

# A step toward semantic indexing of an Encyclopedic corpus

**Philippe Alcouffe<sup>†</sup>, Nicolas Gacon<sup>†</sup>,  
Claude Roux\*, Frédérique Segond\***

<sup>†</sup> Hachette Multimédia  
11 rue de Cambrai, 75019 Paris, France  
{philippe.alcouffe, nicolas.gacon}@hachette-multimedia.fr  
\* Xerox Research Centre Europe  
6 chemin de Maupertuis, 38240 Meylan, France  
{roux, segond}@xrce.xerox.com

## Abstract

This paper investigates a method for extracting and acquiring knowledge from Linguistic resources. In particular, we propose an NLP based architecture for building a semantic network out of an XML on line encyclopedic corpus. The general application underlying this work is a question-answering system on proper nouns within an encyclopedia.

## 1. Introduction

With the increase of demand in information access together with the increase of the volume of information in multimedia (online, offline), the use of indexing and tools for intelligent search becomes crucial. It is now well known that query-answering systems give more precise results when using an abstract and logical graphical representation of the knowledge expressed in the searched text.

Until now, technologies used in reference CD-ROM industry were mostly oriented towards full text search, with the limits linked to the lack of independence between the logical representation of the information and its morphosyntactic representation in the text. A first step towards the thematic indexing of the Hachette-Multimedia Encyclopedia (Hachette, 2000) has been achieved using a tree of 3500 themes. This tree indexes the encyclopaedia via a logical and abstract representation of the information. This work has been one of the most innovative of Hachette Multimedia. Although thematic indexing allows a better abstract representation compared to full text indexing, it is still difficult to encode all relations between units (cf. (Alcouffe, 98) ).

In this paper is in line with previous works on query answering (e.g. CoBrain and Murax (Kupiec, 1993)) and presents the first steps of the joint CIRCE project, which goal is to build an NLP based architecture dedicated to the automatic semantic indexing of the Hachette Encyclopedie.

Because the encyclopedia has been written in order to describe founding principles of knowledge on one hand, and the links between those units on the other one we start from the hypothesis that an Encyclopedic corpus is a description of a relevant and prototypical set of semantics links.

## 2. Encyclopedic corpus

The Hachette Encyclopedia 2000 contains more than 120 000 articles and 17 000 media. It corresponds to a corpus of more than 23 million words. For the articles, the logical representation is XML (Extensible Mark up Language), a mark-up language whose particularity is to identify the information regardless of its presentation (title, italic...). Using XML makes it possible to create as many tags as one wants, which means to integrate new semantic markers within the very same tags.

Using XML turns the encyclopedic corpus into a structured knowledge data-base.

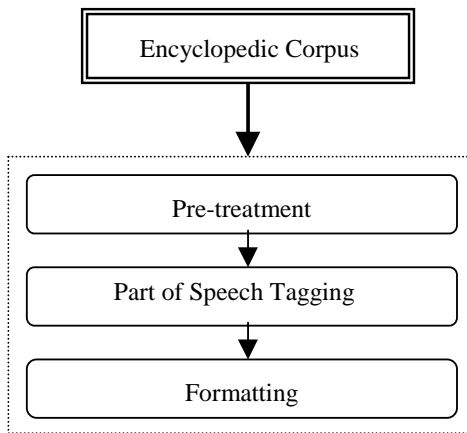
## 3. Feasability study

Having a predicate dictionary is a prerequisite of indexing any text: it is necessary to know which are the relevant relations that will be used to index any text.

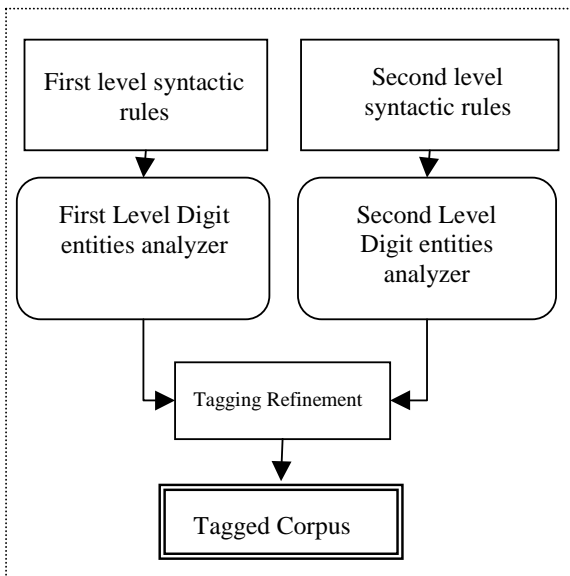
In next sections, we present an extraction methodology and illustrate it with the example of one specific semantic link and its arguments. The association of an event and a predicate is the starting point of our work.

**Process step: (schema)**

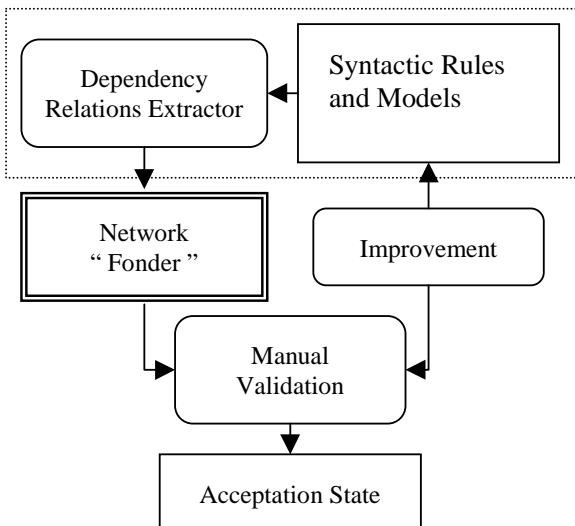
*Step one: Preliminary tagging*



*Step two: Contextual study of numbers*



*Step three: Identification of predicates and arguments*



**3.1. Preliminary tagging**

We tagged the entire Encyclopedic corpus (23 million words) with the XeLDA software (XeLda, Xerox). This allowed us to underline various kinds of encoded data (*step one*). This first step in the information extraction process permitted the shaping of the whole encyclopedia articles for a more precise study on the figures.

*morphosyntactic representation*

« Chef religieux dès 1885, il fonda, en 1906, la ligue pan musulmane de l’Inde »

“Religious leader from 1885, he founded in 1906, the pan musulmane league of India.”

Graphie	Lemme	Cat	Contexte	Pos
Chef religieux	_VIDE	NP	SENTENCE_2	1
dès	dès	PREP	SENTENCE_2	2
1885	1885	DATE_SUR	SENTENCE_2	3
,	,	CM	SENTENCE_2	4
il	il	PRON	SENTENCE_2	5
fonda	fonder	VERB_P	SENTENCE_2	6
,	,	CM	SENTENCE_2	7
en	en	PREP	SENTENCE_2	8
1906	1906	DATE_SUR	SENTENCE_2	9
,	,	CM	SENTENCE_2	10
la	le	DET_SG	SENTENCE_2	11
ligue panmusulmane de l’Inde	_VIDE	NP	SENTENCE_2	12

**3.2. contextual study of the number**

This analysis contributed to build a set of syntactic rules that disambiguates and organizes the encoded data into a hierarchy. XeLDA does not give the possibility to organize the different kinds of numbers, as it attributes the same tag NUM to all following cases:

- a. one, two...
- b. I, XVI...
- c. 13.5, 52 %,...
- d. 15 mars 1915,...

Therefore we created new grammatical categories in order to restrain the search field to numbers.

- NUM\_LETTER (one, two,...)
- NUM\_ROM (XVI,II,...)
- NUM\_SURE (1.7, 35 %,...)
- NUM
- DATE\_SURE

This study on the syntactic context of the figures on the sole basis of grammatical categories enabled us to reduce the ambiguities due to the XeLDA tagging up to 65%.

Examples of rules:

Cat_-1	Cat_1	Deduction
SENT	NOUN_PL	NUM_SUR
NOUN_SG	PUNCT	DATE_SUR

...(mort en mars+NOUN\_SG 1995+NUM)+PUNCT,...

thus by applying the second rule, we obtain the following tagging:

...(mort en mars+NOUN\_SG 1995+DATE\_SURE)+PUNCT,...

### 3.3. Identification of predicates and arguments

We then extract a set of semantic links that includes a date argument, and gather those relations together using synonymy classification (about 30 classes have been defined).

- *To Found = {create, elaborate, erect, build, construct, institute}*

It is worth noticing that it is the first study of this kind conducted on a tagged French encyclopedic corpus.

In order to construct the extraction process, the identification with arguments and to test the methodology we chose the semantic relation: "FONDER" ("TO FOUND"). The methodology is based on two main steps:

- first, building syntactic rules for identifying relevant arguments (dates, places)
- second, link these arguments to the predicates through roles (agent).

We start with the hypothesis that prepositions provide salient information for the recognition of predicate arguments. In fact, prepositions "in" or "to" often mark temporal information while prepositions "from" and "with" comprise information about the agent .

The search of the object complement is based on the syntactical nature of the predicate that determines the location of the object into the syntactical structure of the sentence and on the use of an elimination process of units recognized during the previous stage.

Each relevant component is identified by its functional nature (PERS,CM,...) within a sentence and by its localization ([x]). Below are each of the successive steps of the process based on the previous example.

SENTENCE\_2 [0] PERS *Agha Khan III*  
 SENTENCE\_2 [6] FONDER *fonda* NP\_D [6]  
 SENTENCE\_2 [8] EN\_DATE *1906* [9]  
 SENTENCE\_2 [4] CM  
 SENTENCE\_2 [7] CM  
 SENTENCE\_2 [10] CM  
 SENTENCE\_2 [12] END  
 SENTENCE\_2 [7] OBJET , *en 1906, la ligue pan musulmane de l'Inde* [12]

At the first step : the term « to found » indicates that the sought element is on the right of the verb (NP\_D);

thus, it takes all the elements which are between the position 7 and the last one (12, END). A date introduced by the preposition « in » is marked as well as all the commas (CM).

*The successive steps of the elimination process*

[7] CM  
 [7] OBJET , *en 1906, la ligue pan musulmane de l'Inde* [12]  
 →[8] OBJET *en 1906, la ligue pan musulmane de l'Inde* [12]

The analyzer compares the positions of the objects with the position of the various elements identified before.

[8] EN\_DATE *1906* [9]  
 [8] OBJET *en 1906, la ligue pan musulmane de l'Inde* [12]  
 →[10] OBJET , *la ligue pan musulmane de l'Inde* [12]

[10] CM  
 [10] OBJET , *la ligue pan musulmane de l'Inde* [12]  
 →[11] OBJET *la ligue pan musulmane de l'Inde* [12]

From now on there is no more element to substitute in the object, the analyzer stops and goes back to the last modified object.

In 92% of the studied cases, the extraction of the arguments of the relation "FONDER" has been correct.

Such relation can then be integrated into the CD-ROM of the encyclopaedia to be queried in Natural Language.

### 3.4. conclusion of the feasibility study

Moreover, such a net is crucial to NLP algorithms: the temporal anaphora "la chute du mur de Berlin" ("The destruction of Berlin wall"), identified with a unit, would allow to date correctly information in a sentence such as: "Au lendemain de la chute du mur de Berlin" ("the following day of the Berlin wall destruction").

The extracted semantic relations with arguments is currently incorporated into the NLP search module of the encyclopedia. Thus when we asked "who instituted the pan musulmane league of India", the answer will be "Agha Khan founded the pan musulmane league of India" and the user will be presented with this view of information in the text of the relevant article. Encouraged by this first results we decided to go further and defined a the CIRCE project, a joint research project between Hachette-Multimedia and the Xerox research Centre Europe.

## 4. CIRCE: first experiments

The main result of CIRCE is a semantically indexed version of the Hachette encyclopedia. In other words, it is interested in automatically building a semantic graph out of the different articles of the encyclopedia.

Once this graph have been built, it is stored and traversed by the query-answering system each time readers of the encyclopedia ask a question. When a matching path is found, readers are returned the part of the encyclopedia which answers their question.

The next section describes the different natural language processing modules that are used by the system as well as how we plan to integrate them.

## 4.1. Making the Semantic Graph

The foreseen semantic graph is composed of the words, terms, proper nouns, dates and locations that have been detected in the encyclopedia. They are nodes in the graph which are connected to each other through specific relations that have been inferred from the sentences that are the fabric of these articles. The construction of this graph requires different tools: an entities detector (i.e. proper nouns, locations, titles, dates), a semantic analyzer to enlarge or restrict the semantic notions of each terms, a robust parser to find out which relations connect the different words, deduction semantic rules to define and name semantic relations between entities and a graph manager to store this information.

### 4.1.1. Detecting the entities

An entity can be a proper noun, a location, a title, a date or an event (Battle of Waterloo). We use *ThingFinder* together with an already defined list of entities defined by Hachette Multimedia in order to extract all the entities from the encyclopedia. These entities can be such as:

- *Proper Nouns*: Napoléon, Pasquale Paoli, Charles Le Téméraire, Charles de Gaulle, Le comte de Toulouse
- *Locations*: Les châteaux de la Loire, Paris
- *Events*: La bataille de Tolbiac, La guerre de sept ans
- *Dates*: février 1785, jusqu'en 1914, 3000 à 1500 avant J.C.
- *Measures*: 1163 m., 10 %.

We then compare these entities against sentences in the encyclopedia and mark them with special XML tags. As a result we got an entity-tagged version of the encyclopedia which looks as:

Quand <Personne> Charles le Téméraire </Personne>, maître du puissant <Loc> état bourguignon </Loc>, meurt devant <Loc> Nancy </Loc> <date>en 1477</date>, le <Loc> duché de Bourgogne</Loc>, <Loc> la Picardie </Loc> et le <Loc> Boulonnais </Loc> tombent dans l'escarcelle de <Personne>Louis XI </Personne>

*When Charles le Téméraire, master of the powerful Burgundy State, dies in front of Nancy in 1477, the domain of Burgundy, Picardie and the Boulonnais falls in the hands of Louis XI.*

Where the tag <Personne> indicates a person name., <date> a date and <Loc> a location.

The next step consists of syntactically parsing the encyclopedia in order to extract functional dependency relations. The parsing is done thanks to the Xerox *robust parser*.

### 4.1.2. Parsing the encyclopedia

The robust parser is specially designed to extract syntactic relations between words in a sentence. This parser (currently developed at XRCE) splits a text in

sentences. Each sentence is then processed and relations such as *subject*, *object*, *complement* are extracted for each word<sup>1</sup>.

The robust parser extracts the following functional dependencies from the previous sentence:

*SUBJ(tombent,Picardie)*  
*SUBJ(tombent,Boulonnais)*  
*SUBJ(meurt,Charles)*  
*SUBJ(tombent,duché)*  
*VPPMOD (tombent,dans,escarcelle)*  
*VPPMOD(meurt,devant,Nancy)*  
*VPPMOD (meurt,en,1477)*  
*NNPP(maître,du,état)*  
*NNPP(duché,de,Bourgogne)*  
*NNPP(Nancy,en,1477)*  
*NNPP(escarcelle,de,Louis)*

Where:

*SUBJ* stands for the subject relation between a verb and a noun.

*VPPMOD* is an indirect complement of a verb. This relation is extracted when a verb is followed by a PP that belongs to the sub-categorization frame of the verb.

*NNPP* is noun complement relation that links to NPs.

Functional dependency relations that involve an entity serve as input to the writing of semantic deduction rules.

### 4.1.3. Writing the semantic deduction rules

The next step consists of building semantic links between entities and to semantically type the link.

What we now want to do is to link Charles le Téméraire and Nancy via a semantic link that has the type "location-death).

In order to do this we use the robust parser output to write semantic deduction rules.

For example, if we examine the following syntactic functions:

*SUBJ(meurt,Charles)*  
*VPPMOD(meurt,devant,Nancy)*  
*VPPMOD (meurt,en,1477)*

We see, that we can connect the verb *meurt* (dies) with three different objects that are of a specific type.

We have recorded in our file that *Charles* was a person, that *Nancy* was a location and *1477* a date.

We can use this information to write the following rules:

if (*SUBJ(mourir,person)* and  
*VPPMOD(mourir,prep,location)*) then  
*LocationDeath(person,location)*.

if (*Subj(mourir,person)* and  
*VPPMOD(mourir,prep,date)*) then  
*DateDeath(person,location)*.

---

<sup>1</sup> At the moment the robust parser extracts around 30 functional dependency relations

If we apply these two deduction rules to the output of the *Robust Parser*, the result is:

*LocationDeath(Charles,Nancy)*  
*DateDeath(Charles,1447)*

The robust parser provides a formalism to write these deduction rules. First designed for syntactic purpose we can extend it and use it to write semantic rules.

In order to go one step further into semantics it is now important to semantically disambiguate words using the Xerox *semantic disambiguator* (Brun, Segond, 2000 & Dini et al. 1998).

#### 4.1.4. Disambiguating words

The functional dependencies extracted by the robust parser provide a first layer to store and manage information, but this step is usually not sufficient to answer every query about a specific subject.

Very often, the query contains words that are explicitly used in the documents. For example, the documents may describe a *king* as a *sovereign*, while the query will use the word *king* to refer to this person. We need tools to connect the words that have similar meanings.

The Xerox semantic disambiguator uses a database of semantic disambiguation rules which have been extracted from dictionary content<sup>2</sup>. Rules extraction has been performed using the dictionary as a semantically tagged corpora and functional dependency relations obtained with the robust parser. Semantic disambiguation rules are then applied to any new text via an applicator which uses both a linguistic strategy and the semantic of dictionary tags (in our case SGML) in order to prioritize rules application<sup>3</sup>.

Performing word sense disambiguation helps the system in choosing through a list of possible synonyms the most appropriate ones according to the meaning of a given word appearing in a given context. The next section describes how the system uses synonyms list in order to retain in the resulting graphs only the appropriate semantic links.

#### 4.1.5. Using synonyms

As words are polysemous it is important to choose between the different possible meanings in order to retain in the final graphs only the appropriate semantic links.

Indeed, consider the sentence :

Le Roi de France *vit le jour* en Gascogne. (*The King of France was born in Gascogne*).

In this example, the expression *voir le jour* is semantically equivalent to the word *naquit* (*naître, to be born*).

If a query contains the word *naître* and the sentence in the text uses the expression *voir le jour*, the system should be able to find a correct match between these expressions.

The final graph for the previous sentence is:

<- voir le jour ->  
France -> King <- naître -> Gascogne

As we can see the disambiguation process restricts this enrichment of the semantic graph only to the synonyms of a specific meaning of a word as revealed by the context.

Indeed in the sentence “L’armée comprenait des Basques et des Bretons” (*The army was composed of Basques and Bretons*) the meaning of *comprendre* is to *contain* not to *understand*. The graph will be enriched with words like *composer* or *contenir*, but will not contain words like *voir* or *compréhension*.

To summarize, the overall CIRCE process is as follows: the robust parser output is reassembled as a semantic graph where concepts are directly mapped over words. Next, the names of the syntactic relations are mapped over conceptual relations as *agent, patient* etc. Each sentence is analyzed in the same way and the resulting graphs are then merged to build one single graph for each article. Entities previously detected by the *ThingFinder* serve as bridges in this graph to gather scattered information among the article. For example, if *Charles le Téméraire* is mentioned several times through different sentences, the name will be used as a pivot to unify the different descriptions given by the text about this entity.

## 5. Conclusion

This paper presented the first steps of the automatic semantic indexing of an encyclopedia. We showed that natural language processing tools, such as robust parsing and semantic disambiguation, are now ready to perform such a task in a fine-grained way. Issues that we have not addressed yet concern the building of semantic resources such as general ontologies as well as the treatment of anaphora.

Since we mainly work with an encyclopedia with a focus on proper nouns, locations and dates, the ontologies we are firstly interested in concern specific classes of verbs (such as *mourir, naître, écrire, vaincre* etc.).

Anaphora is a widely spread phenomenon throughout the encyclopedia. There are divided into different sorts, some of them are common nouns utilized as such (*le président des Etats-Unis a dit...*, *Le directeur de la banque de France a affirmé* etc.), others are pronouns or relative reference expressions (*dans le même pays... trois ans avant...*). They can occur within a sentence or at the paragraph level. A typical example is as below:

En 580, Carthage défend les Phéniciens de Motyé et de Palerme contre les Grecs, dont elle défait les armées à Sélinonte, sur la côte sud-ouest de la Sicile.

Un demi-siècle plus tard, elle s’allie aux Étrusques d’Italie occidentale et expulse de Corse les Phocéens de Marseille.

<sup>2</sup> In our case we used the Oxford-Hachette French dictionary (Oxford-Hachette, 1993).

<sup>3</sup> For more details about the semantic disambiguation system see (Brun, Segond, 2000).

In 580, Carthage helps the Phenicians of Motyos and Parlerma against the Greeks. Carthage defeats the armies at Sélimente on the South-West coast of Sicily. Half a

century later, it allies with the Etruscans of West Italy and expels from Corsica the Phoceans from Marseilles.

## 6. References

- Aït-Mokhtar S., J.-P. Chanod, 1997. Subject and Object Dependency Extraction Using Finite-State Transducers. *ACL'97 Workshop on Information Extraction and the Building of Lexical Semantic Resources for NLP Applications*, Madrid July 7th-12th 1997.
- Alcouffe P., 1998. From thematic index to semantic links: querying multimedia reference CD-ROM as knowledge bases, *LREC Proceedings*, Granada, 1998.
- Brun C., F. Segond, 2000. Semantic encoding of electronic documents, *To appear in International Journal of Corpus Linguistics*, Kluwer.
- Chanod J.-P., Tapanainen, P., 1996. A Robust Finite-State Grammar for French *ESSLLI'96 Workshop on Robust Parsing, August 12-16, 1996, Prague, Czech Republic*.
- CoBrain <http://www.Cobrain.com>
- Dini L., V. Di Tomaso, F. Segond, 1998. Word Sense Disambiguation with Functional Relations, *Language Resource and Evaluation Conference*, Granada, May 98.
- Dini L., V. Di Tomaso, F. Segond, 1998. Error Driven Word Sense Disambiguation, *COLING/ACL, Montreal 98*.
- Hachette, 2000. L'encyclopédie Hachette 2000, *Hachette-Multimedia*.
- Kupiec J., 1993. MURAX : a robust linguistic Approach For Question Answering Using an On-Line Encyclopedia, *ACM-SIGIR Proceedings*, Pittsburgh, 1993.
- Oxford-Hachette, 1993, *The French Dictionnary*, Oxford University Press.
- Proux D., Y. Chenevoy. 1997. Natural Language Processing for Book Storage : Automatic Extraction of Information from Bibliographic Notices. *In Proceedings of the Natural Language Processing Pacific Rim Symposium*.
- XeLDA <http://www.xrce.xerox.com/ats/xelda>