Using POS Annotations and EMdF Text Modeling for Semantic Web Search

Rodica Diaconescu[♣], Paul Hirschbühler[♣], Dan Ionescu[♣♣], France Martineau[♣], Mircea Trifan[♣♣]
Faculty of Arts[♣], School of Information Technology and Engineering[♣♣]
University of Ottawa, Ottawa, Ontario, Canada[i]

Abstract: **There is a large research effort for building a Semantic Web for complementing the current text-based web with machine understandable semantics. This paper introduces a new and complex algorithm for automatic Natural Language text exploration while exploring a corpus. The algorithm aims at searching and retrieving information from a corpus. At the same time the new algorithms can be used to store semantic based web repositories. The corpus is first explored by a morphologic and syntactic parser processing. An intermediate text corpus is structured according the Enhanced Monad dot Feature (EMdF) model. For the parsed text above an EMdF model is directly mapped and stored in an RDF database. Information retrieval on the initial text corpus is performed using queries on the RDF database. These queries resolve co- occurrences, verbs identification, and false positive elimination such that the results obtained are well-formed relations. The algorithms introduced are illustrated on examples of Semantic Web information retrieval on a medieval French corpus based on "Mémoires" of Philippe de Commynes.**

I INTRODUCTION

There are intensive research and standardization activities in regards to the passage from the present form of the Web to a global repository of machine-understandable information [13]. The exploration of this information located on a universal Web of semantic assertions, however, raises a series of open issues related to a common model onto which any prospective application has to be mapped, to the mathematical for representing the knowledge contained in the universal Web, and mainly to the exploration of it. To structure and work with the knowledge contained in the information stored into this large repository of information Description Logic [1] and related ontologies [14] were introduced and built for the hope of finding a solution to the above problem. Despite serious efforts made in the above area, search mechanisms for information retrieval capable of providing answers to queries embedding goals with meaning are not yet available. The reason of such a lack of search tools which can overpass the level of shallow semantics, is that understanding of more profound levels of the information semantics relates to the expressive power of a natural language be it English, French or any other language used to store information on the Semantic Web.

Natural Language Processing/Understanding (NLP) [20] domain is rich in methods and algorithms for mapping an NL text in symbols understandable by computers. In this direction parsing, Part of Speech (POS) tagging, syntactic bracketing, and disfluency annotation are important steps taken towards NLP automatic text understanding [3], [7], [2]. A series of methods have been proposed and used for improving the result of parsers of natural language texts to obtain a complete grammar, thus helping machines to understand them [11]. Results of applying machine learning and text classifications were obtained for the same NL understanding problem. However, due to the ambiguity of a natural language there are still unanswered questions in regards to the precision of Semantic Web search results. The existence of co-occurrences and false positives in any semantic query results are existing barriers which have to be overcome when trying to explore the information based on its meaning.

In this paper a new information retrieval and text exploration technique is introduced for the purpose of filtering out co-occurrences and false positives when exploring a text corpus. This new method combines the power of results obtained in the linguistic domain such as the morphologic and syntactic text annotation with the recent developments in the area of using ontologies and the associated Resource Description Framework (RDF) [13] [17] databases for storing the information on the Semantic Web. The Description Logic [1] helps developing robust and well verified ontologies the elements of which can be marked on the explored text. In addition to the above a powerful text modeling technique called EMdF is used to index uniquely the text and map it in syntactic objects eventually stored in an RDF database. First a text corpus built using the "Mémoires" of Philippe de Commynes is analyzed morphologically and syntactically [9], the results of this phase being stored in a text database using the EMdF text modelling technique [16]. The modeled text is mapped to an RDF database using Sesame, an open source RDF database with support for RDF Schema inference and querying language [4]. The powerful query language called SeRQL allows performing "select" queries against triplet relations (subject, predicate, object). These queries resolve issues related to co-occurrences, verbs identification, and false positive elimination such that the results obtained are well-formed relations. In Section II a complete architecture for a semantic repository and information retrieval system (SIRRS) is given. In Section III the MCRI Corpus [9] and the process of morphologic tagging and syntactic annotations is described using examples from "Mémoires" of Philippe de Commynes. In Section IV the definition and usage of ontologies for the purpose of Semantic Web search is given and exemplified on the French nobility as in the time of Louis XI, the king of France. In section V an innovative technique for mapping text model expressed using EMdF to RDF is introduced. In section VI the power of the RDF representation of the morphologic and syntactic annotated text in conjunction with the EMdF model is described in terms of the SeRQL query language as used with the Sesame RDF database. It is shown on examples that the combination of POS tagging, EMdF, and RDF storage provides the power to eliminate co-occurrences and the false positives, which are major hurdles while exploring the Semantic Web for a more profound semantic. Finally in Section VII conclusions are provided.

II SEMANTIC INFORMATION REPOSITORY AND RETRIEVAL SYSTEM (SIRRS)

The overall architecture of the SIRR system is given in the Figure 1 below. SIRRS contains two major paths and associated processes.

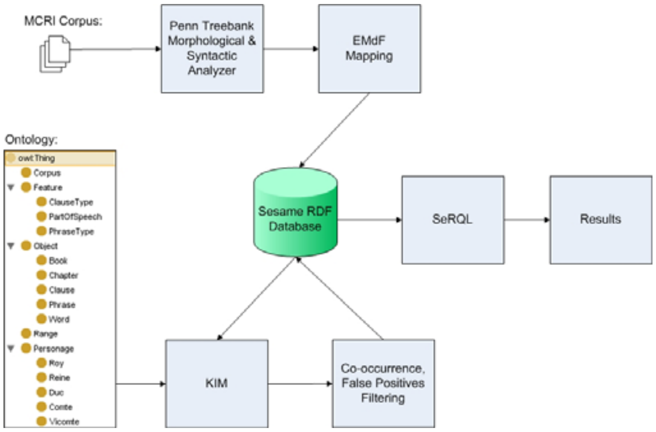


Figure 1. SIRRS Architecture

One path is used for Information Storage where a given text is first analyzed from a morphology and syntax rules point of view and tagged accordingly [11]. The tagged text is then mapped via a monadic dot feature format [16] into an RDF database [4]. The other path of the system consists of the text exploration process guided by an ontology [14] which for this time was defined and built a-priori. The ontology in this case was introduced using Protégé [18], a publicly available software capable of describing it in an OWL file [19]. Using Protégé concepts related to the French nobility as encountered in the corpus based on the medieval masterpiece of Philippe de Commynes: "Mémoires" [5] were introduced in the above ontology. This Corpus was used in the MCRI project as described in Section III. Based on the above ontology the corpus was explored using KIM for identifying entities, while the EMdF text stored in the RDF database is used to detect co-occurrences. Using SeRQL and the grammar structure provided by syntactically annotated text corpus false positives are identified and filtered out of the information retrieval results.

III THE MCRI CORPUS: AN OVERVIEW

The Project «Modeling Change: The Paths of French[1]» (MCRI) [10], has as primary goal to build a French corpus representative of the complexity of French varieties, both historically and geographically. It covers the period of time between the XII-th to the XVIII-th century and gathers texts from northern regions of France, and from French colonization in North America (New-France). The French Corpus comprises 2,500,000 words, and has 3 subdivisions: (i) Medieval Corpus (Old and Middle French), (ii) Renaissance Corpus and (iii) Colonization of North America corpus (New-France). The medieval part of the Corpus consists of 1 million words, from which 250,000 are already morphologically tagged and 51,100 are syntactically annotated. The texts are initially subjected to a basic encoding, in order to standardize the formats (indication of the title, author, date, origin, type of text, beginning and end of text, etc.) according to the Text Encoding Initiative standard protocol [15] and they will be available on the Project site using ARTFL software [12].

The MCRI corpus is morphologically and syntactically annotated using methods and tools developed for the Penn-Helsinki Parsed Corpus of Middle and early Modern English [11]. The goal of building the French Corpus above is related to the development of a linguistic change model based on statistical profiles of the linguistic changes.

A. The MCRI Corpus morphologically tagged

The first step in building a parsed corpus is the Parts of Speech (POS) tagging. At the base of the POS tagging system of MCRI Corpus is the Transformation Based Learning paradigm (TBL) [3], refined by the Fast Transformation-Based Learning Toolkit (fnTBL) [7]. The TBL starts with an un-annotated text as input which passes through the 'initial state annotator', when it assigns tags to the input in a heuristic fashion. The output is a temporary corpus, which is then compared to a goal corpus, i.e., the correctly annotated training corpus. For each time the

[1] «Modeling the change: the paths of French» 2004-2009 is subsidized by the SSHR Canada, projects of the Major Collaborative Research Initiatives (MCRI), project Director F. Martineau.

temporary corpus is passed through the learner, the learner produces one new rule, the single rule that improves the annotation the most compared with the goal corpus. By this process, the learner produces an ordered list of rules and the temporary corpus is replaced with the analysis that results when the rules are applied to it. At the end, texts are submitted to manual correction.

For assigning morphological tags to the input corpus, which in this paper is the MCRI, Commynes' "Mémoires", a compositional tag set based on PPCME [11] was used. This tag set was also elaborated in accordance to the rich French morphology and on results of other French tagging software such as *Grace* [8] and *Cattex* [6].

The output was compared with the goal text a 20,000 words part of "Cent Nouvelles Nouvelles Anonymes" [9]. The POS tagging of the goal text was the result of POS tagging guideline for the MCRI Corpus, conceived after a thorough study on morphological particularities of French [9].

The MCRI guideline tag set employs from 1 to 3 fields (each one materialized by 1-3 capital letters), easily usable by the syntactic software (i.e., *filles*/NCPL=Noun, Commun, Plural "girls"). Tags specific for French, as (DF) for the Partitive Determinant (i.e., *je bois du*/DF *vin* "I drink wine"), or (L) for the auxiliary *aller* "go" expressing the future have been added. The MCRI tag set also indicates the name number, (i.e., *les*/D *petites/*ADJ *filles/*NCPL "the small girls"), knowing that the noun, as core of the nominal group, will transmit this label to the syntactic phrase NP. A morphological tagging sample on a paragraph drawn from Commynes' "Mémoires" is given below.

```
Car/CONJO les/D biens/NCPL ne/CONJO les/D
honneurs/NCPL ne/NEG se/PRO despartent/VJ
point/ADVNEG à/P l'/D appetit/NCS de/P
ceulx/PRO qui/WPRO les/PRO demandent/VJ
./PONFP
```

"Because the benefits and the honours are not allocated by the wish of the ones who are asking for them".

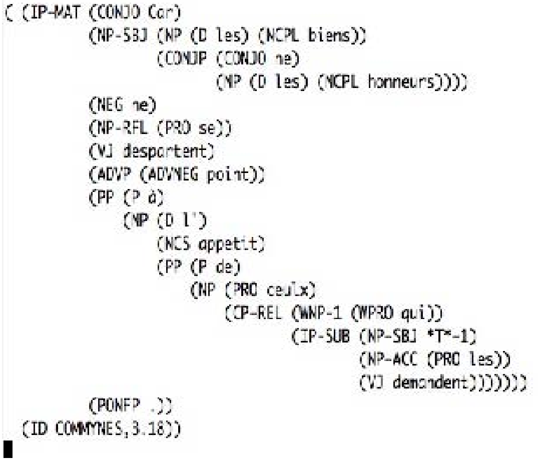
Fig. 2. Morphologically Annotated Sample

B. The MCRI Corpus syntactically annotated

The MCRI syntactic bracketing system adopts the general principles of analysis and component representation of the Middle English Corpus [11] and refined by Bikel's parsing software [2], which incorporates a consistent treatment of related grammatical phenomena. However, the automatic syntactic annotation has to be corrected by hand. This stage adds corrections and information: it changes syntactic tags, it appends subcategory information, it changes attachment level or it breaks up sentences when required. The final editing is partially automated.

The adoption of the system of syntactic bracketing of Middle English for a French corpus required the adaptation of the system existing to a grammar which differs in many regards from the Middle English's one. The final version of the MCRI syntactic bracketing guideline could be considered complete only after the corpus will have been completely parsed, since new situations add continuously. The MCRI syntactic bracketing process was initially developed starting from the "Mémoires" of Philippe de Commynes, as formation text, and corrections have been added manually.

The manual syntactic labelling of the text annotated is carried out using a tool of annotation functioning of linux, which is included in GNU Emacs. A syntactic bracketing of a token containing a token (IP-MAT, i.e., main clause with/without subordinates) with its embedded relative clause (CP-REL) is given in Figure 3.



```
( (IP-MAT (CONJO Car)
          (NP-SBJ (NP (D les) (NCPL biens))
                  (CONJP (CONJO ne)
                         (NP (D les) (NCPL honneurs))))
          (NEG ne)
          (NP-RFL (PRO se))
          (VJ despartent)
          (ADVP (ADVNEG point))
          (PP (P à)
              (NP (D l')
                  (NCS appetit)
                  (PP (P de)
                      (NP (PRO ceulx)
                          (CP-REL (WNP-1 (WPRO qui))
                                  (IP-SUB (NP-SBJ *T*-1)
                                          (NP-ACC (PRO les))
                                          (VJ demandent)))))))
          (PONFP .))
  (ID COMMYNES,3.18))
```

Fig 3. Syntactically annotated sample

The tool for seeking linguistic information in the annotated corpus is "Corpus Search", written in Java and can be run in a Linux Mac, Unix, or Windows operating system. Its basic functions are intuitive, and it can seek on its results. The search results are recorded in a file output. A possible search on null subjects of the MCRI parsed corpus (Commynes) is given in Figure (4a) below. Its results are as in (4b).

(4) a. node: IP*
query: ((IP* idoms NP-SBJ*) AND (NP-SBJ* iDoms *proimp*|*con*|*pro*))
(4) b. source files
commynes.mjc
hits/tokens/total
990/897/2472

Figure 4. A corpus query and its results

IV. USING ONTOLOGY TO DESCRIBE THE DOMAIN

In order to make statements and make queries about a subject domain, there is a need for a conceptualization of that domain. This conceptualization requires naming and describing entities that might exist in that domain together with their relationships. The result is a vocabulary for representing and communicating knowledge about the above domain. When this conceptualization is explicit it is called ontology. An ontology is a logical theory to relate the intended meaning of a formal vocabulary. The ontology includes hierarchies of object classes (taxonomies) and their relationships. Ontology is one of the basic steps towards a Semantic Web. It contains the meaning that a computer system ascribes to terms in that vocabulary. However, the process of constructing an ontology is rather difficult and time consuming. Attempts to automate the process are made [21], however, a human supervision is mandatory. For the purpose of the research presented in this paper the ontology of the French nobility as contained in the "Mémoires" of Philippe de Commynes is a-priory defined and built. In the Figure 5 below, the ontology built to illustrate the new method built for information retrieval is given. As a good illustration example considered was the hierarchy of nobility described in "Mémoires" of Philippe de Commynes. The ontology is captured in the OWL language and an example for the Roy class is provided below.

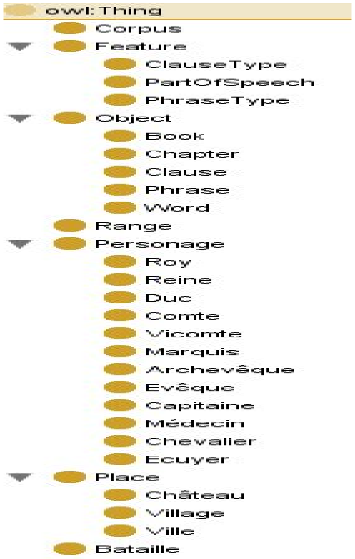


Figure 5. The Ontology of the French nobility as depicted in "Mémoires" by Philippe de Commynes

```
<owl:Class rdf:ID="Roy">
     <rdfs:subClassOf rdf:resource="#Personnage"/>
</owl:Class>
<owl:ObjectProperty rdf:about =
        "#hasLifeSpan">
        <rdfs:domain>
   <owl:Class rdf:about="#Personnage"/>
        </rdfs:domain>
        <rdfs:range rdf:resource="#Range"/>
</owl:ObjectProperty>

<Roy rdf:ID="r_15">
<hasName rdf:datatype =
    "http://www.w3.org/2001/XMLSchema#string"
    > Louis X|</hasName>
    <hasNobilityTitle rdf:datatype =
"http://www.w3.org/2001/XMLSchema#string"
    > Roy </hasNobilityTitle>
</Roy>
```

Figure 6. Excerpt of the OWL file generated by Protégé

There are two types of classes in the ontology created as described above: classes that deal with the EMdF text model: Corpus, Feature, Object Range and their subclasses and classes ascribed to entities from the text like: Personnage, Place, Bataille ("battle") and their subclasses. The EMdF related classes and their properties are described in the next section.

For the purpose of demonstrating the information retrieval technique proposed in this paper, this ontology describing the nobility ranks, their places, and the battles will be used to find answer to questions such as who were the relatives of Louis XI who had properties on the Loire region for example, and who were killed in battles.

V MAPPING A MORPHOLOGIC AND SYNTACTIC ANNOTATED TEXT TO EMDF AND TO RDF

The Resource Description Framework was recommended by W3C as a standard for the notation of meta-data on the World Wide Web. The RDF Schema [22] provides the means to specify a vocabulary and to model object structures. These techniques were built for enabling the enrichment of the Web with machine understandable or processable semantics, thus giving birth to the so called Semantic Web. The most primitive concept in RDF is the triple object-attribute-value A(O; V) where O is the object, A is the attribute and V is the value, or in other words a labelled edge relationship: [O] – A→ [V] where [A] and [V] are nodes and –A→ is the labelled edge. This notation is useful because any object from one triple can play the role of a value in another triple, which allows composition of nodes and edges allowing, eventually, building nested and/or chained graphs. Thus information stored in a series of Web sites or the same site can be linked through relationships expressing a deeper semantic between the elements of the information such in the following example:

```
hasName
('http://www.mrci.ca/commynes/phillippe',"Ph
ilippe de Commynes ")
hasWritten
('http://www.mrci.ca/twain/mark',
'http://www.books.org/ISBN0872491307')
title
('http://www.books.org/ISBN0872491307',
" Mémoires")
```

Figure 7. Using RDF to represent information on the Semantic Web

As relations such as hasName, hasWritten and title are attributes on the labelled edges which are used to encode the relationships among the content of various web sites, it is clear that applying a proper graph search technique the queries for the more profound information contained on various web sites can be done.

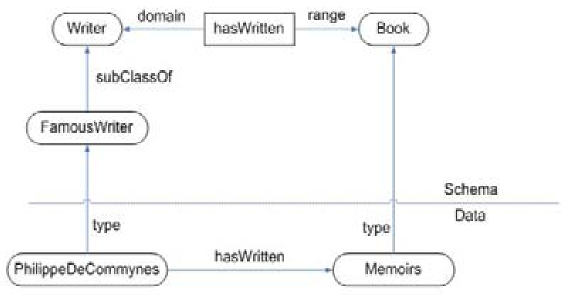


Figure 8

The previous schema as shown in Figure 8 above maps into an RDF set of triplets:

```
(type Book Class)
(type Writer Class)
(type FamousWriter Class)
(subClassOf FamousWriter Writer)
(type hasWritten Property)
(domain hasWritten Writer)
(range hasWritten Writer)
(type PhilippeDeCommynes FamousWriter)

(type ISBN0872491307 Book)
(hasWritten PhilippeDeCommynes
ISBN0872491307)
```

Figure 9. The RDF representation of the fact that Philippe de Commynes is a Famous Writer, who has written the "Mémoires" catalogued under the ISBN0872491307

The Extended Monads-dot-Features (EMdF) model is based on the definition of monads, textual objects e.g., words, phrases, etc., with features and the relations between these objects. A monad is an absolute, indivisible position in a text, which corresponds to a word and is represented by an integer.

TABLE 1
Using monads to model text

hasMonad	Word		Phrase	Phrase	Sentence
/hasRange	hasSurface	hasPartOf Speech	hasPhraseType	hasPhrase Type	hasSentence Type
1	Je	PRO	NP-SBJ		IP-MAT
2	me	PRO	NP-ACC		
3	suys	EJ			
4	mys	VPP			
5	en	P		PP	
6	ce	D	NP-ACC		
7	propoz	NCS			

Monads are grouped as objects, e.g. words are grouped into sentence objects, sentence objects are grouped into paragraph objects etc. Each object class has a number of features, e.g. part of speech for each word. No restrictions are imposed on the user as to how to relate objects to objects. In TABLE 1 an EMdF database corresponding to one possible analysis of the sentence in Figure 10 below is given.

Je me suys mis en ce propoz

```
( (IP-MAT (NP-SBJ (PRO Je))
          (NP-ACC (PRO me))
          (EJ suys)
          (VPP mys)
          (PP (P en)
              (NP (D ce) (NCS propoz)))

(ID COMMYNES,67.883))
```

"I have placed myself into this topic"

Figure 10. The MCRI sentence to be represented according to the EMdF model in TABLE 1

There are three object types: Word, Phrase, and Sentence each with the features listed in the table heading. The monad-granularity is Word. The text is encoded by the hasSurface feature on Word object type. Examples of other types of properties are provided in Figure. 11.

hasClauseType : Clause -> ClauseType
hasPhraseType : Phrase -> PhraseType
hasPartOfSpeech : Word -> PartOfSpeech
hasSurface : Word -> String
hasRange: Clause, Document, Phrase -> Range
hasLast : Range -> Int hasFirst : Range -> Int
hasDocument : Corpus -> Document
hasMonad : Object -> Int

Figure 11. Encoding other types of properties while modeling texts with EMdF

VI USING AN RDF DATABASE AND ITS QUERY LANGUAGE FOR INFORMATION RETRIEVAL

Sesame is an open source Java framework for storing, querying and reasoning with RDF and RDF Schema. It can be used as a database for RDF and RDF Schema, or as a Java library for applications that need to work with RDF internally. Examples of tasks to be performed with Sesame are to read a big RDF file, to find the relevant information for the application, and use that information. Sesame provides the necessary tools to parse, interpret, query and store all this information, embedded in the application or in a separate database or even on a remote server.

SeRQL ("Sesame RDF Query Language", pronounced "circle") is a new RDF/RDFS query language for Sesame. It combines the best features of other (query) languages (RQL, RDQL, N-Triples, N3) and adds some of its own. Some of SeRQL's most important features are:

- Graph transformation.
- RDF Schema support.
- XML Schema datatype support.
- Expressive path expression syntax.
- Optional path matching.

The following SeRQL statement is used to query all pronouns from IP-MAT type of sentences, from the database:

```
  SELECT
  W
  FROM
  {W} rdf:type {ex:Word},
  {W} ex:hasPartOfSpeech {ex:PRO},
  {W} ex:hasMonad {M},
  {S} rdf:type {ex:Sentence},
  {S} ex:hasRange {R},
  {R} ex:hasFirst {F},
  {R} ex:hasLast {L}
WHERE
  L < MAND M < F
USING NAMESPACE
ex = <http://www.owl-ontologies.com/Ontology1160580224.owl#>
```

Figure 12 A Sesame query language (SerQL) example

The following example illustrates how co occurrences of “Personnages” and “Batailles” within sentences can be obtained.

```
SELECT
    P, B, S
FROM
    {P} rdf:type {ex:Personnage},
    {P} ex:hasRange {RP},
    {RP} ex:hasFirst {RPF},
    {RP} ex:hasLast {RPL}
    {B} rdf:type {ex:Bataille},
    {B} ex:hasRange {RB},
    {RB} ex:hasFirst {RBF},
    {RB} ex:hasLast {RBL}
    {S} rdf:type {ex:Sentence},
    {S} ex:hasRange {R},
    {R} ex:hasFirst {F},
    {R} ex:hasLast {L}
  WHERE
    ( L < RPF AND RPL < F ) AND
    ( L < RBF AND RBL < F )
  USING NAMESPACE
    ex = <http://www.owl-
  ontologies.com/Ontology1160580224.owl#>
```

Figure 13 Resolving for co-occurrences using morphologic and syntacti text annotation, EMdF text modeling, RDF mapping and SerQL

To filter out false positives through SeRQL information retrieval method introduced in this paper, the following example given in Figure 13 and 14 illustrate the procedure. The example considers the verb "aller" (“go”) which is tagged as a conjugated future auxiliary LJ or as a conjugated movement verb as VJ.

```
( (IP-MAT (CONJO et)
         (NP-SBJ *con*)
         (LJ va)
         (VX compter))
        (NP-ACC (NP-ACC (D ces)
                  (NCPL nouvelles)
         (PP (P ^de)
              (NP (D ^le) (CODE  {TEXT:du})
                   (NPRS Liége))))
 (PON ,) ………………………
(ID COMMYNES,134.1849))
```

“and (he) will tell those stories of Liège”

Figure 13. The verb « aller » as an auxiliary

```
( (IP-MAT (CONJO Et)
           (VJ va)
           (ADVP-TMP (ADV tousjours))
           (NP-SBJ (Q quelque)
                   (NCS humblet)
           (CP-REL (WNP-1 (WPRO qui))
                   (IP-SUB (NP-SBJ *T*-1)
                           (VJ va)
                           (PP (P à)
                                (NP (NCS part))))))
(PONFP .))
(ID COMMYNES,92.1237))
```

“And walks always some unfortunate who goes aside”

Figure 14. The verb “aller” as a main verb

To filter the sentences with the desired meaning the tagging information is taken into account and modeled using the Resource Description Framework and introduced in the Sesame database.

```
SELECT
  W, S
FROM
  {W} rdf:type {ex:Word},
  {W} ex:hasPartOfSpeech {ex:VJ},
  {W} ex:hasMonad {M}
  {S} rdf:type {ex:Sentence},
  {S} ex:hasRange {R},
  {R} ex:hasFirst {F},
  {R} ex:hasLast {L}
WHERE
    L < MAND M < F
USING NAMESPACE
ex = <http://www.owl-
ontologies.com/Ontology1160580224.owl#>
```

Figure 15. False positives are eliminated using the morphologic and syntactic annotation which distinguishes the function of the verb.

This procedure above can be used in many of the specific searches as for example for the verb "vouloir" (want) which can also can be tagged as a conjugated modal (followed by anther verb) MDJ or as a conjugated main verb VJ. Clearly, this procedure introduced above entails a considerable work and efforts for creating a correctly annotated corpus, for mapping the information contained in an annotated corpus in an RDF database, and for the EMdF text modeling in the same RDF database. However, some of the parts of the workflow of an automated process which leads to a search of information based on profound semantic queries can be programmed. These works will make though the subject of future paper.

VII CONCLUSIONS

This paper showed that a combination of results obtained in apparently disjoint research domains - the linguistic and grammar theory of natural languages and the research and standardization efforts related to the Semantic Web - can be used for overcoming hurdles in the process of information retrieval from the Semantic Web when more profound semantic of the information is sought. The methodologies, techniques and methods related to the Part of Speech annotation, text modeling, and meta-data of RDF databases combined together can make the search mechanisms more powerful, and the query results more accurate. It was shown that using the information embedded in a morphologically and syntactically annotated corpus, the power of a text model-in the present paper the EMdF,- and the query power and flexibility of RDF databases some annoying query artefacts such as the co-occurrences and the false positives can be filtered out. As such, an increase in the meaning and accuracy of the results is obtained. A Semantic Information Repository and Retrieval System (SIRRS) was built and its principle and algorithms proven correct. The SIRRS processes can be automated. This will make the subject of further papers.

AKNOWLEDGMENT

The research was supported in part by the NSERC Grant A-16800 and by the SSHREC MCRI Grant 412-2004-1002.

REFERENCES

[1] F. Baader, D. Calvanese, D. McGuinness, D. Nardi, P. Patel-Schneider, “The Description Logic Handbook”, 2003.
[2] D. Bikel, Multilingual Statistical Parsing Engine http://www.cis.upenn.edu/~dbikel/software.html.
[3] E. Brill, “Transformation-Based Error-Driven Learning and Natural Language”, *Computational Linguistics,* vol. 21, pp. 543-566 December 1995.
[4] J. Broekstra, A. Kampman, F. van Harmelen, ”Sesame: A Generic Architecture for Storing and Querying RDF and RDF Schema”, 2002
[5] Ph. De Commynes: “Mémoires”, ed. J. Calmette, Paris, “Les Belles-Lettres”, 1964.
[6] ”Étiquetage morpho-syntaxique CATTEX”, http://bfm.enslsh.fr/IMG/html/Principes_d_etiquetage_ de_la_BFM.html.
[7] R. Florian and G. Ngai, “Fast Transformation-Based Learning Toolkit”, 2001, http://nlp.cs.jhu.edu/~rflorian/fntbl/index.html.
[8] ”Format de description lexicale pour le français”, GRACE, http://aune.lpl.univ-aix.fr/projects/grace/.
[9] F. Martineau, C.R. Diaconescu and P. Hirschbühler, (in press), “Le Corpus Voies du français: de l’élaboration à l’annotation”, *Le Nouveau Corpus d'Amsterdam*, in P. Kunstmann et A. Stein (réd.) Stuttgart: Steiner. 2007.
[10] ”Modéliser le changement: les voies du français” “Modeling French: the paths of French”. http://www.voies.uottawa.ca/
[11] “Penn Parsed Corpora of Middle English and Modern English texts”, http://www.ling.upenn.edu/hist-corpora/
[12] ”The Project for American and French Research on the Treasury of the French” http://humanities.uchicago.edu/orgs/ARTFL/.
[13] “Semantic Web”, http://www.w3.org/2001/sw/
[14] S. Staab, R. Studer, “Handbook on ontologies”, 2004
[15] “Text Encoding Initiative standard protocol” http://www.tei-c.org/.
[16] U. Petersen, “Emdros -- a text database engine for analyzed or annotated text”, 2004.
[17] Shelley Powers: “Practical RDF” O’Reilly, 2003
[18] http://protege.stanford.edu/
[19] http://www.w3.org/TR/owl-features
[20] L. Quintano, S. Abreu, I. Rodrigues:” Relational Information Retrieval through Natural Language Analysis” Lecture Notes in Computer Science-Vol. 2543/2003 -Web Knowledge Management and Decision Support: 14th International Conference on Applications of Prolog, INAP 2001, Tokyo, Japan, October 20-22, 2001.
[21] Y. Sure, A. Gomez-Perez, W. Daelemans, M-L. Reinberger, N.Guarino, N.F. Noy:”Why Evaluate Ontology Technologies? Because it works” IEEE Intelligent Systems, vol.19, issue 4, July 2004, pp 74-81
[22] http://www.w3.org/TR/2000/CR-rdf-schema-20000327/
[23] B. Popov, A. Kiryakov, D. Ognyanoff, D. Manov, A. Kirilov KIM - A Semantic Platform For Information Extraction and Retrieval Journal of Natural Language Engineering, Vol. 10, Issue 3-4, Sep 2004

i The list of authors is in alphabetical order, not in the order of their contribution to this paper, for example Section III was produced by F. Martineau, R. Diaconescu, and P. Hirschbühler. Section II, IV, V, VI and VII were contributed by M. Trifan, and D. Ionescuu.