Cognitive grammar for indexing and writing

Sahbi SIDHOM , Mohamed HASSOUN, Richard BOUCHÉ
Department S.I.I.- CERSI
ENSSIB, 17-21 Boulevard 11 November 1918
69623 Villeurbanne - Lyon, FRANCE
e-mail: {sidhom,hassoun,bouche}@enssib.fr

Resumen :

Nuestro proyecto surge de la reflexión sobre la manipulación de documentos multimedia, la representación textual de su contenido y el analisis en vistas de creación de un índice.
Nuestro estudio permetido a identificar un modelo sintáxico que surge del análisis de un conjuto de resumenes de l'INA en Francia.
La gramática de dicho modelo llamada "cognitiva" es menos compleja par los problemas de lengua y es reutilizable como ayuda para la redacción de textos resumidos.

Palabras clave:

Multimedia, INA Francia, ficha documentativa, sintagma nominal, gramática, ayuda a la redacción, tratamiento automatica de languas, analizador morfo-sintáctico.

Summary

Our research project comes from the reflection regarding the multimedia document manipulation, the textual representation of their contents and the automatic analysis in view of their indexing. Our study has allowed to identify a syntactic model stemming from the summaries corpora analysis of the I.N.A. France (Institut National de l'Audiovisuel de France). The grammar of this model called " cognitive " is a little complex to deal with language phenomena. It could be re-used in writing summary texts .

Keywords

multimedia, I.N.A. France, documentary note, French nominal phrase, grammar, writing assistance, natural language processing, morpho-syntactic parser.

Introduction

The multimedia allows an unique support to contain text, photo, audio and video documents, which until then were exploited separately.

We pass from the writing document civilization to the multimedia where each type of communication has replied to constraints and specific possibilities of its material support.

The mixture of media has been simplified , but we know badly today how to represent the non-textual document contents.

I. Representation of the content of a document

A documentary information system (library, center of documentation, databases, etc.) includes a part of human activities and a part of automatic process. Concerning textual documents case, the integration of a document of this type is done by reference to a knowledge owned by the system. This knowledge is represented thanks to different automatic tools like classifications, documentary languages or algorithms coming from artificial intelligence paradigms [4].

This integration is in general the automation of some intellectual tasks. Its result appears by a structured database. The conception of such a database relies strongly on the hypotheses made on the mechanism of information retrieval [2].

For the other type of documents (non-textual), the interpretation is described otherwise :

Text , image, audio and video interpretation systems are built on a set of symbolic signs and on particular grammars [1]. These systems inform if we can interpret them, that is to say to decode their symbols.

We are authorized to decode the text by making linguistic references, but what about the decoding of images, audio or video sequences ?

A. specific Aspects of the decoding : I.N.A. France

The analysis of multimedia documents preserved or produced by information services as stations or television channels needs the greatest care if we want to facilitate the information location and to reply to demands from extracted documents formulated by users as journalists.

Concerning audio reporting, documentary video or photography, these documents have a great potential of reuse. The analysis of these different document types is necessary to identify and to characterize their contents.

The I.N.A. has developed a grid of audiovisual document analysis then the method has been re-adapted for other types of documents.

The interest of this grid of analysis [8] is to produce a set of symbolic signs that will serve to interpret them by a list of descriptors on the content of the document, to verify them by a specialized audiovisual thesaurus, and finally to produce of a set of sentences to construct summaries (short and long) about the document.

Produced summaries are added in the bibliographical note associated to the document.
The short summary is identified in the note by the recording name " chapeau " , and for the long summary, by the recording name " résumé" .

B. Process of representation of the content

The content representation of a document shows complex problems that some scientists try to solve by applying methods, by adapting some types of models, or by using hybrid approaches coming from other research works.

Specialists of the documentation in specialized organisms have found efficient and adequate methods for this task considered as " too intellectual and tiring " .

For the case of I.N.A., in these bibliographical notes in addition to characteristic and descriptive information of the document, textual content representations are incorporated as " chapeau " for a general and brief summary, " résumé " for a more elaborated and complete summary, and " résumé producteur " for an elaborated summary by the document producer or by the production company [8].

For our project, this content representation is fundamental to unify the nature of processing of the different types of documents in view of their indexing.

To realize this task, we work on the speech of the natural language, especially the French language, and we try to construct a description of its morpho-syntactic organization.

Such a logical step consists firstly in the recognition and the categorization of words, then to the collecting of these words in syntagma (syntactic group). But, these syntagmatic constructions have a major and important part in the informative structure of the speech [7].

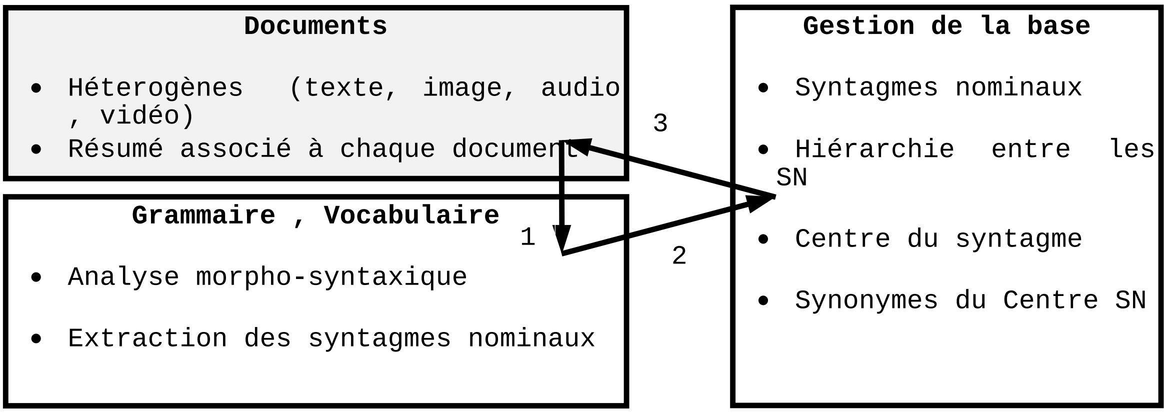
C. Varieties of documents

Multimedia data are obtained, either by direct creation on a numerical support thanks to specialized software, or by transfer from a traditional support (photographs, paper , audio tapes, video, etc.) to a numerical support. Then, they are distributed via computer systems (typically, the internet) [5].

The documentation center of I.N.A., specialized in areas of communication and audiovisual is a unique center in France due to its valuable resources (paper works, study and research reports, photographic documents, news reporting, etc.). Since the law of 1992, June 20th it has been in charge of the legal deposit for all the emissions of French radio-television.

In such an environment, technologies allowing the automatic extraction of relevant documents will be brought to play an essential role in the information society.

II. Process of knowledge extraction



Documents

- Héterogènes (texte, image, audio , vidéo)
- Résumé associé à chaque document

Grammaire , Vocabulaire

- Analyse morpho-syntaxique
- Extraction des syntagmes nominaux

Gestion de la base

- Syntagmes nominaux
- Hiérarchie entre les SN
- Centre du syntagme
- Synonymes du Centre SN

The undertaken research work focuses on the idea of a tool allowing the indexation (→[1]) of multimedia documents from a textual content description. We would like to define grammatical and lexical constraints (→[2]) of writing of this descriptive, in such a way that its analysis could be made automatically and end to an indexation.

On the organizational point of view, we do not wish to enumerate all grammatical rules of language rewriting, i.e. to make the list of all possible language constructions, but it seems more judicious that the study of the corpus leads us to this list.

III. Linguistic model *vs.* indexation

We have sought a linguistic model to make obviousness the information content of textual recording and to try and find linguistic units whose reference to the reality is stable.

The workgroup of SYDO-Lyon* scientists (Système Documentaire at Lyon) has proposed, for such linguistic units, a recognition grammar of the French concept " *syntagme nominal : SN* " that we could try and translate : *French Noun Phrase* (NP).

One of their work objectives [3],[9] was to show that is quite possible appointing a linguistic meaning to a set of determined syntactic structures.

In this case, the NP can be understood as a purely syntactic element of the speech, formally identifiable and comparable to subject and complement phrases, and as a " referent ", or " référent " in french.

A referent is an element able to appoint palpable or not palpable objects of the real world.

SYDO-Lyon : This work-group is composed by scientists of different laboratories and research centers at Lyon universities as laboratoire d'informatique documentaire Lyon I, CERSI enssib, CRLS Lyon II and ERSICO Lyon III.

A. The linguistic model

The first hypothesis of the SYDO model is that parts of the speeches constructed around the noun (or NP | SN) are those that are carry reference to objects of the speech universe and therefore those that it is necessary to identify.

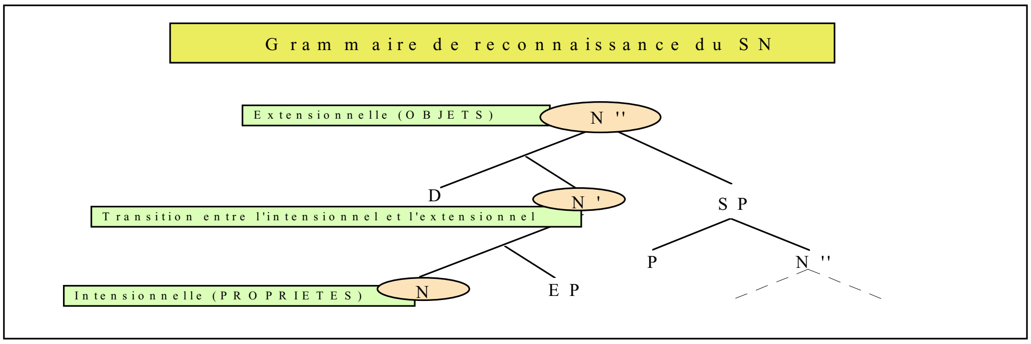
The proposed linguistic model reflects the mechanism allowing the passage of predicate words (in intensionnal logic) to the NP (the closing of a class of objects).

The whole is done by passing by transitory stages that correspond to free complex predicates then linked to the speech universe (in extensionnal logic).

The conceived model has for aim [7] :

1. to allow the identification of NP (level N"),
2. to determine the structure of these NP by highlighting the relationships between its constituents. The storage of NP representation (§ II., $_{\rightarrow}^{3}$) will facilitate the information retrieval.
3. to show the word transition mechanism (level N'). The word as predicate functioning in an intensionnal logic (level N) and its transition to the referential value unit or the NP, i.e. extensionnal logic (N").

The grammar of the NP articulates around three levels :


Grammaire de reconnaissance du SN
Extensionnelle (OBJETS)
N''
D
N'
SP
Transition entre l'intensionnel et l'extensionnel
P
N''
Intensionnelle (PROPRIETES)
N
EP

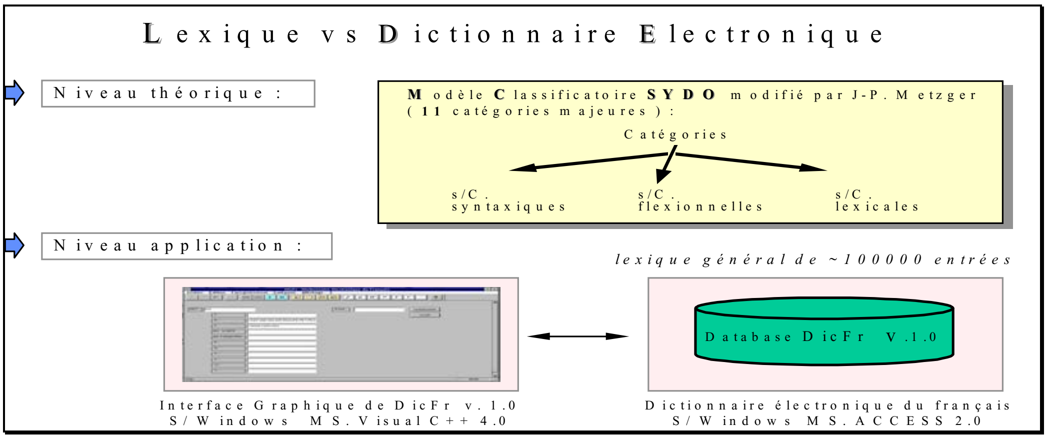
B. the model architecture : morpho-syntactic parser

The morpho-syntactic parser consists of three essential phases. The first includes morphological categories that take French Grammar into consideration. The second phase comprises the lexicon that provides necessary information for the syntactic parsing. This lexicon includes variables that have been organized in inflected, syntactic and lexical classes. The third phase consists of the syntactic parsing that is composed of re-writing syntactic rules of NP.

- **Morphological Analysis**

The morphological parser is a tool of recognition of forms in a text. It processes separately each text forms and provides all possible cutting out in a couple (basis, flexion) and interprets them.

The extracted basis of the word gives access to the lexicon (the dictionary) [6] and therefore to associated information (inflected, syntactic and lexical information).


Lexique vs Dictionnaire Electronique
Niveau théorique :
Modèle Classificatoire SYDO modifié par J-P. Metzger
(11 catégories majeures) :
Catégories
s/C. syntaxiques
s/C. flexionnelles
s/C. lexicales
Niveau application :
lexique général de ~100000 entrées
Database DicFr V.1.0
Interface Graphique de DicFr v.1.0
S/Windows MS.Visual C++ 4.0
Dictionnaire électronique du français
S/Windows MS.ACCESS 2.0

- **Syntactic analysis**

The goal of the syntactic parsing is the decomposition of the text in NPs and emphasis, inside each NP, on others NPs and words roundup.

At parser entry, the series (set) to analyze is a sequence of couples (lexical entry, category). The result expected is the NPs recognition.

IV. Statistical study of the corpus

We have tried to study the stability of textual descriptive (summaries) by a statistical analysis of its grammatical components.

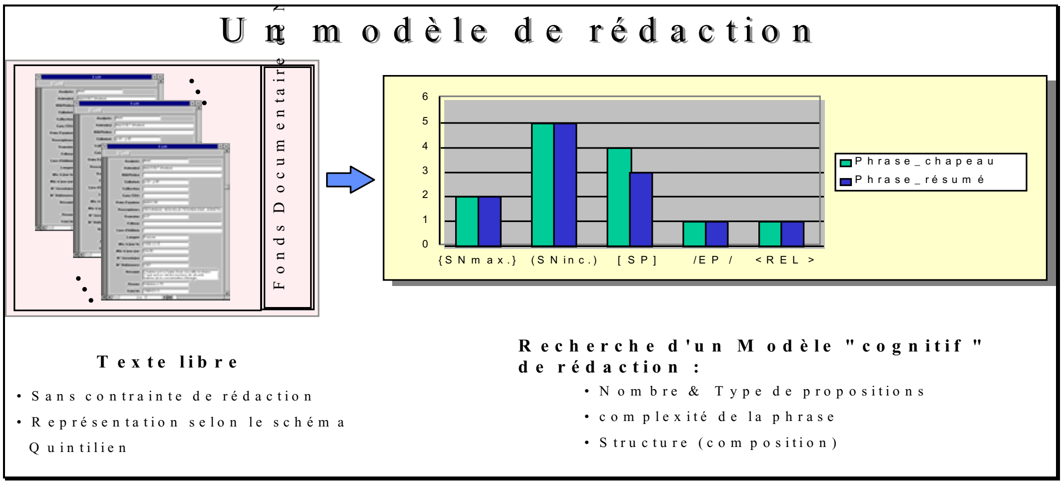
Syntactic elements that have been used for this study are maximal NPs (SNmax.), NPs included in a maximal NP (SNinc.), prepositional phrases (SP = prep. + SN), prepositional expansions (EP = prep. + noun) and relative sentences (REL = rel. + phrase).

The corpus is constituted approximately of 200 bibliographical notes of audiovisual documents (text, radio and television emissions). Each note contains at least two summary-records among “ chapeau ” , “ résumé ” and “ résumé producteur ” .

The statistic analysis on the note corpus has revealed a grammatical stability in the textual descriptive (summaries). Being the cover of this last, a stability one can notice. Then, we have observed that the INA documentalists have no editorial, structural or syntactic constraints for the writing of these summaries, they just have to apply the grid of content representation (or grid/method of audiovisual document analysis).

The reusing of this grammatical stability, meaning of text production, allows to construct a model of summarized text writing with a restrictive NP grammar and properly shaped phrases. Syntactic structures of this grammar is not as complex as we thought.

We consider that this “ cognitive ” model will serve both as an indexing tool and as a help for controlled writing.



Un modèle de rédaction

Texte libre

- Sans contrainte de rédaction
- Représentation selon le schéma Quintilien

Recherche d'un Modèle "cognitif" de rédaction :

- Nombre & Type de propositions
- complexité de la phrase
- Structure (composition)

Conclusion

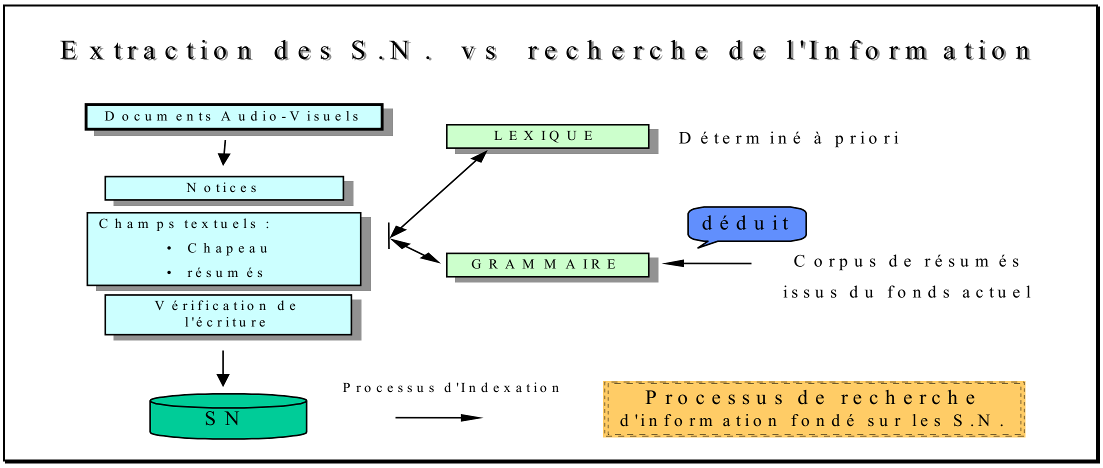
If one compares to keywords with which one usually indexes documents, on the one side the NP is a coherent descriptor, as formally defined, on the other side a real and objective descriptor, as coherent with language phenomena (speech).

The interest of this work is that, without syntactic or editorial constraints in the production of summary texts at INA, identifying a stable syntactic model and its grammar was shown a rather complex structure.

The identified model reflects an intellectual activity and a homogeneous way of writing behavior at INA.

This identified “ cognitive ” model will serve, once implemented, to index by NP extraction, to carry out information retrieval based on the NP and to give a help for controlled writing.

The information retrieval process in our study is based on the SNmax., on the SNinc., on the center of the SN. If the first attempt fails, next tries will go to its synonyms and to the center of the SN.



Thanks

We thank INA France for its agreement and cooperating at the constitution of our corpus of audiovisual documentary notes.

Bibliography

[1] Anne-Marie Guimier-Sorbets, *Des textes aux images: accès aux informations multimédia par le langage naturel*, Documentaliste-Sciences de l’information, 1993, vol.30 n°3.

[2] Esen Ozkarahan, *Mutltimedia document retrieval*, Information Processing and Management, Vol.31, N°1, pp.113-131, 1995.

[3] Geneviève Lallich-Boidin, Gérard Henneron, Rosalba Palermiti, Analyse du français : achèvement et implantation de l’analyseur morpho-syntaxique, Les cahiers du CRISS N°16, novembre 1990.

[4] Jaques Maniez, *L’évolution des langages documentaires*, Documentalistes-Sciences de l’information, 1993, Vol.30, N°4-5.

[5] Jean Guy Meunier, La lecture et l’analyse de textes assistées par ordinateur: quelques fondements théoriques, Cahiers de recherche LANCI Université du Québec Montréal, N° 95.1,1995.

[6] Max Silberztein, INTEX overview,1998,*http://www.ladl.jussieu.fr/INTEX/overview.html* (visited 02/18/98).

[7] Richard Bouché, *Le syntagme nominal : une nouvelle approche des bases de données textuelles*, Meta XXXIV.3.1989.

[8] Sahbi SIDHOM, *Automatic indexing of multimedia documents based on the extraction of nominal phrases*, Proceedings of 5th ISKO Conference, 25-29 august 1998 Lille, France, Ed. Ergon Verlag , p.424 (Poster).

[9] Widad Mustapha-Elhadi, *La contribution de la terminologie à la conception théorique des langages documentaires et à l’indexation automatique*, Meta XXXVII.3.1992.