Document Structure Identification as a means for Relevant Indexation

Emmanuel Giguet,
Nadine Lucas,
Grégoire Cousin

In information retrieval systems, the indexation task is usually conducted irrespective of the document structure. We introduce here a module which allows preprocessing of documents prior to indexation techniques. It detects the physical lay-out of text and labels text zones, such as title and text-body. The method relies on pattern recognition and applies to a wide range of documents. This ensures the correct routing of selected parts of a document towards adequate tools for analysis. Examples of text identification and segmentation at rough and minute grain are presented. Document structure identification offers an opportunity to balance the calculation of the inputs of an index according to the nature of the zones in which the terms appear (title or text-body for instance).

1 INTRODUCTION

Our laboratory has been involved in a project of technology transfer with DATOPS company for eight months. The aim of this project, called LINGUIX, was to improve the capacities of statistical analysis of textual flows of DATOPS by embedding linguistic components. DATOPS' technology allows the detection and follow-up of key-words in strategic information environments. It deals with heavily evoked topics, i.e. the trends of opinion, and also the detection of weak tendencies such as the emergence of opinions, or technological prospects.

We developed linguistic analysis tools to supplement the range of our partners' statistical indexers, namely a language identification tool, a syntactic parser for French and English (Giguet & Vergne, 1997; Giguet, 1998; Vergne & Giguet, 1998; Vergne, 1999), a rhetorical parser for English (Lucas, 1993; Lucas *et al.*, 1993), and a parser for metaphors (Ferrari, 1996; Ferrari, 1997). In addition to these tools, we produced a routing module which takes into account the structure of documents. This module makes it possible to balance the calculation of the index entries according to the nature of the zones or areas in which the terms appear. It is a generic module, since it can be customized for any application, and according to language and genre. It also enables the selection of the measure of granularity of entering and outgoing textual flows.

The structure of this paper is as follows. Firstly, we will briefly introduce the drawbacks of full-text indexing method and its expected advantages if we take into account the document structure. Secondly, we will describe the problematic of the identification of the document structure within raw texts and marked-up texts. Thirdly, we explain the method we developed in order to compute the structure of such documents ; we give implementation details of the module. Fourthly, we will illustrate step by step the behavior of the tools with the help of two examples. Finally, we will provide some conclusive remarks.

2 LIMITATIONS OF FULL-TEXT INDEXING

Having looked at linguistic components embedded in various indexing systems, we extended our work not only to the reproduction of well-known routines such as noun phrase, currency or named entity detection. We tried to open up some new prospects by introducing more sophisticated tools. Since the very beginning of these researches, we believe that we could not reach this goal without knowledge on the structure of the documents. Indeed, in every linguistic components we designed, there is good reason to give up blind full-text indexing. We will below illustrate it using our linguistic parsers :

For trivial uses such as noun phrase detection, our syntactic parser is robust enough in order to process without knowledge about text structure. However, we remark that certain parts of documents can be of no use for indexation purposes. For instance, in news dispatches, the signature includes the source of the document and copyright information that we might want to ignore while indexing. Furthermore, other uses of the parser gives better results if the process only considers textual areas. For instance, indexing “subjects” can be interesting but it needs the identification of textual areas (e.g., we have to distinguish paragraphs from tables or titles).

The rhetorical parser works closely with the document structure. Indeed, its aim is to recursively decompose a speech into labeled parts (namely introducer, pivot and adjunct) and to qualify them according to their structure. In the indexation system, such a process allows the weighting of indexed terms depending on the type of the part in which it occurs. Moreover, when higher speed analysis is required, the indexation task can also skip or delay the analysis of adjuncts.

The document structure can also be useful for the parser of metaphors. The underlying idea of the use of such a parser in an indexation system of textual flows is to be able to detect phenomena such as tendencies (e.g., on a financial market) or excitation (e.g., of some people about a particular event), then, the weight of a target term can be balanced depending on both the strength of the metaphor and the nature of the text area in which it occurs (e.g., main title, paragraph).

To summarize, knowledge on document structure enables our system to forward the different parts of a document to relevant parsers and to provide balanced values for the content of different types of text areas in the indexation task. Moreover, we saw that the finer the linguistic components, the more we gain with knowledge about document structure.

3 FROM PHYSICAL LAYOUT TO DOCUMENT STRUCTURE

In previous section, we tried to show that different reasons could be pointed out in order to account for the structure of the documents whilst indexing. Finding good reasons in itself is not that difficult: the real challenge is to retrieve the document structure from the physical layout of the original text. By retrieval of the structure, we mean specifically the identification of the absolute and relative locations of all the text areas of the document, as well as the identification of their nature (e.g., sections, sub-sections, titles, paragraphs, lists, tables). This problem lies in the *pattern recognition* research field (Chali *et al*, 1996 ; Pascual, 1991 ; Pascual & Virbel, 1996 ; Barnard & Logan, 1996).

When an application is meant to process a single type of text, taking into account its physical structure may be trivial : it is only a matter of exhaustive description, depending on the richness of physical rendering (e.g., italic, capitals) and of its exploitation to emphasize the logical frame. For instance, in raw texts, the absence of formatting instructions makes it difficult to locate the titles or the tables. In an SGML or XML text, if the elements (i.e., tags) were correctly defined and used when designing the document (Barnard *et al*, 1995 ; Ide & Véronis 1995), the identification of the type of the various text areas can be much easier.

When the documents come from various sources as it is the case for the documents which are retrieved from the Internet, it is necessary to find a customizable identification and extraction system. Indeed, each source of documents, each publishing software, each user has its own formatting method. On the Internet, the most common technique used is tagging, i.e., the insertion of HTML elements in raw texts. We can point out that in practice, the tagging recommendations are nearly never followed. For instance, physical tags are used in order to format titles while logical tags, designed for this purpose, exist (see table 1). What matters most to the author is the visual rendering.

Within the framework of the LINGUIX project, we were mainly faced with the analysis of news dispatches from worldwide wires such as Reuters News Service, the Agence France Presse, Washington Post... In these wires documents, part of the formatting is made automatically, which ensures the presence and the regularity of certain tags, but not always their relevance. In addition, the system had to be robust enough in order to deal with documents coming from any Web

site. The common point of this variety of documents is the fact that they include an HTML-like tagging, more or less specifically enriched, with a major importance attached to aesthetics rather than to logical considerations.

News wire	Beginning of dispatch title	End of dispatch title
CNN Interactive	<FONT SIZE=6>	</FONT>
RBB	<H5>	</H5>
PR Newswire	<FONT SIZE=+1><STRONG>	</STRONG></FONT>

Table 1 : Variability of title marks according to the news wire

4 METHOD

The document processing protocol is as follows: each file contains a single document, which can be a news dispatch or a web page, with a meta-information indicating the type of the document. This meta-information was set downstream by a file retriever and depends on the nature, and on the source of the document.

This meta-information is the basis we rely on in order to extract the text areas. For that purpose, we use a matching list between a type of document and an identification and extraction procedure of text areas. Each procedure makes use of both formal and positional criteria to mark up the documents.

The formal criteria are, for instance, the use of capitals for the title in Reuters news dispatches and the use of the STRONG tag (e.g., bold rendering) for the title in CNN news dispatches. The use of positional criteria makes it possible to select locations where it is relevant to look for a specific formal criterion.

Finally, we assign a value to each text area which is detected in the document. For instance, we assign the value “title”, “body” or “signature” to subdivide a text ; then, we extract the areas of interest, such as the “body” which is then routed to a sentence parser.

A format parser was implemented to carry out this task. It associates a logical frame with a type of document. It also anticipates the addition of new features when a new type of document pops up and does not match any the existing formats, and it allows an easy update of the detection rules of the text areas when a format changes. The output of this routing module, which also acts as an extractor, are given in an independent language, which can be interpreted by the other parsers.

4.1 General idea

As it has been stated earlier, the technique used to retrieve the textual zones relies on both formal and positional criteria. In other words, the technique generates a tree structure of the document by superimposing a XML-like tagging on it. With the XML notation, we bound text areas using opening and closing tags, and we label these areas thanks to the name of the element.

The mark up of opening and closing pair of tags is very flexible. It can be carried out simultaneously in a single action. It can also be carried out at different times, in different actions. Hence, it is possible to mark up simultaneously a pair of tags only if we can exhaustively characterize the form of a zone. If this form is too variable, it is sometimes easier to bound it by characterizing each of its borders in turn.

4.2 Implementation

The recognition of the logical structure of a given document or of one of its constituents (text area) is obtained through the following steps :

- the physical features of the items belonging to the immediately lower level in the logical structure are assessed and matched to a more abstract level;
- the item belonging to the lower level is then in turn examined, thus proceeding down the ladder of the logical structure;
- items (text areas) relevant for a given processing are extracted and routed towards a specific parser, e.g. sections or paragraphs can be directed to the rhetorical parser and sentences to the sentence parser.

Let us consider how we deal with the recognition of the constituents. It is based on pattern recognition. The parser does not solely rely on physical features, such as capital letters to detect a title. It also makes use of the position in order to characterize a specific value. For instance, a sequence of capital letters at the beginning of a text zone will be interpreted differently from a sequence of capital letters followed by a line break at the beginning of a text : the former will be analyzed as a "section title", while the latter will be analyzed as a "general title".

Patterns are written under the form of regular expressions and recognition of a text area triggers insertion or deletion of opening and closing XML tags. Many actions of this type on the document may be necessary before complete identification is achieved.

The use of regular expressions makes it easier to deal with various physical features and allows the setting of the formal criteria. The position criterion is best handled through the insertion of tags in the text.

We described the recognition of next lower constituents, we will now proceed to the recognition of the tree-like organization of these constituents.

This recognition is made through successive parses. At the beginning of the process, the higher constituent is the text in itself. Then the next lower constituents are assessed, as stated before. Each parse goes down a step in the ladder, and detects in turn the immediately lower constituents until there are none, thus allowing the description of the logical tree-like structure.

5 EXAMPLES

In order to show how the parser proceeds through the above cited steps, to retrieve the logical structure of documents, we will describe step by step two examples. The first explains how the logical structure is retrieved from a news dispatch from RBB written in Roman alphabet and in Spanish. Note that at this stage, the choice of language is open, we currently work on English and French but the module handles Spanish as illustrated below. The segmentation provides a broad division in title and body and a division of the body in paragraphs.

The second explains recognition of smaller units: paragraphs and words from a document written in Russian (using the Cyrillic alphabet). This illustrates the ability to deal with different alphabets (through Unicode), as well as the ability to deal with different types of text.

It illustrates the ability of the format parser to deal with various languages and different measures in the granularity of typographic items.

5.1 Structure identification of a news dispatch (in Spanish)

5.1.1 Step 1 : identification of the document source

The original document is identified as an RBB dispatch – source underlined below. The structuring rules for this source are therefore selected. These rules are listed in the box below.

```
TEXTUAL_AREA          (WHOLE_BODY)
SEARCH_PATTERN        ([0-9][0-9][0-9][0-9]-[0-9][0-9]-[0-9][0-9]:
<BR>\r?\n[ ]*)
REPLACE_BY            $1<DispatchTitle>
SEARCH_PATTERN        (</H5>)
REPLACE_BY            </DispatchTitle>$1<DispatchBody>
SEARCH_PATTERN        (<P><!- ..CO: ->)
REPLACE_BY            </DispatchBody>$1

TEXTUAL_AREA          (DispatchBody)
SEARCH_PATTERN        (<P>)
REPLACE_BY            </Parag>$1<Parag>
SEARCH_PATTERN        (<Parag>)(\r?\n\r?\n?)$
```

```
REPLACE_BY            $2
SEARCH_PATTERN        ^([\r\n ]*)</Parag>
REPLACE_BY            $1
```

The default workbench WHOLE_BODY is automatically defined on the document - in bold in the dispatch below.

```
<WHOLE_BODY><HTML><BODY>
<H5> 1999-04-14: <BR>
ESPAÑA: CASA CONFÍA EN EL EFECTO AIRBUS PARA SEGUIR CRECIENDO.
<I>(EXPNSI)</I>
</H5>
<P>
P.VARELA.Madrid<P>
Mientras la Sociedad Estatal de Participaciones Industriales(Sepi)
negocia la entrada de un socio extranjero en el capital de Cons-
trucción Aeronáuticas (Casa), la compañía ha presentado los mejo-
res resultados en sus 75 años de historia. <P>
La participación del 4,2% que la compañía española tiene en el
consorcio europeo Airbus generó el 40% del total de las ventas de
Casa durante 1998. La compañía que preside Alberto Fernández
quiere aumentar su cuota de participación en los futuros proyectos
de Airbus, que tienen un reflejo directo en la facturación. <P>
De momento, ha conseguido la asignación de una cuota del 10% en el
nuevo A-340-500/600 y negocia una participación similar en el fu-
turo A3XX. Casa tuvo un volumen de negocio de 1.008 millones de
euros (167.747 millones de pesetas) durante 1998, un 9,3% más que
el registrado el año anterior. Nuestra intención es hacer crecer
nuestras ventas en el mercado militar hasta que representen un 40%
del total de la facturación de la compañía", aseguró ayer Alberto
Fernández. El máximo responsable de Casa espera que los ingresos
que produzca el Eurofighter, el avión de combate europeo, supongan
el 15% de la facturación en cinco años. <P>
Durante 1998, el beneficio neto alcanzó 47 millones de euros
(7.798 millones de pesetas), un 20 por ciento más que en 1997.<P>
(c) Recoletos Cia. Editorial, S.A., 1999. <P>

EXPANSION (SPANISH LANGUAGE) 14/04/1999 P4 <P>
<P><!- ..CO: ->
<!- AIRBSI AIRBUS INDUSTRIE (FRA)< ->

</BODY></HTML>
<!--SIEVE_TYPE="Datops, RBB"-->
<!--SIEVE_DATE="1999.4.14"--></WHOLE_BODY>
```

5.1.2 Step 2 : Dissociation of title and dispatch-body

Workbench : WHOLE_BODY – bold, double-underline.

The first rule uses positional information: the fact that the dispatch title follows the date, itself followed by a line break. We search the occurrences of this pattern - underlined below -, to insert the opening tag DispatchTitle – in bold.

The second rule uses disruption in style between the dispatch title and the rest of the document – expressed by the tag </H5> underlined below –, in order to insert the closing tag DispatchTitle and the opening tag DispatchBody – in bold.

The third rule defines the end of the dispatch body by searching the last paragraph break <P> – underlined below –, and inserts the closing tag DispatchBody – in bold.

```
TEXTUAL_AREA          (WHOLE_BODY)
SEARCH_PATTERN        ([0-9][0-9][0-9][0-9]-[0-9][0-9]-[0-9][0-9]:
<BR>\r?\n[ ]*)
REPLACE_BY            $1<DispatchTitle>
SEARCH_PATTERN        (</H5>)
REPLACE_BY            </DispatchTitle>$1<DispatchBody>
SEARCH_PATTERN        (<P><!- ..CO: ->)
REPLACE_BY            </DispatchBody>$1
```

<WHOLE BODY><HTML><BODY>
<H5> 1999-04-14:
<DispatchTitle>ESPAÑA: CASA CONFÍA EN EL EFECTO AIRBUS PARA SEGUIR CRECIENDO.
<I>(EXPNSI)</I>
</DispatchTitle></H5>**<DispatchBody>**
<P>
P.VARELA.Madrid<P>
Mientras la Sociedad Estatal de Participaciones Industriales (Sepi) negocia la entrada de un socio extranjero en el capital de Construcción Aeronáuticas (Casa), la compañía ha presentado los mejores resultados en sus 75 años de historia. <P>
La participación del 4,2% que la compañía española tiene en el consorcio europeo Airbus generó el 40% del total de las ventas de Casa durante 1998. La compañía que preside Alberto Fernández quiere aumentar su cuota de participación en los futuros proyectos de Airbus, que tienen un reflejo directo en la facturación. <P>
De momento, ha conseguido la asignación de una cuota del 10% en el nuevo A-340-500/600 y negocia una participación similar en el futuro A3XX. Casa tuvo un volumen de negocio de 1.008 millones de euros (167.747 millones de pesetas) durante 1998, un 9,3% más que el registrado el año anterior. Nuestra intención es hacer crecer nuestras ventas en el mercado militar hasta que representen un 40% del total de la facturación de la compañía", aseguró ayer Alberto Fernández. El máximo responsable de Casa espera que los ingresos queproduzca el Eurofighter, el avión de combate europeo, supongan el 15% de la facturación en cinco años. <P>
Durante 1998, el beneficio neto alcanzó 47 millones de euros (7.798 millones de pesetas), un 20 por ciento más que en 1997. <P>

EXPANSION (SPANISH LANGUAGE) 14/04/1999 P4 <P>
</DispatchBody><P><!- ..CO: ->
<!- AIRBSI AIRBUS INDUSTRIE (FRA)< ->

</BODY></HTML>
<!--SIEVE_TYPE="Datops, RBB"-->

<!--SIEVE_DATE="1999.4.14"--></WHOLE_BODY>

5.1.3 Step 3 : Subdivision of dispatch-body in paragraph

Workbench : DispatchBody – bold, double-underline.

The first rule uses the fact that the rendering of the pattern <P> – underlined – is a paragraph break. It inserts a closing tag Parag followed by an opening tag Parag – in bold.

This rule produces unwanted tags at the beginning and end of the workbench, the two following rules correct this side effect and delete the unwanted tags – striked out below.

```
TEXTUAL_AREA          (DispatchBody)
SEARCH_PATTERN        (<P>)
REPLACE_BY            </Parag>$1<Parag>
SEARCH_PATTERN        (<Parag>)(\r?\n\r?\n?)$
REPLACE_BY            $2
SEARCH_PATTERN        ^([\r\n ]*)</Parag>
REPLACE_BY            $1
```

<WHOLE_BODY><HTML><BODY>
<H5> 1999-04-14:
<DispatchTitle>ESPAÑA: CASA CONFÍA EN EL EFECTO AIRBUS PARA SEGUIR CRECIENDO.
<I>(EXPNSI)</I>
</DispatchTitle></H5><DispatchBody>
~~</Parag>~~<P><Parag>
P.VARELA.Madrid</Parag><P><Parag>
Mientras la Sociedad Estatal de Participaciones Industriales (Sepi) negocia la entrada de un socio extranjero en el capital de Construcción Aeronáuticas (Casa), la compañía ha presentado los mejores resultados en sus 75 años de historia. </Parag><P><Parag>
La participación del 4,2% que la compañía española tiene en el consorcio europeo Airbus generó el 40% del total de las ventas de Casa durante 1998. La compañía que preside Alberto Fernández quiere aumentar su cuota de participación en los futuros proyectos de Airbus, que tienen un reflejo directo en la facturación.
</Parag><P><Parag>
De momento, ha conseguido la asignación de una cuota del 10% en el nuevo A-340-500/600 y negocia una participación similar en el futuro A3XX. Casa tuvo un volumen de negocio de 1.008 millones de euros (167.747 millones de pesetas) durante 1998, un 9,3% más que el registrado el año anterior. Nuestra intención es hacer crecer nuestras ventas en el mercado militar hasta que representen un 40% del total de la facturación de la compañía", aseguró ayer Alberto Fernández. El máximo responsable de Casa espera que los ingresos que produzca el Eurofighter, el avión de combate europeo, supongan el 15% de la facturación en cinco años. </Parag><P><Parag>

```
Durante 1998, el beneficio neto alcanzó 47 millones de euros
(7.798 millones de pesetas), un 20 por ciento más que en 1997.
</Parag><P><Parag>
(c) Recoletos Cia. Editorial, S.A., 1999. </Parag><P><Parag>

EXPANSION (SPANISH LANGUAGE) 14/04/1999 P4 </Parag><P><Parag>
</DispatchBody><P><!- ..CO: ->
<!- AIRBSI AIRBUS INDUSTRIE (FRA)< ->

</BODY></HTML>
<!--SIEVE_TYPE="Datops, RBB"-->
<!--SIEVE_DATE="1999.4.14"--></WHOLE_BODY>
```

5.1.4 Step 4 : identification of the dispatch structure

All the rules have been triggered. The text zones DispatchTitle, DispatchBody and Parag – in bold – are extracted, they become leaves or nodes of the logical tree and the root is WHOLE_BODY.

```
<WHOLE_BODY><HTML><BODY>
<H5> 1999-04-14: <BR>
<DispatchTitle>ESPAÑA: CASA CONFÍA EN EL EFECTO AIRBUS PARA SEGUIR
CRECIENDO.
<I>(EXPNSI)</I>
</DispatchTitle></H5><DispatchBody><P><Parag>
P.VARELA.Madrid</Parag><P><Parag>
Mientras la Sociedad Estatal de Participaciones Industriales
(Sepi) negocia la entrada de un socio extranjero en el capital de
Construcción Aeronáuticas (Casa), la compañía ha presentado los
mejores resultados en sus 75 años de historia. </Parag><P><Parag>
La participación del 4,2% que la compañía española tiene en el
consorcio europeo Airbus generó el 40% del total de las ventas de
Casa durante 1998. La compañía que preside Alberto Fernández
quiere aumentar su cuota de participación en los futuros proyectos
de Airbus, que tienen un reflejo directo en la facturación.
</Parag><P><Parag>
De momento, ha conseguido la asignación de una cuota del 10% en el
nuevo A-340-500/600 y negocia una participación similar en el fu-
turo A3XX. Casa tuvo un volumen de negocio de 1.008 millones de
euros (167.747 millones de pesetas) durante 1998, un 9,3% más que
el registrado el año anterior. Nuestra intención es hacer crecer
nuestras ventas en el mercado militar hasta que representen un 40%
del total de la facturación de la compañía", aseguró ayer Alberto
Fernández. El máximo responsable de Casa espera que los ingresos
que produzca el Eurofighter, el avión de combate europeo, supongan
el 15% de la facturación en cinco años. </Parag><P><Parag>
Durante 1998, el beneficio neto alcanzó 47 millones de euros
(7.798 millones de pesetas), un 20 por ciento más que en
1997.</Parag><P><Parag>
(c) Recoletos Cia. Editorial, S.A., 1999. </Parag><P><Parag>

EXPANSION (SPANISH LANGUAGE) 14/04/1999 P4 </Parag><P>
```

```
</DispatchBody><P><!- ..CO: ->
<!- AIRBSI AIRBUS INDUSTRIE (FRA)< ->

</BODY></HTML>
<!--SIEVE_TYPE="Datops, RBB"-->
<!--SIEVE_DATE="1999.4.14"--></WHOLE_BODY>
```

5.2 Structure identification of a Russian text extract

In order to illustrate a different feature of this tool, here is an example of rough structure identification of a text extract. There is a difference in script (Cyrillic). There is also a difference here in the segmentation measure which is paragraphs and words (minute grain). Below, in the box are the rules applied on the text extract.

```
TEXT_AREA               (WHOLE_BODY)
SEARCH_PATTERN          ^
REPLACE_BY              <Parag>
SEARCH_PATTERN          $
REPLACE_BY              </Parag>
SEARCH_PATTERN          ([\r\n]*)
REPLACE_BY              </Parag>$1<Parag>

TEXT_AREA               (Parag)
SEARCH_PATTERN          ^
REPLACE_BY              <W>
SEARCH_PATTERN          $
REPLACE_BY              </W>
SEARCH_PATTERN          ([ ]*)
REPLACE_BY              </W>$1<W>
```

<WHOLE_BODY>Огромные изменения произошли в государственном устройстве страны. Всего три года назад Россия впервые в своей многовековой истории обрела Конституцию, принятую открытым и свободным волеизъявлением народа. Она стала одним из важнейших факторов политической стабильности, реализовав принцип сильной президентской власти, соответствующий и потребностям общества, и условиям нынешней переходной ситуации. Конституция ограничила возможность власти подмять общество и стала реальным инструментом, с помощью которого граждане отстаивают свои права. Прежде чем менять Конституцию, нужно научиться ее уважать, научиться жить по ней, используя весь ее потенциал.

В прошлом году впервые в истории России на демократических выборах переизбран действующий глава государства. Выборы законодательной и президентской власти прошли в срок и в соответствии с избирательным законодательством. Важным событием в жизни страны стали выборы органов власти субъектов Российской Федерации и органов местного самоуправления. Еще раз подтверждено: нам удалось привести в действие демократический механизм смены и преемственности власти.

Необходимо объективно проанализировать опыт всех прошедших выборов и внести изменения в избирательное законодательство. Главная задача - усилить гарантии избирательных прав граждан. Полагаю, что более важную роль здесь должна играть система избирательных комиссий.**</WHOLE_BODY>**

The selected rules bound paragraphs and "words" in turn. Here is the result once all the rules have been triggered. Some extra rules would be needed to properly tokenize the punctuation marks.

```
<WHOLE_BODY><Parag><W>Огромные</W> <W>изменения</W>
<W>произошли</W> <W>в</W> <W>государственном</W> <W>устройстве</W>
<W>страны.</W> <W>Всего</W> <W>три</W> <W>года</W> <W>назад</W>
<W>Россия</W> <W>впервые</W> <W>в</W> <W>своей</W>
<W>многовековой</W> <W>истории</W> <W>обрела</W>
<W>Конституцию,</W> <W>принятую</W> <W>открытым</W> <W>и</W>
<W>свободным</W> <W>волеизъявлением</W> <W>народа.</W> <W>Она</W>
<W>стала</W> <W>одним</W> <W>из</W> <W>важнейших</W>
<W>факторов</W> <W>политической</W> <W>стабильности,</W>
<W>реализовав</W> <W>принцип</W> <W>сильной</W>
<W>президентской</W> <W>власти,</W> <W>соответствующий</W>
<W>и</W> <W>потребностям</W> <W>общества,</W> <W>и</W>
<W>условиям</W> <W>нынешней</W> <W>переходной</W> <W>ситуации.</W>
<W>Конституция</W> <W>ограничила</W> <W>возможность</W>
<W>власти</W> <W>подмять</W> <W>общество</W> <W>и</W> <W>стала</W>
<W>реальным</W> <W>инструментом,</W> <W>с</W> <W>помощью</W>
<W>которого</W> <W>граждане</W> <W>отстаивают</W> <W>свои</W>
<W>права.</W> <W>Прежде</W> <W>чем</W> <W>менять</W>
<W>Конституцию,</W> <W>нужно</W> <W>научиться</W> <W>ее</W>
<W>уважать,</W> <W>научиться</W> <W>жить</W> <W>по</W> <W>ней,</W>
<W>используя</W> <W>весь</W> <W>ее</W> <W>потенциал.</W></Parag>
<Parag><W>В</W> <W>прошлом</W> <W>году</W> <W>впервые</W> <W>в</W>
<W>истории</W> <W>России</W> <W>на</W> <W>демократических</W>
<W>выборах</W> <W>переизбран</W> <W>действующий</W> <W>глава</W>
<W>государства.</W> <W>Выборы</W> <W>законодательной</W> <W>и</W>
<W>президентской</W> <W>власти</W> <W>прошли</W> <W>в</W>
<W>срок</W> <W>и</W> <W>в</W> <W>соответствии</W> <W>с</W>
<W>избирательным</W> <W>законодательством.</W> <W>Важным</W>
<W>событием</W> <W>в</W> <W>жизни</W> <W>страны</W> <W>стали</W>
<W>выборы</W> <W>органов</W> <W>власти</W> <W>субъектов</W>
<W>Российской</W> <W>Федерации</W> <W>и</W> <W>органов</W>
<W>местного</W> <W>самоуправления.</W> <W>Еще</W> <W>раз</W>
<W>подтверждено:</W> <W>нам</W> <W>удалось</W> <W>привести</W>
<W>в</W> <W>действие</W> <W>демократический</W> <W>механизм</W>
<W>смены</W> <W>и</W> <W>преемственности</W>
<W>власти.</W></Parag>
<Parag><W>Необходимо</W> <W>объективно</W> <W>проанализировать</W>
<W>опыт</W> <W>всех</W> <W>прошедших</W> <W>выборов</W> <W>и</W>
<W>внести</W> <W>изменения</W> <W>в</W> <W>избирательное</W>
<W>законодательство.</W> <W>Главная</W> <W>задача</W> <W>-</W>
<W>усилить</W> <W>гарантии</W> <W>избирательных</W> <W>прав</W>
<W>граждан.</W> <W>Полагаю,</W> <W>что</W> <W>более</W>
<W>важную</W> <W>роль</W> <W>здесь</W> <W>должна</W> <W>играть</W>
<W>система</W> <W>избирательных</W>
<W>комиссий.</W></Parag></WHOLE_BODY>
```

6 CONCLUSIONS

The automatic indexing of documents is a new avenue in our research team. Having studied the linguistic components, in a way that is common in the various indexing systems, we deliberately chose not to limit ourselves to the reproduction of well-known routines. On the contrary, we tried to introduce more sophisticated tools opening some new prospects. The module of document structure identification we just introduced is a good example of this modest effort. It allows the correct routing of selected parts of a document towards adequate tools for analysis and offers an opportunity to balance the calculation of the inputs of an index according to the nature of the zones in which the terms appear (title or text-body for instance). In that respect, it is not a tool for linguistic analysis which would stand as a competitor of the traditional statistical indexers, but it is a tool for preprocessing which allows a more selective calculation of the index entries.

This tool is robust and effective enough to be used today in a chain of indexing of documents retrieved from Internet. It is mainly used for the structuring of news dispatches coming from multiple and multilingual sources, as well as the structuring of more traditional documents.

REFERENCES

BARNARD, David T. ; BURNARD, Lou ; GASPART, Jean-Pierre ; PRICE, Lynne ; SPERBERG-MCQUEEN, C.M. (1995) : Hierarchical Encoding of Text: Technical Problems and SGML Solutions; Computers and the Humanities 29,3.

BARNARD, David T. ; LOGAN, George M. (1996) : Complementary Approaches to Representing Differences Between Structured Documents. Proceedings of ALLC-ACH '96, University of Bergen, Norway, June.

CHALI, Yllias; PASCUAL, Elsa ; VIRBEL, Jacques (1996) : Text structure Modeling and Language Comprehension processes. Proceedings of ALLC-ACH '96, University of Bergen, Norway, June.

FERRARI, Stéphane (1996) : Using textual clues to improve metaphor processing. Proceedings of the 34th Annual Meeting of ACL, Student Session. University of California, Santa Cruz, U.S.A., June.

FERRARI, Stéphane (1997) : Méthode et outils informatiques pour le traitement des métaphores dans les documents écrits. Thèse de doctorat. Université de Paris-Sud, décembre.

GIGUET, Emmanuel ; VERGNE, Jacques (1997) : From Part-of-Speech Tagging to Memory-based Deep Syntactic Analysis. Proceedings of the Inter-

national Workshop on Parsing Technologies (IWPT'97), MIT, Boston, Massachussets, USA, September.

GIGUET, Emmanuel (1998) : Méthode pour l'analyse automatique de structures formelles sur documents multilingues. Thèse de Doctorat. Université de Caen, France, décembre.

IDE, Nancy ; VÉRONIS, Jean (1995) : The text encoding initiative: Background and context. Dordrecht, Kluwer Academic Publishers.

LUCAS, Nadine (1993) : Syntaxe du paragraphe dans les articles scientifiques en japonais et en français. Dans Parcours linguistiques de discours spécialisés, éd. Moirand *et al*., Berne…Paris, Peter Lang

LUCAS, Nadine *et al.* (1993) : Discourse analysis of scientific textbooks in Japanese : a tool for producing automatic summaries. Department of Computer Science Tokyo Institute of Technology, 92TR-004.

PASCUAL, Elsa (1991) : Représentation de l'architecture textuelle et génération de texte. Thèse de Doctorat, Université Paul Sabatier, Toulouse.

PASCUAL, Elsa ; VIRBEL, Jacques (1996) : Semantic and Layout Properties of Text Punctuation. Proceedings of the ACL workshop on Punctuation in Computational Linguistics, Santa Cruz, June.

VERGNE, Jacques ; GIGUET Emmanuel (1998) : Regards Théoriques sur le "Tagging". In proceedings of the fifth annual conference Le Traitement Automatique des Langues Naturelles (TALN 1998), Paris, France, juin.

VERGNE, Jacques (1999) : Étude et modélisation de la syntaxe des langues à l'aide de l'ordinateur, Analyse syntaxique automatique non combinatoire. Habilitation à diriger des recherches, Université de Caen.

Emmanuel Giguet is a researcher in computational linguistics at the GREYC CNRS UPRESA 6072, Caen University, 14032 Caen Cedex, France. He can be reached at Emmanuel.Giguet@info.unicaen.fr, see http://www.info.unicaen.fr/~giguet.

Nadine Lucas is a researcher in linguistics at the GREYC CNRS UPRESA 6072, Caen University, 14032 Caen Cedex, France. She can be reached at Nadine.Lucas@info.uncaen.fr, see http://www.info.unicaen.fr/~nadine.

Grégoire Cousin is a doctoral student at the GREYC CNRS UPRESA 6072, Caen University, 14032 Caen Cedex, France.

The LINGUIX project was funded by the French Ministry for Higher Education and Research (ref. 98K6411).