the same segment in order to avoid the redundancies of verbs.

And the *intention search* which based on the intentional model in order to facilitate the process of recognition of the intentions of a textual corpus.

In the next section we present the various stages which it carries out for the intention search.

V. The Intention Research in Analyzer

This section is dedicated to an illustration of the intention research sequence of a textual corpus mono-domain in scientific publications.

The domain we considered for this study is publications articles in computer science. One of the reasons why we chose to work with scientific articles is the practical value of better document retrieval environments for scientists. Scientific papers are different from other text types with respect to their overall structure, an aspect we are particularly interested in. And as we read this kind of documents all the day, we consider that we are the experts.

This kind of documents is justified by the fact that Intentions in scientific documents are deterministic. This means that we consider that for each fragment there is one and only one intention. The analysis was limited to the introductions of the textual corpus to validate and facilitate the comprehension of our proposals. In the following we present the author task and the analyzer task which enable us to recognize the relevant intention.

i. Author Task (Manual Segmentation)

We propose to apply the method of segmentation to textual documents of which one knows the logical structure. The analysis is made on each logic element such as the introduction, the chapters, and the conclusion. It is a question of delimiting, in each analyzed element, the segments of text corresponding to the intentions. The manual annotation follows a method which consists in following a certain number of stages based on a manual analysis of documents corpus. It is a question of identifying the segment (called fragment).

One idea is to make segmentation on units of sense which bring a semantic and pragmatic, clean and autonomous unit. Sentences are regrouped in manner to constitute a semantic unit in which one suspect that the author wishes to make make something by his reader. This can be a physical or mental action. These documents are therefore segmented manually according to our method of segmentation (Fig. 2). Each segment corresponds to an intention, and it is constituted of whole of sentences which one identifies the verbs.

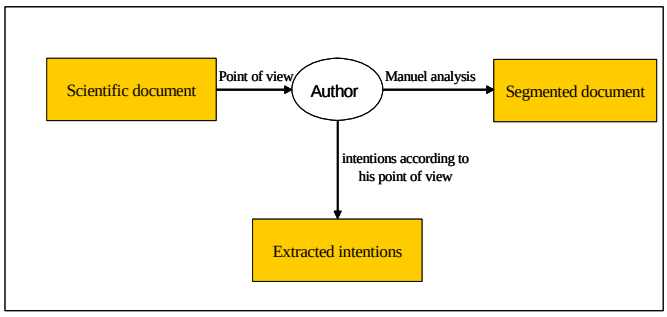

Fig. 2 Manuel segmentation of documents according to the author

ii. Analyzer Task

The second task is entirely taken in hand by the system. It requires six stages as is shown in Fig. 3.

1) *Initialization* is a task which set up the necessary resources to all other following operations. It is the first task being launched and is carried out only before starting the other tasks. Initialization in the analyzer system, allow us to introduce a corpus annotated by an expert, and to enrich the knowledge base starting from the knowledge described in the input text.

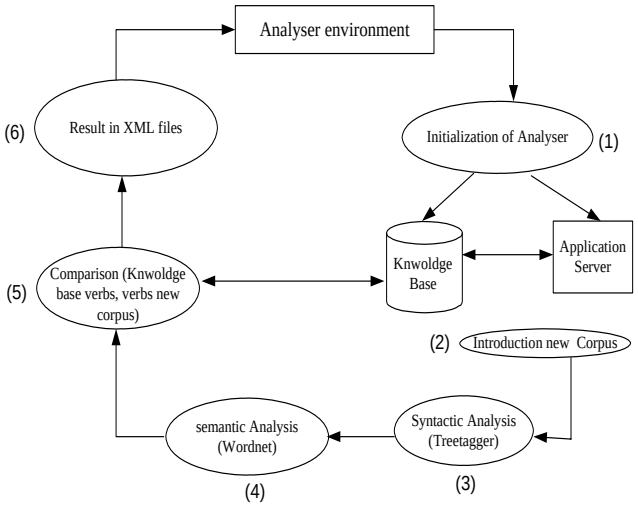

Fig. 3 Life cycle of analyzer

This base contains a list of the intentional verbs of the segments, and the statistical elements which indicate the occurrences of the verbs in same intentions (relative Frequency = F_r) and also the elements which indicates the occurrences of the verbs in a segment(absolute Frequency = F_a).

The original idea here is to constitute at the beginning a knowledge base introduced by the human expert in which the system learns that for each intention of the introduced segments there are representative verbs of the intention (in terms of their absolute frequency compared to the set of the input texts) and relative (compared to the set of the introduced segments which has the same intention).

This initialization is an important phase because it allows the system to recognize the verbs according to their frequency by using an ontology of the intentions.

2) *Introduction of a new corpus*, this stage makes it possible to the reader to introduce a new corpus not segmented, and used it like the entry of the system, in order to make an analysis to segment it according to the authors intentions.

3) *Syntactic Analysis (Treetagger)* allows the system to make a syntactic analysis on each logic element of the document. The analyzer recognizes the sentences and the verbs using the ontology of the verbs.

Our system analyzes this corpus, sentence by sentence in order to make the segmentation according to author intentions. We took each sentence as an entry in a software which called (TreeTagger) for extract the verbs from this sentence.

TreeTagger is a tool for annotating text with part-of-speech and lemma information which has been developed within the TC project at the Institute for Computational Linguistics of the University of Stuttgart. The TreeTagger has been successfully used to tag German, English, French, Italian, Dutch, Spanish, Bulgarian, Russian, Greek, Portuguese, Chinese and old French texts and is easily adaptable to other languages if a lexicon and a manually tagged training corpus are available. The free version of the software currently allows such an analysis in the main European languages.

The input of sets of the sentences in Treetagger generates as an output a textual file composed by a sets of infinitive verbs. Initially, the use of this software enabled us to create a list of verbs for each segment from our segmented documents.

4) *Semantic Analysis (Wordnet)* after the generation of the textual files which contain lists of the verbs, the semantic analysis uses Wordnet to find the synonyms of the verbs in the annotated corpus in the same segment in order to avoid the redundancy of the verbs, and at the same time to find the other synonyms of these verbs.
   Wordnet is a lexicographical ontology for the English language developed by Princeton University. It is represented in the form of lists linked between them to create a network. It is used for a dictionary (WordWeb2), an expert system (SearchAide), a software for automatic annotation of texts, etc.
   Consequently the system treats the result obtained by Wordnet, after having made a counting of verbs which are repeated in the same segment.
   The basic assumption of this step is that the number of occurrence of verbs in the segments (of each intention in the same given logical element) is represented by a percentage (relative Frequency). That means that the intention I is represented by a list of verbs (Vi) which one calculates the percentage of occurrences in the segments corresponding to the same intention in the same logical entity of the document (for example, intention I1 of the logical entity " Introduction "). The analyzer recognizes the sentences and the verbs using the ontology of intentions. And one calculates the absolute frequency of each segment i.e. the number of occurrences of the verbs in this segment. The verbs and their frequencies absolute and relative are stored in the knowledge base with intentions Ii which were identified manually. We combined a certain number of scores of relevance available for each verb. In this stage, we present a statistical method based on learning which extracts the relevant verbs from the sentences of a document, i.e. those which have relevance scores.
5) *Comparison (Knowledge Base verb, new corpus verb)* will be made according to most relevant frequencies of these verbs in the knowledge base and that of the verbs of a new document (which is an estimation of the verbs probability that repeat in the knowledge base with big frequencies). In this stage we obtained a segmentation by sentences, and we used then the principle of regrouping sentences by intentions. For that, all the contiguous sentences which have verbs at the same intentions are regrouped at the same segment.
6) *Result on XML file*, this stage allows the generation of an XML file containing the results of segmentation accompanied by intentions (Fig. 4).

VI. Ontology of Intentions

There are several methods devoted to construction ontologies from corpus of text. The majority is based on the contents of texts to build ontology; the texts are then the principal source of knowledge for the information acquisition [13]. All the concepts of the model as well as their relations exclusively come from an analysis of texts, without external contribution of knowledge.

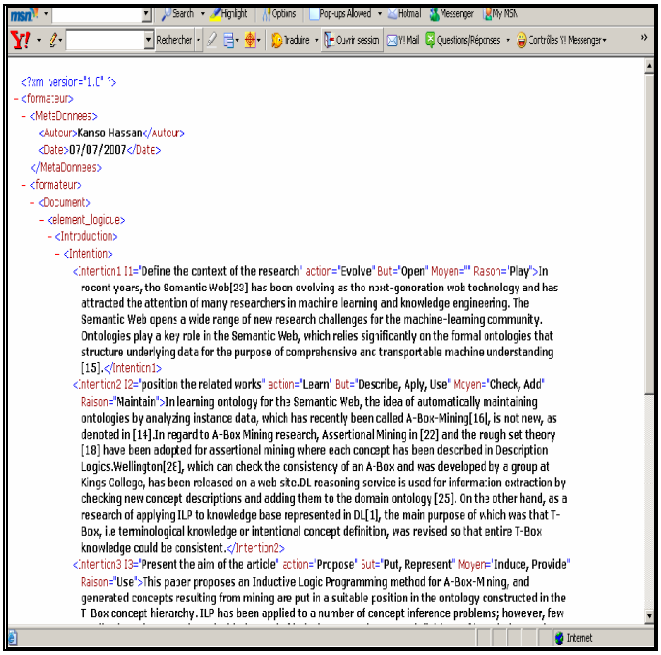

Fig. 4 Analyzer output on XML file

The authors [16] register in this step even if it is recognized that the texts can not constitute the single source of knowledge. This endogenous approach decomposes into several stages: the extraction of terms referring to the basic concepts (conceptual primitives [13]), then lexical relations which they maintain (creation a terminological base [15], [16], [8] in order to make emerge the first relations inter-concepts. The following stage is based on the analysis of semantic relations between terms to extract new relations between concepts as well as new concepts in order to precise a semantic network of concepts. The semantic network must be validated by domain expert to specify the significant relations (normalization [12]).

The result corresponds to the definition [7] “an ontology is a hierarchical structure of a set of terms of domain ", that is an ontology that can be represented in a formal or semi-formal

way.

The use of ontologies in our case will facilitate the recognition of the intentions. Ontologies will help the authors of the documents to specify their intentions, and they intervene in the intentions recognition of document [9] [10]. The knowledge representation of an intentional ontology that we use is as follows. An existing thesaurus contains the intentions of a precise domain; these intentions were put by an expert of the domain during the method of system initialization. The conception of ontology is generated in parallel with the execution of different the stages from the analyzer, After having introduced the corpus of a document, the extraction of the concepts and verbs of each segment, with the assistance of thesaurus intentions, the analyzer identifies the intentional verbs.

The use of the synonyms, provoke a reduction of the sets of verbs of a segment. From a corpus in a precise domain, we made an extraction of the terms and concepts of each segment in order to build relation between different verbs and concepts. The structuring of the ontology of verbs and the concepts dependent between them is done by the use of a generic ontology (Wordnet), and the existing thesaurus with the scientific corpus.

The schema (Fig. 5), show also the taking into account of updating the thesaurus for every introduction of a new corpus.

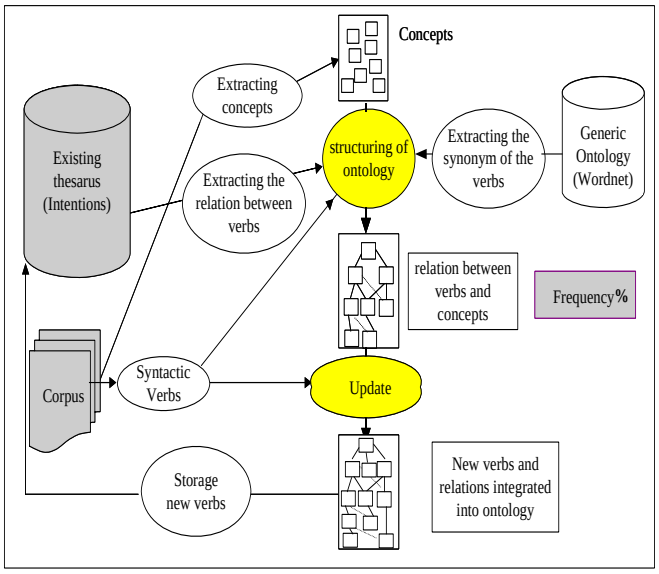

Fig. 5 Architecture of construction the intentions ontology of a corpus mono-domain

According to our definition of intention, an intention is composed of acts considered as action, goal, means and reason. We consider that each act of an intention is represented by the most significant verb of a segment. Each segment is dependent with a single intention. The name of an action is obtained from that the verb from which it rises; we could therefore recover the list of used actions by analyzing the chain verbe1 till verb N. However, this analysis should be carried out with each analysis of a new corpus, it is therefore preferable to memorize the link between an action and a verb from which it rises. We propose a new formalization for the intentions recognition of the textual corpus.

Lemme
We are interested in this work in the intentional verbs, i.e. the verbs which translate intentional actions $(V_I \Rightarrow A_I)$

T is a set of p text segments S : $T = \{S_i , (i=1,...,p)\}$, V_s is a set of n verbs that represents S: $V_S = \{ V_1, ……., V_n\}$, $V_{max}(s, v, s_y)$ is a function related with three variables (s : segment, v : verb and s_y : is a set of synonymous of the verb v) present the verb which have a maximum occurrence in a segment S.
S_y is a set of synonyms of each verb, and C_I is a set of concepts extracted from the segment S_i ; $C_I = \{c_1, \ldots, c_n\}$. Each occurrence of a verb is associated a node of an ontology tree.

$$I = \{ V_{max} \cup C_I \}$$

The result obtained is the union of an intentional verb with its concept which makes it possible for the users to identify the intention of a segment.

VII. Conclusion

Dealing with “intentions” in electronic documents is a very hard task. This article presents a research challenge: how to build a semi-automatic system for author’s intentions recognition based on written documents? The main result we can consider here is that one has to restrict the application domain and that we need very rich ontological representation.

The presented analyzer allows to recognize intentions, it is based on a semi-automatic method of extraction the intentional information from a corpus mono-domain. Ontologies are used with a knowledge representation language for the machine and are exploited with possibilities of inference. The objective thus of this ontology is not only the intentions recognition but also enrichment the knowledge base by new detected intentions.

The utility of our research is to improve the performances of the information research systems, or in other words, to make the organization of these corpus in order to facilitate the information access. Through this system the readers will be able to find documents not only in terms of the underlying concepts but also in terms of authorial intention.

The perspective for our work is the taking into account the intentional structure in the domain of information research, in order to answer the user needs. We hope that the search engines take into account this type of structure (intentional) in research techniques.

References

[1] A. El Hore and S. Tazi, “Apprentissage par observation à travers les intentions du raisonnement”, Conference on Information Technologies, MCSEAI’06, 7-9 december 2006.

[2] A. El Hore and S. Tazi, "Explaning and Indexing Solutions for Physics Learning", IEEE, Conférence International CELDA'05, p. 532-538, 13-17 décembre 2005.
[3] A. El Hore and S. Tazi, "Pero a planning system for the explanation of problem solving in physics", Mixed Language Explanation in Learning Environment (MLELE'05) in conjunction with AIED'05, Amsterdam, p. 80-81, 18-22 juillet 2005.
[4] A. El Hore and S. Tazi, "Planning and explaining solutions for physics learning", IEEE international conference on Machine Intelligence (http://www.acidcaicmi2005.org/) The 2nd ACIDCA-ICMI'2005. Tozeur – Tunisia, November 5-7, 2005.
[5] A. El Hore and S. Tazi, "Planning solution for physics learning", International conference CAPS'05 http://europia.org/ICHSL05/) CAPS 05 , Marrakech – Morocco, November 22-25, 2005 .
[6] B. Grosz, and S. Kraus, "Collaborative plans for complex group action", Artificial Intelligence, 86(2):269-357, 1996.
[7] B. Swartout and P. Ramesh and K. Knight and T. Russ, "Towards Distributed Use of Large- Scale Ontologies" Proceedings of 10th Knowledge Acquisition Workshop (KAW), 1996.
[8] G. Lame, "Knowledge acquisition from texts towards an ontology of French law", Proceedings of EKAW, 53-62, 2000.
[9] H. Kanso, C. Soulé-Dupuy and S. Tazi, "Reconnaissance des intentions de communication écrite dans des corpus de documents scientifiques", 9ème édition de la conférence H2PTM : Hypertextes, Hypermédias 29,30 et 31 Octobre 2007 - Hammamet, Tunisie.
[10] H. Kanso, C. Soulé-Dupuy and S. Tazi, "Representing author's intentions of scientific documents", Funchal, Madeire (Portugal), 12-16 Juin 2007.
[11] I. Pacherie, "La dynamique des intentions", Dialogue, XLII, 3, 2003, pp. 447-480.
[12] J. Bouaud and B. Bachimont and J. Charlet and P. Zweigenbaum, "Methodological Principles for Structuring an Ontology", Proceedings of IJCAI, 1995.
[13] J. Nobécourt, "A method to build formal ontologies from texts", Proceedings of EKAW, 21-27, 2000.
[14] J. Searle, "Intentionality", Cambridge: Cambridge University Press, 1983.
[15] N. Aussenac-Gilles and D. Bourigault and A. Condamines and C. Gross, "How can knowledge acquisition benefit from terminology", Proceedings of EKAW, 1995.
[16] N. Aussenac-Gilles and B. Biébow and S. Szulman, "Corpus analysis for conceptual modelling", Proceeding of EKAW'2000, 13-20, 2000.
[17] Y. Al-Tawki, "Création par réutilisation de documents décrits par les intentions de l'auteur", Doctorat de l'Université de Toulouse 1, 2002.