Highlighting Document's Structure

Sylvie Ratté, Wilfried Njomgue, and Pierre-André Ménard

***Abstract*—**In this paper, we present symbolic recognition models to extract knowledge characterized by document structures. Focussing on the extraction and the meticulous exploitation of the semantic structure of documents, we obtain a meaningful contextual tagging corresponding to different unit types (title, chapter, section, enumeration, etc.).

***Keywords*—**Information retrieval, document structures, symbolic grammars.

I. Introduction

DOCUMENT management (classification, indexation, storage) is practised for a long time, in particular in libraries. In recent years, due to technological progress and the transition from paper to numerical form, the amount of the textual resources has become gigantic [1], making it difficult to exploit manually. In order to make a better use of this large resource, information retrieval (henceforth IR) systems were created, aiming at improving the quality and efficiency of knowledge extraction.

Our project objective is to produce the prototype of a system that would extract business rules from corporate texts and translate them into a visual software engineering model. Within this perspective, it is not only a question of extracting knowledge, but also to be able to visualize correctly extracted relevant information. The upstream of this project, object of this paper, is to extract knowledge characterized by document structure and contents. We will focus here on the extraction and exploitation of the semantic structure of documents in order to obtain a meaningful contextual tagging corresponding to different unit types (title, chapter, section, enumeration, etc.).

This paper is organized as follow. First, we will underline previous researches in information retrieval that have taken into account document structure to improve results. Then, we will describe the models we propose for the detection of some document structures. A group of experiments, to test these models, is then presented. Finally, we will mention problems that have to be solved in our future research.

II. Previous Researches

Some of the deficiencies displayed by traditional techniques of IR can be directly related to the fact that they do not use or barely use the structure of documents (see [2]). However, every human reader admits that some entities like title, summary, or subtitle can convey very relevant information. At this point, it is appropriate to distinguish between the logical structure and the physical structure of documents. The organization of a document in chapters, sections, titles, paragraphs, concerns its logical architecture whereas its physical structure or presentation is characterized by its layout. In IR, on a macroscopic scale, the title of a chapter announces us what will follow and, in a way, summarizes the contents of what will follow. The words extracted from this title are more relevant than others words in the document. The transition from a paragraph to another is synonymous of changing an idea. The logical structure is the result of the author's will in the organization of the document. On the other hand, the physical structure is the consequence of external constraints due to the design layout. It also translates in a visual way the logical architecture of the document. In this paper, the expression "document structure" will refer only to the logical structure of documents.

Nowadays, researchers are unanimous to recognize that the exploitation of the textual document's structure would be a significant additional asset in information retrieval. Schlieder and Meuss [3] affirm in these words: "to be unaware of document's structure is equal of being unaware of its semantics". Generally, document's information is more or less dissimulated in its structure; this prescribed the correct interpretation of the document [4]. Thus, some researches exploited XML[1] documents which have the advantage of offering a structure which facilitates their representation and their exploitation in various contexts [5].

This logical structuring of XML documents is very well defined, hence allowing the use of some conversion techniques to extract it. For example, the indexing method used in the system CONCERTO [6] takes into account the structure of the documents but do not propose a method to identify correctly these structures. Researches of [2] on ontology construction are also based on already tagged document structures. In this system, a Perl program detects the

Manuscript received August 31, 2007. This work was supported in part by the Natural Sciences and Engineering Research Council of Canada under a grant obtained by Sylvie Ratté.

Sylvie Ratté is with École de Technologie Supérieure, Software and IT Engineering Department, 1100 Notre-Dame West, Montréal, QC, Canada, H3C 1K3 (phone: +1 514-396-8612; fax: +1 514 396-8405; e-mail: sylvie.ratte@etsmtl.ca). She is the director of the Cognitive and Semantic Engineering Laboratory (office A-3442).

Wilfried Njomgue is with the Cognitive and Semantic Engineering Laboratory of the Software and IT Engineering Department, 1100 Notre-Dame West, Montréal, QC, Canada, H3C 1K3 (e-mail: wnjomgue@gmail.com).

Pierre-André Ménard is with the Cognitive and Semantic Engineering Laboratory of the Software and IT Engineering Department, 1100 Notre-Dame West, Montréal, QC, Canada, H3C 1K3 (e-mail: pmenard@gmail.com).

[1] XML : eXtensible Markup Language, http://www.w3c.org/XML

logical architecture of documents in order to produce XML documents which served as input to build a flora ontology. In many fields of IR, document structure's semantics is fundamental. However, few researches are devoted to their identification.

The unstable results obtained by extraction tools during the experiments carried out by [2] are specifically due to the fact that the structure of documents is not taken into account. In fact, throughout textual research, document structure rarely has a specific treatment [7]. Frequently, this structure is even regarded as an obstacle to the content analysis of documents. Beside many systems include a pre-treatment phase that eliminates any form of structure, in order to take into consideration only words [6], what we have called "document's content".

Knowing that information is the result of both content and structure, it is of interest to all researchers in IR to find methods which use simultaneously both kind of information. This is the objective of the overall project: first to highlight the existing structural models (title, section, sub-section, chapter, paragraph, list of enumeration, etc), second, to exploit them and finally to process the "total" contents of documents. We are now going to explain the techniques we used to uncover this documentary structure and how it can be exploited in the downstream of the project.

III. Recognition Models

Among elements which compose the logical architecture of document, we were interested initially in those that are visually prominent, such as title, section, chapter, and enumeration list. The creation of our models is based on the construction and the syntax of logical structure of French documents. The models were elaborated and tested relying on a substantial corpus of corporative documents (see section 4).

Models presented here are realized with Intex[2] and Nooj[3] but can easily be adapted to other tools that can correctly identify tokens, types of punctuations, end of sentences, and numbers.

A. Recognition Model for Sections and Chapters

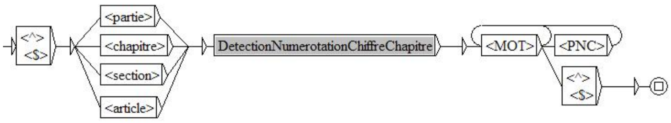


Fig. 1 Graph of detection of chapters and sections. <MOT>, <PNC> respectively identify any word and any form of punctuation used in French literature

The graph of Fig. 1 recognizes any linguistic form made up with any word which has as lemma "partie" , "chapitre", "section", or "article", followed by an alphanumeric numbering (for example II, II.1, A, A.1,etc.), and ended with some sequence of words (<MOT>) or a punctuation (<PNC>). The signs <^> and <$> respectively mark of the beginning and end of sentences.

The sub graph included in this graph ("DetectionNumerotationChiffreChapitre") is presented in Fig. 2. The graph contributes to the detection of various form of alphanumeric numbering. The sub-graph « ChiffresRomains » lists all the Romans numeral from zero to 2999. The sub-graph « AlphabetMajuscule » is the list of all capital letters of the French alphabet. Since we could not find a reference for chapter that use a small letter (« Chapter **a** : …. », « Chapter **i**…. »), we have included only capitalized ones (« Chapter **A** : …. » or « Chapter I…. »).

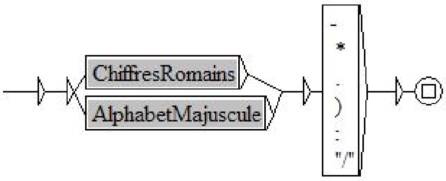


Fig. 2 Sub-graph « DetectionNumerotationChiffreChapter » incorporated in the graph of Fig. 1

B. Recognition Model for the Title of a Document

Very often, the title of a document is the first sentence of that same document. In this position, it is regarded as an expression:

- starting with a capital letter (<PRE>) and not ending with a point,
- written in capital letter (<MAJ>),
- where all words start with a capital letter,
- being center aligned and having various characteristics quoted above.

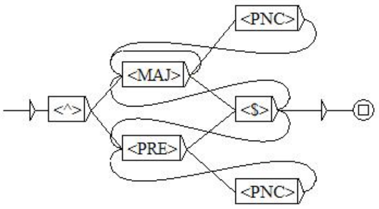


Fig. 3 Graph of detection of the title of a document <PRE>, <MAJ> symbols indicated respectively any word starting with a capital letter and any word written in capital letter

C. Recognition Model for an Enumeration List

As mentioned before, in most IR systems, the logical structure of document is absent or, as in the case of enumeration list, defective. In example (1), a standard IR system will extract the words « family », « father », « mother », « children » without establishing any relationship between them.

A family consists of: (1)
- a father,
- a mother,
- children.

[2] http://intex.univ-fcomte.fr
[3] http://www.nooj4nlp.net

In order to solve this problem, some systems use the notion of co-occurrence between words to preserve the relationship [8]. The model we propose is illustrated in Fig. 4. It is partially based on the linguistic analysis presented in [9] although these authors are using semi-structured documents. The sub-graph « DetectionNumerotation » recognizes all mixed alphanumeric numbering. To avoid any conflict with the detection of chapter structure (A.1, I/, IV/, etc.), we hypothesise that these mixed alphanumeric numbering would not be in capital letter (1.2, a.1, i/, 1.a, iv/, etc.).

Before building other types of complex extraction models such as those necessary to recognize tables, whose extraction is much more complex, we have tested our models on a large French corpus.

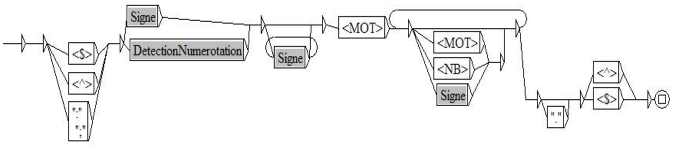


Fig. 4 Graph of detection of a list of enumeration in a document. <NB> symbol designs any number

IV. Experiments and Results

To realize these experiments, we obtained from Ecole de technologie supérieure of Montreal substantial corpus of documents. This database consists of various types of documents such as constitutive documents[4], meeting documents[5], administrative documents, human resources documents[6], documents of communication[7], legal documents like specifications and procedural documents [10]. Documents used in these experiments are semi-structured or unstructured documents. Semi-structured documents have a certain logic and semantics configuration like a graph (title of the document, section and sub-sections). In the case of unstructured documents, we know neither the context, nor the way information is fixed.

In order to try out these models and to display the results, we have used Nooj which is an improved version of the terminological extractor Intex. Nooj allows a fast and interactive development of automaton, transducers and grammars to analyze texts (represented as directed graphs). It also analyses and allows the integration of several levels of sub graphs. A linguistic automaton identifies expressions in the texts, while the transducer associates specific labels to any words in the text. Grammars are represented by finite state automaton. Nooj is still under development and is the object of regular and constant update. Unfortunately, its online help is not completed, making it at times difficult to use. Consequently, we mainly worked with Intex, then using the migration option of graphs from Intex to Nooj to complete the work. It is also worth noting that Nooj can take into account different format of texts (.pdf, .doc) in addition to .txt.

It is difficult to make a complete evaluation of the results obtained by estimating precision, recall, noise, silence, and F-measure, which are parameters generally used within the IR community to evaluate most IR system.

At this point, it is necessary to mention that the results presented are the consequence of the recognition of the expressions expressed by grammars. In other words, a correct grammar will inevitably lead to the awaited answer and any other grammar will produce useless results. If the expected answer is not correct, the error would be on the construction of the grammar. Under theses conditions, it is appropriate to re-inspect, to correct the grammar according to the undetected expressions. This also underlines the difficulty of evaluating the quality of the obtained results since we did not have access to a well tagged corpus.

However, considering the examples illustrated in Figs. 5, 6 et 7 below, the results of this experiments are encouraging. In all these figures, Intex marks in blue the expression recognized by the grammar.

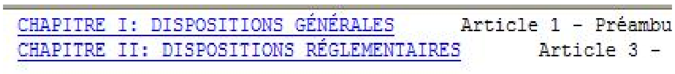

CHAPITRE I: DISPOSITIONS GÉNÉRALES Article 1 - Préambu
CHAPITRE II: DISPOSITIONS RÉGLEMENTAIRES Article 3 -

Fig. 5 Results of the detection of the title of a chapter

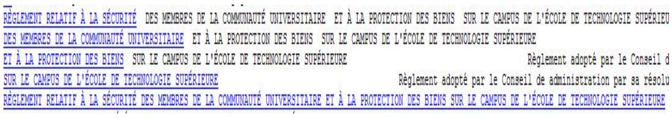

RÈGLEMENT RELATIF À LA SÉCURITÉ DES MEMBRES DE LA COMMUNAUTÉ UNIVERSITAIRE ET À LA PROTECTION DES BIENS SUR LE CAMPUS DE L'ÉCOLE DE TECHNOLOGIE SUPÉRIEU
DES MEMBRES DE LA COMMUNAUTÉ UNIVERSITAIRE ET À LA PROTECTION DES BIENS SUR LE CAMPUS DE L'ÉCOLE DE TECHNOLOGIE SUPÉRIEURE
ET À LA PROTECTION DES BIENS SUR LE CAMPUS DE L'ÉCOLE DE TECHNOLOGIE SUPÉRIEURE Règlement adopté par le Conseil d
SUR LE CAMPUS DE L'ÉCOLE DE TECHNOLOGIE SUPÉRIEURE Règlement adopté par le Conseil de administration par sa résolu
RÈGLEMENT RELATIF À LA SÉCURITÉ DES MEMBRES DE LA COMMUNAUTÉ UNIVERSITAIRE ET À LA PROTECTION DES BIENS SUR LE CAMPUS DE L'ÉCOLE DE TECHNOLOGIE SUPÉRIEURE

Fig. 6 Results of detection of the title of the document

The recognition model for titles of chapters and documents are respectively presented in Figs. 5 and 6 while Fig. 7 illustrates the result of applying the recognition model for enumeration lists.

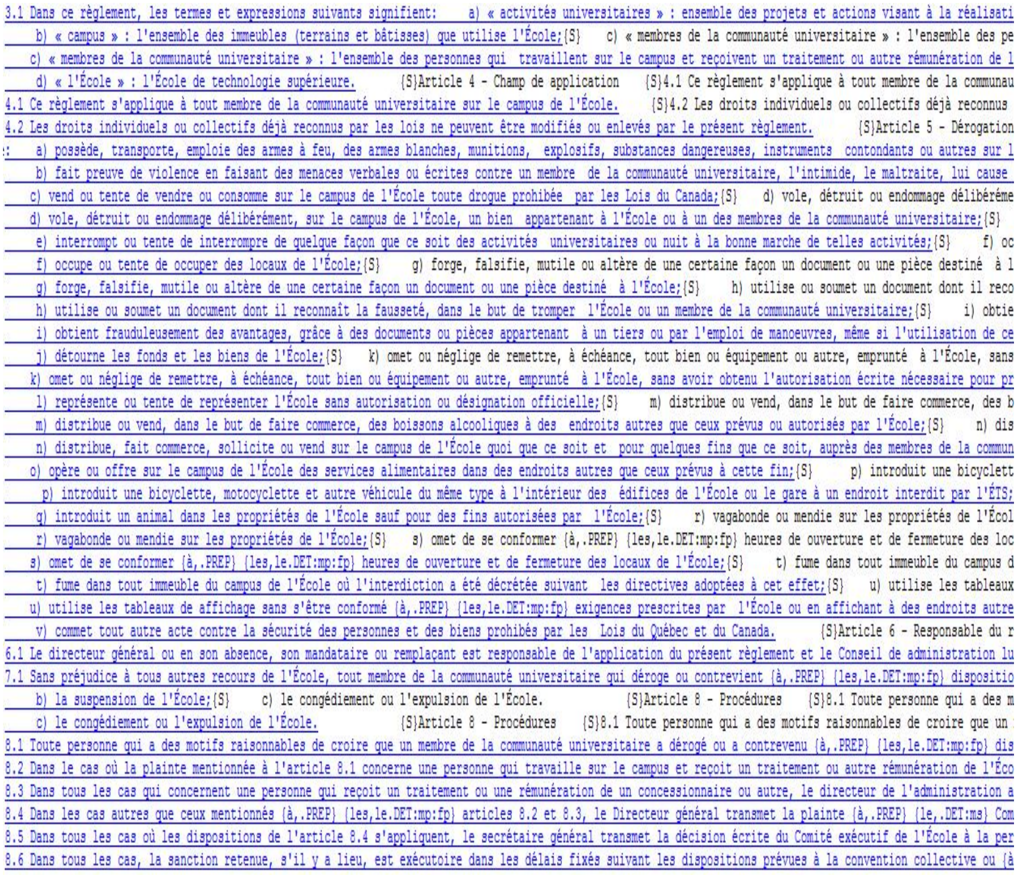
In Fig. 6, the longest expression (at the bottom) presents the title of the document as a whole. The four preceding lines show up because the title of the document was written in pieces; each underlined expression appearing on a different line.

[4] *Constitutive documents* are documents related to the existence of an organization and its legal bases.

[5] *Meeting documents* are documents related to meetings (convocation, minutes, reports)

[6] *Human resources documents* are documents related to staff management.

[7] *Documents of communication* include all types of documents being used mainly to establish and maintain internal and external relationships necessary for the development of the organization. There are internal communications (note, report/ratio, bulletin or newspaper) and external (activities of marketing, public relations with the press: press releases, booklets, catalogues, leaflet, posters, etc.).



```
3.1 Dans ce règlement, les termes et expressions suivants signifient:     a) « activités universitaires » : ensemble des projets et actions visant à la réalisati
     b) « campus » : l'ensemble des immeubles (terrains et bâtisses) que utilise l'École;{S}     c) « membres de la communauté universitaire » : l'ensemble des pe
    c) « membres de la communauté universitaire » : l'ensemble des personnes qui  travaillent sur le campus et reçoivent un traitement ou autre rémunération de l
     d) « l'École » : l'École de technologie supérieure.          {S}Article 4 - Champ de application    {S}4.1 Ce règlement s'applique à tout membre de la communau
4.1 Ce règlement s'applique à tout membre de la communauté universitaire sur le campus de l'École.     {S}4.2 Les droits individuels ou collectifs déjà reconnus
4.2 Les droits individuels ou collectifs déjà reconnus par les lois ne peuvent être modifiés ou enlevés par le présent règlement.        {S}Article 5 - Dérogation
:    a) possède, transporte, emploie des armes à feu, des armes blanches, munitions,  explosifs, substances dangereuses, instruments  contondants ou autres sur l
     b) fait preuve de violence en faisant des menaces verbales ou écrites contre un membre  de la communauté universitaire, l'intimide, le maltraite, lui cause
    c) vend ou tente de vendre ou consomme sur le campus de l'École toute drogue prohibée  par les Lois du Canada;{S}    d) vole, détruit ou endommage délibéréme
    d) vole, détruit ou endommage délibérément, sur le campus de l'École, un bien  appartenant à l'École ou à un des membres de la communauté universitaire;{S}
     e) interrompt ou tente de interrompre de quelque façon que ce soit des activités  universitaires ou nuit à la bonne marche de telles activités;{S}     f) oc
     f) occupe ou tente de occuper des locaux de l'École;{S}     g) forge, falsifie, mutile ou altère de une certaine façon un document ou une pièce destiné  à l
     g) forge, falsifie, mutile ou altère de une certaine façon un document ou une pièce destiné  à l'École;{S}     h) utilise ou soumet un document dont il reco
     h) utilise ou soumet un document dont il reconnaît la fausseté, dans le but de tromper  l'École ou un membre de la communauté universitaire;{S}     i) obtie
     i) obtient frauduleusement des avantages, grâce à des documents ou pièces appartenant  à un tiers ou par l'emploi de manoeuvres, même si l'utilisation de ce
     j) détourne les fonds et les biens de l'École;{S}    k) omet ou néglige de remettre, à échéance, tout bien ou équipement ou autre, emprunté  à l'École, sans
    k) omet ou néglige de remettre, à échéance, tout bien ou équipement ou autre, emprunté  à l'École, sans avoir obtenu l'autorisation écrite nécessaire pour pr
     l) représente ou tente de représenter l'École sans autorisation ou désignation officielle;{S}     m) distribue ou vend, dans le but de faire commerce, des b
     m) distribue ou vend, dans le but de faire commerce, des boissons alcooliques à des  endroits autres que ceux prévus ou autorisés par l'École;{S}     n) dis
     n) distribue, fait commerce, sollicite ou vend sur le campus de l'École quoi que ce soit et  pour quelques fins que ce soit, auprès des membres de la commun
    o) opère ou offre sur le campus de l'École des services alimentaires dans des endroits autres que ceux prévus à cette fin;{S}      p) introduit une bicyclett
      p) introduit une bicyclette, motocyclette et autre véhicule du même type à l'intérieur des  édifices de l'École ou le gare à un endroit interdit par l'ÉTS;
     q) introduit un animal dans les propriétés de l'École sauf pour des fins autorisées par  l'École;{S}     r) vagabonde ou mendie sur les propriétés de l'Écol
     r) vagabonde ou mendie sur les propriétés de l'École;{S}    s) omet de se conformer {à,.PREP} {les,le.DET:mp:fp} heures de ouverture et de fermeture des loc
    s) omet de se conformer {à,.PREP} {les,le.DET:mp:fp} heures de ouverture et de fermeture des locaux de l'École;{S}     t) fume dans tout immeuble du campus d
     t) fume dans tout immeuble du campus de l'École où l'interdiction a été décrétée suivant  les directives adoptées à cet effet;{S}    u) utilise les tableaux
    u) utilise les tableaux de affichage sans s'être conformé {à,.PREP} {les,le.DET:mp:fp} exigences prescrites par  l'École ou en affichant à des endroits autre
     v) commet tout autre acte contre la sécurité des personnes et des biens prohibés par les  Lois du Québec et du Canada.       {S}Article 6 - Responsable du r
6.1 Le directeur général ou en son absence, son mandataire ou remplaçant est responsable de l'application du présent règlement et le Conseil de administration lu
7.1 Sans préjudice à tous autres recours de l'École, tout membre de la communauté universitaire qui déroge ou contrevient {à,.PREP} {les,le.DET:mp:fp} dispositio
     b) la suspension de l'École;{S}     c) le congédiement ou l'expulsion de l'École.          {S}Article 8 - Procédures    {S}8.1 Toute personne qui a des m
     c) le congédiement ou l'expulsion de l'École.            {S}Article 8 - Procédures    {S}8.1 Toute personne qui a des motifs raisonnables de croire que un
8.1 Toute personne qui a des motifs raisonnables de croire que un membre de la communauté universitaire a dérogé ou a contrevenu {à,.PREP} {les,le.DET:mp:fp} dis
8.2 Dans le cas où la plainte mentionnée à l'article 8.1 concerne une personne qui travaille sur le campus et reçoit un traitement ou autre rémunération de l'Éco
8.3 Dans tous les cas qui concernent une personne qui reçoit un traitement ou une rémunération de un concessionnaire ou autre, le directeur de l'administration a
8.4 Dans les cas autres que ceux mentionnés {à,.PREP} {les,le.DET:mp:fp} articles 8.2 et 8.3, le Directeur général transmet la plainte {à,.PREP} {le,.DET:ms} Com
8.5 Dans tous les cas où les dispositions de l'article 8.4 s'appliquent, le secrétaire général transmet la décision écrite du Comité exécutif de l'École à la per
8.6 Dans tous les cas, la sanction retenue, s'il y a lieu, est exécutoire dans les délais fixés suivant les dispositions prévues à la convention collective ou {à
```

Fig. 7 Results of detection of enumeration lists in a document

Fig. 7 presents a list of all enumerations in one specific document.

Despite the difficulties already mentioned, we made an evaluation between the results obtained automatically and those obtained manually with 46 documents. The results are presented in Table I.

TABLE I
PRECISION RATE OF DETECTION OF DOCUMENT'S TITLE, ENUMERATION LIST AND CHAPTER

	Titles	Enumeration list	Chapters
Precision	66.67%	92.88%	34.72%
	31/46	1332/1434	50/114

The small success rate concerning chapters' detection is due to the fact that in our unstructured documents the title of chapter was not always preceded by the key word "chapter". For that first experiment, this only key word was essential to emphasize this structure. The automaton associated with this model can easily be modified.

Beyond these convinced results, some difficulties remain to be solved, in particular the conception of recognition models for other structures such as paragraphs and tables, both included in our future works.

As mentioned before, in order to avoid conflict between the detection enumeration lists and chapters, we associate the capital letters to the detection of expressions for chapters and sections.

I – Ecole de technologie supérieure (1) (2)
i - Ecole de technologie supérieure (2). (3)

In these two cases of enumeration, we distinguished the first form (2) from the second (3) by the capital letter for the enumeration. So (2) will be recognized as a title while (3) will be tagged as an enumeration. Of course, if the first form is preceded by the word chapter, it will be clearly identify as a chapter's title.

One of the encountered problems and not the least is the fact that we use a terminological extractor for the recognition of textual expressions. But, terminological extractors hardly worry to extract certain special characters such as white character, carriage return or line feed character. Those characters are important to us for the detection of paragraphs among others. Our idea here is to set up a program that will tag these linguistic markers whose will help us in this task of detection of the documentary structure.

The question is to know if we could build a grammar which takes into account all the aspects of the documentary structure without having to manage conflicts between them. The first idea is to proceed with a priority list, carrying out the task of extraction from generic structures (title, chapter) to more specific ones (enumeration lists).

V. Conclusion

The efficient exploitation of unstructured document, although complex, is crucial to IR. In this paper, we have defined symbolic models to recognize document's structures. The recognition models are use in the upstream of a document analysis project in software engineering that will transform natural language text to visual representation models.

Considering the fact that, in all organizations (not forgetting the Web itself), there is more unstructured documents than structured ones[8], this research is justified since it offers the foundation for a tool to transform unstructured documents into a useful and coherent material ready for IR. Furthermore, they could also be easily translated afterwards in XML format, thus facilitating their exploitation by any tools [7].

This research could also be indirectly useful from an industrial point of view since the main cost of any documentary project comes from the definition and the maintenance of document structure [11].

We intend to refine and expand the proposed models to take into account tables, among other documentary structures. These units abound in relevant information that is generally lost during IR process. Their specificities require a detailed attention [12]. As we all know, the information within a cell in a table can only be properly interpreted taking into account the column and the relevant line. Usual IR tools, extract the information within the cell independently of its column or its line, thus loosing crucial semantic information.

Finally, it is worth noting that these experiments can be easily adapted to other languages such as English since most of the models make a parsimonious use of specific tokens (e.g. "chapitre"/ "chapter", "partie"/ "part").

We are currently testing the models to verify, to what extend, we can extract each recognized units (and its following content) as a mini-document that could be, in turn, indexed and categorized.

Acknowledgment

The authors would like to thank the "Service des Archives" of the École de technologie supérieure to have given us access to the document repository of the university with a special thank to Josée Guillemette, archivist, for her valuable collaboration and interest in this project.

References

[1] P. Lyman, and H. R. Varian, "How Much Information", 2003, Retrieved from http://www.sims.berkeley.edu/how-much-info-2003 on August 30th, 2007.

[2] F. Role, and G. Rousse, "Construction incrémentale d'une ontologie par analyse du texte et de la structure des documents", in *Document numérique,* Lavoisier, 2006, Vol. 9, No 1, p. 77-91.

[3] T. Schlieder, and H. Meuss, "Querying and ranking XML documents", *Special Topic Issue of the Journal of the American Society of Information Science on XML and Information retrieval*, 2002.

[4] Y. Prie, "Sur la piste de l'indexation conceptuelle de documents. Une approche par l'annotation", in *Document Numérique, numéro spécial "L'indexation", Lavoisier,* December 2000, Vol. 4, No 162, pp. 11-35.

[5] H. Zargayouna, "Indexation sémantique de documents XML", 2005, Ph.D. Thesis, Université Paris-Sud, France.

[6] D. Kerkouba, "Une méthode d'indexation automatique des documents fondée sur l'exploitation de leurs propriétés structurelles. Application à un corpus technique", 1984, Ph.D. Thesis, Grenoble, France

[7] X. Tannier, "Recherche d'information dans les documents XML" in rapport de recherche 2006-400-007, Centre Génie Industriel et Informatique (G2I) de l'Ecole Nationale Supérieure des Mines de Saint-Etienne, France, 2006.

[8] W. Njomgue, "Le système MAID : Multi-Approches pour l'Indexation des Documents au sein de l'Intranet de Suez-Environnement", Ph.D. Thesis, 2005, Université de Technologie de Compiègne, France.

[9] S. Aït-Moktar, V. Lux, and E. Banik, "Linguistic Parsing of Lists in Structured Documents" in *Proceedings of the 2003 EACL Workshop on Language technology and the Semantic Web (3rd Workshop on NLP and XML, NLPXML-2003),* Budapest, Hungary.

[10] L. Gagnon-Arguin, and H. Vien, "Typologie des documents des organisations – De la création à la conservation", *Collection gestion de l'information, Presse de l'Université du Québec*, 2005.

[11] R. Abascal, M. Beigbeder, A. Benel, S. Calabrotto, B. Chabbat, P-A. Champin, N. Chatti, D. Jouve, Y. Prie, B. Rumple, and E. Thivant "Modéliser la structuration multiple des documents" in *Rapport d'activités 2002-2003 des recherches collectives sur la « multistructuralité » des documents,* Institut des Sciences du Document Numérique (ISDN), France, 30 September 2003.

[12] S. Douglas, M. Hurst, and D. Quinn, "Using Natural Language processing for Identifying and Interpreting Tables in Plain Text" in *Proceedings of the Fourth Annual Symposium on Document Analysis and Information Retrieval,* 1995, pages 535-546, Las Vegas, NV, USA

[8] Another possible explanation for the significant number of unstructured documents compared to structured one is due to the fact that the first function of a document is to be read by the human.