What is this Text about?

Nicolas Hernandez, Brigitte Grau

What Is This Text About?

Combining topic and meta- descriptors for text structure presentation

Nicolas Hernandez
LIR Group
LIMSI–CNRS laboratory
91403 Orsay, France
hernandez@limsi.fr

Brigitte Grau
LIR Group
LIMSI–CNRS laboratory
91403 Orsay, France
grau@limsi.fr

ABSTRACT

Most work in text retrieval aims at presenting the information held by several texts in order to give entry clues towards these texts and to allow a navigation between them. Besides, a lesser interest is dedicated to the definition of principles for accessing content of single documents. As most information retrieval systems return documents from an initial request made of words, a usual solution consists of presenting document titles and highlighting words of the request inside a passage or in the whole document. Such a presentation does not allow a rapid reading and systems cannot satisfy themselves with it. Our studies lead us to provide indicative and informative view of texts as in summarization systems. We offer the user different levels of abstraction of a text: the first is a global overview, where global topics are indicated and positioned in the text. The second level of abstraction goes deeper in the topic description by adding local topics and information about the argumentative role of the segments. In this paper, we will detail the extraction of thematic descriptors and meta-descriptors that relies on recurrence -respectively in a text or in the corpus- and how their characterization provides the segment structuring.

Categories and Subject Descriptors

H.3.1 [**Information storage and retrieval**]: Content Analysis and Indexing—*Abstracting methods, Linguistic processing*; H.5 [**Information interfaces and presentation**]: Miscellaneous; I.2.7 [**Artificial intelligence**]: Natural Language Processing—*Discourse, Text analysis*

General Terms

Algorithm and Documentation

Keywords

Dynamic summarization, Text visualization, Meta-descriptors and topical descriptors identification, Text structure

1. INTRODUCTION

What is this text about? Which descriptors are relevant for indexing it and further retrieving its topics? Is this text relevant for my purpose, does it contain the answer I am looking for and in which part?

Even if these questions often characterize user's needs, answering them often requires a lot of times, as information retrieval systems seldom offer means for a rapid reading and a guided exploration of a document. Most work in text retrieval aims at presenting the information held by several texts [4, 28, 3] in order to give entry clues towards texts and to allow a navigation between them. Besides, less interest is dedicated to the definition of principles for accessing content of single documents. Jacquemin et al. [16] offer a 3D visualization of large documents, as PhD thesis, that encounters for the overall document logical organization and for the similar parts from a topic point of view. It remains to characterize each part of the documents more precisely in order to inform about their content, and this problem is mainly tackled in automatic summarization [24, 27]. Apart from the size and the number of documents they deal with, all these systems face up the same major problem: it is difficult, in textual information presentation, to represent text but with text [11]. The abstract view commonly assumed relies on the organization of the information according to a thematic criterion. Multi-document presentation systems visualize thematic classes of documents in 2D or 3D paradigms [11]. Nevertheless, the deeper an application has to examine single documents of relatively small size (10 to 20 pages), the more graphical means are difficult to use for abstracting text.

As most information retrieval systems return documents from an initial request made of words, a usual solution consists of presenting document titles and highlighting words of the request inside a passage or in the whole document (KWIC). Such a presentation does not allow a rapid reading and systems cannot satisfy themselves with it. If the TileBars interface of [12] entails a user to target relevant passages by adding to the document a graphical bar indicating the text segments that contain the request words, it is not enough for giving an idea of the whole text. Thus the fundamental problem consists in defining what information has to be shown so that a user will rapidly get rid of irrelevant texts or irrelevant passages and will be guide towards

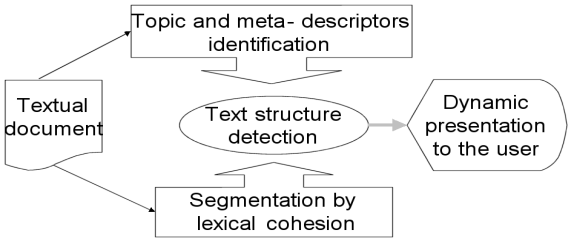

Figure 1: System overview.

relevant ones. Our studies lead us to provide indicative and informative view of texts as in summarization systems [24]. In these approaches, the indicative part corresponds to eliciting text topics, and the informative part to a summary obtained by extracting the prominent descriptive sentences. However whether such a static view of texts is relevant in a first approach of the texts in order to decide of keeping them or not, it has to be refined for guiding the reading.

This remark leads us to offer the user different levels of abstraction of a text. The first level is a global overview, where global topics are indicated and positioned in the text. We not only show the different segments as in [1, 17], but we place them in a thematic structure [26, 29, 23, 14]. The second level of abstraction goes deeper in the topic description by adding local topics and information about the argumentative role of the segments. This last information is given by selecting argumentative and rhetorical cue phrases in each segment, i.e. their meta-descriptors [15, 24, 27]. The third level gives the detail of chosen segments. In addition, the links between each entity are kept so that they allow navigating from a level to another, or, inside a same level, from a topic to another.

In this paper, after presenting an overview of our system, we will compare it to related works (section 3). Then we will detail how thematic descriptors and meta-descriptors are extracted (section 4) and how their characterization and they combination provides the segment structuring (section 5). We develop our model for scientific or technical texts, avoiding any handmade database and any restriction on the text domains. Finally, we present results we obtained with our approach in section 6, and make a conclusion.

2. SYSTEM OVERVIEW

Figure 1 shows the architecture of our system. A text is first pre-processed by a POS tagger because the segmentation task and the extraction phase rely on lemmatized words in order to bypass word inflections. The tagging also provides the morpho-syntactic category of each word.

The segmentation process leads to decompose the text in consecutive segments that are cohesive and coherent. As in [10], our method ([20]) relies on the repetition of same words when a same topic is developed, and on the concentration of these words in a part of text if they are typical of this local topic. Thus, it finds breaks between two basic units if they do not share enough words that mainly appear in these units. We will see in section 5.1 that we do not always interpret a cohesive segment as a topic segment. We only consider content words as unit descriptors (noun, verb and adjective), and we weight them according to the *TF.IDF* scheme, with *tf* the occurrence number of the word in the text, and *idf* the number of paragraphs that contain this word. We then evaluate the proximity of two units by computing the Dice coefficient between their two vectors and by applying a threshold relative to the variation of this value to the mean of all the coefficient values for each two consecutive units.

Scientific texts are generally structured in rather equilibrated paragraphs, and a paragraph tends to develop only one topic. Thus we chose them as basic units and text segment are then made of several consecutive paragraphs.

In order to exhibit the text structure, we identify in each segment their topic descriptors and their meta-descriptors. Topic descriptors are dynamically recognized, without any *a priori* on what are potential topics and which relations they may entertain. Topic descriptors are the noun phrases that are the most prominent at the segment level (local topics), and at the multi-segment level (global topics). At the multi-segment level, descriptors correspond to topics that are developed along several segments. As two consecutive segments differ in their local topics (by construction), we consider each segment develops a different point of view on the same global topic. Topic descriptors are mainly recognized by their occurrence number, incorporating in this computing an anaphora resolution process.

Meta-descriptors are recognized in texts from a list we acquired automatically from the whole corpus, considering all the expressions that are used in all the texts, independently of the domains, are those cue phrases of the general language that are used for highlighting a passage, or for making a link explicit. We will describe here the acquisition phase. These descriptors will play an explicative role in the visualisation phase, by specifying the argumentative role of the topic or the segment in the global text structure.

This global structure relies on the thematic segments and the links we found between them. These links come from the global topics they share. These topics can also be viewed as constituents of lexical chains that cover parts of texts made of several segments. Their distribution on the text leads to propose a text structure, and we will show we can exhibit different kinds of text structure.

The last process makes use of all this information, represented in XML format, and proposes a dynamic visualization of a text in different abstraction levels, with relations between all the entities our system recognizes: segments, global and local topics descriptors. This visualization allows the user a rapid reading of the whole text by presenting topics and their repartition in the thematic structure. If the user wants details about a specific part, she can request a summarization of a segment (or of linked segments) with the printing of local topics and meta-descriptors, or request the whole segment content(s).

3. BACKGROUND

For years, advantages of taking into account structures of texts have been shown in several areas such as information organization, summarization, and access to information.

Text Description

Boguraev et al. [1] introduce the notions of "capsule overview" – topically homogeneous text segment – and of "topic stamp" – most salient phrases of a segment under consideration – in order to offer slight but intelligible text overviews. They adopt a TextTiling-like algorithm [10] for text segmentation

and based the phrase salience definition on linguistic heuristics such as predominant grammatical roles, privileged positions in the sentence, phrase recurrence, etc..

Similarly, Kan et al. [17] perform a linear segmentation, but this time by using lexical chains. They do not intend to describe segment contents but to locate significant segments presenting key information about the article as a whole. For each segment, significance is computed by summing the relevance of terms occurring in it; it means the *TF*SF* (term frequency * how many segments does the term occur in) of each term and its coverage (a score following a harmonic series is added for each term occurrence) – both are normalized by the maximum value. In addition to segment significance, segment function is determined by text-genre cues such as global position, segment significance, lexical markers, etc..

In a sense, the two following approaches [25, 27] are similar to the previous one. Indeed, both of them aim at providing indicative information about text content, in particular about the argumentative structure of texts.

SumUM, [25], is an automatic two-steps summarizer for scientific documents based on selective text analysis. The selection is based on lexical and syntax recognition patterns such as *goal of author + define + [GOAL]* [1]. First, it provides an indicative view of the document, proposing major themes of the text, and then, an informative view by expanding a particular aspect selected by the user. Though this system obtained the best results in the evaluation program DUC 2002, it requires costly knowledge development.

The second approach [27] takes also interest in global rhetorical organization of scientific articles in order to make more coherent automatic summarization. They propose a learning-based method to identify and classify new "indicator markers". So used, those indicator markers permit to retrieve abstract-worthy sentences. But even if their results are encouraging, their approach to extract indicator phrases requires costly manual pre-processing.

Text Structure and Topic Hierarchy

Dominant discourse analysis theories [19, 8, 2] do not permit simple computational way to determine the detailed discourse structure of texts.

Starting from the definition of topicality based on lexical recurrence [9], some authors have proposed two kinds of robust approaches: the automatic construction of topic hierarchies and the topic text structure detection. The former are based on relations between terms like subsumption or lexical relations. A thorough review of automatic methods to build topic hierarchies is described in [18]. Although topic hierarchies provide good corpus overviews, they need to integrate more information for single text presentation in order to take more into account the text linearity and segment embedding.

Salton et al. [26] are one of the first work that observes topic structure by automatic means. They computed a graph of lexical similarities between paragraphs. Lexical homogeneity between contiguous paragraphs defines text segments and sets of mutually related paragraphs (more than 3) generate text themes. Text structure classification is defined by observing the congruence between text segments and text themes, and text traversal strategies are so defined according to.

[1] "GOAL" is the retrieved expression marked by "goal of author + define"

In comparison with this system, the following two systems have been designed to detect specific text structure. Even if they present correct results they remain text-genre dependant.

Yaari [29] proposes a method for segmenting expository texts based on bottom-up hierarchical agglomerative clustering (HAC). More than a linear segmentation, the algorithm identifies a nested outline of the text. Based on paragraphs as elementary segments, the HAC algorithm merges iteratively the two most lexically similar consecutive segments while more than one segment remains.

In [14] we got interested in nested thematic structures of scientific texts. In order to detect them, we sought for the two most lexically similar non-consecutive segments – this permit us to localize the theme inside those passages – and we reproduced the same process recursively on embedded segments. This algorithm presents correct results over texts where announcements and resumptions are common.

Moens et al. [23] propose to observe generic topical cues, amongst which the content terms, to automatically detect the thematic structure of any given text. They argue that combination of information about lexical chains and topically coherent passages – consecutive sentences sharing the same topical term recognized thanks to cues such as the position in the sentence, the persistency with previous sentence, bound pronouns, etc. – helps in finding the most probable thematic structure (detecting topic shifts, nested topics, sequential topics, topic returns). The authors give one example of heuristic they used: topic returns can be detected as large gaps in the positions of two members of a lexical chain that are into topical focus. Their approach goes further by generating a table of contents of the text by selecting most informative terms in the lexical chain of dominent sentences of each segment.

4. TEXT DESCRIPTORS

We said that a text is fully described if the topics of each of its parts are identified, if the role of each segment can be understood and if the overall organization is made explicit. For this purpose, we defined two kinds of descriptors, topic descriptors and meta-descriptors, for characterizing functional or argumentative roles of segments.

4.1 Meta-descriptors

Meta-descriptors are those expressions an author uses to make explicit an argumentative link between parts of text, in order to indicate which role plays a segment or a topic in the global discourse. These roles are such as introduction, result, conclusion. Meta-descriptors are also those rhetorical expressions that enable to highlight an important passage.

Meta-descriptors correspond to formulaic phrases, either fixed phrases as *in the same way* or *in particular*, or flexible phrases, as *have to be implemented* or *the goal of* that can be found in text with slight variations. These expressions do not depend on text topics and are recurrent in all the documents of a scientific corpus. This last characteristic leads us to extract them automatically from our corpus rather than building a list manually. The automatic process entails to collect expressions without any *a priori* about their validity as meta-descriptor, and, as we will see further, without any *a priori* about the form of the meta-descriptor.

As shown by the previous examples, meta-descriptors do not correspond to a fixed category of phrase and their ex-

F	Expressions	N	Comment
158	example ,	126	Not a candidate term
142	for example	113	Not a candidate term
135	for example ,	108	Not a candidate term
130	. for example	104	Not a candidate term
127	. for example ,	101	Is a candidate term !
102	example of	81	Is a candidate term !
84	example , the	67	... ?
76	an example	60	... ?

Table 1: Automatic filtering.

traction cannot be achieved by the application of syntactic patterns. So we choose to work with n-grams, i.e. sequences of n consecutive tokens, made of 1 up to 5 tokens. Tokens are those provided by a POS tagger, the TreeTagger [13]. The extraction process follows two phases: i) the collecting phase, ii) the selection phase. The collecting phase consists of making a list of unique instances of n-grams from each text, and to count the number of documents that hold each of them. A threshold allows the system to eliminate the less frequent expressions, thus the words linked to the text domain. This first step produces a first list that is sufficiently relevant for a manual selection. Nevertheless, the second step aims at selecting automatically the right length of each expression. We assume that if two expressions, with one enclosing the other, are in the same neighborhood in term of frequency, they both refer to the same meta-descriptor. The smaller expression corresponds to a variation of the longest expression that is then chosen. We made several tests (see [15]), and we finally retained expressions with a frequency equal or greater to 7 after the first step. Closeness of two expressions is stated when their difference of frequency is less than 20% of the greater number.

Table 1 shows the filtering step with F for the frequency of the expression, N for the inferior boundary of the interval that defines the neighborhood. One can see that *example,* is embodied in *for example,* and as they are in the same neighborhood (135 belongs to the interval [126, 158]), *example,* is no more considered as a candidate. By applying this principle gradually, we finally select the longest expression along all the embodied expressions in the same neighborhood (by transitivity). In the series shown in table 1, we finally keep *. for example,*. It does not signify that we always select the longest expression; if two of them are not close enough, they are both kept. We can note the important role played by the punctuation when defining an expression. Punctuation translates a constraint on the position of the meta-descriptor in a sentence.

The final list we obtained contains 1063 markers from a corpus of 80 scientific articles containing each one about 5000 tokens. Table 2 shows an excerpt of the list (words of the meta-descriptors are lemmatized). Another interest of our approach is the possibility of considering some tokens as flexible depending on their part-of-speech assignation (determiner, pronoun, adverb, etc.). This permits us to generalize forms and to bring together several variants of a same form.

When analyzing a text, we recognize all the meta-descriptors in it. They represent a great part of its total vocabulary (between 30% and 40%). Then, we will present in section 6 some heuristics leading to filter them, according to their contextual use.

by mean of	point of view	in particular
combination of	. the result	as follow :
the representation of	this approach be	the output of
the basis for	rate of	in case of
identification of	in this section ,	it be necessary
have be implement	be obtain by	be implement in
available .	a theory of	as oppose to
as far as	a result of	a consequence
a collection	the meaning of the	an algorithm
the sequence of	. this mean that	the degree
the default	. in this paper we	see section
sort of	focus on	the condition on
the model	. the method	the long
the importance of	the goal of	the figure
property of the	the ability to	suggest a
solution be	so as	small number of
show to	in the sense	in the same way

Table 2: Excerpt of our final list of fixed phrases

4.2 Topic Descriptors

Topic descriptors aim at identifying both the topics the whole text is about, and the topics parts of texts are about. This becomes essential when considering different application tasks, as indexing a text or searching for information either by searching documentation about some topics or looking for a precise information. Text topics are extracted from the text currently analyzed for presentation and typically correspond to noun phrases of the text, considering these expressions are the better descriptors for topics, as in [1]. The topic extraction is not based on a pre-built list of selected topics, as for meta-descriptors, as a topic can be relevant in some text and not in other texts, according to the global thematic and the discourse organization. Thus, it is a dynamic process, based on the recurrence and the distribution of the noun phrases inside the thematic segments. Recurrence is the main criterion for deciding of the relevance of a term as for segmenting texts [10]. It is justified when dealing with scientific texts where developing topics needs to make use of a specific vocabulary without synonyms or ambiguities. The distribution criterion is not used for weighting the selected terms as in text segmentation. However, the interpretation given to a topic distributed over several segments remains the same. Thus, frequent terms mainly present in one segment are characterized as local topics. When they appear in several segments, they are considered as global topics, i.e. topics that typify a greater passage. These characterizations will lead us to propose a discourse structure that links non consecutive segments in a same thematic (see section 5). As we just saw it, term recurrence is a strong criterion for deciding which noun phrase identifies a topic. For this reason, recurrence is not only computed by counting identical terms and the system also accounts for their anaphoric references.

Topic extraction follows the steps illustrated in figure 2. The first step aims at retrieving topic referring expressions that denote the text entities by applying syntactic patterns. These patterns identify simple noun phrases (a noun with eventually adjectives), complex noun phrases (phrases with the preposition of) and pronouns (personal, reflexive and demonstrative pronouns). Among these expressions, the system tries to solve the anaphoric expressions made of pronouns and demonstrative noun phrases. We chose to work only on the non ambiguous anaphoric expressions, as our

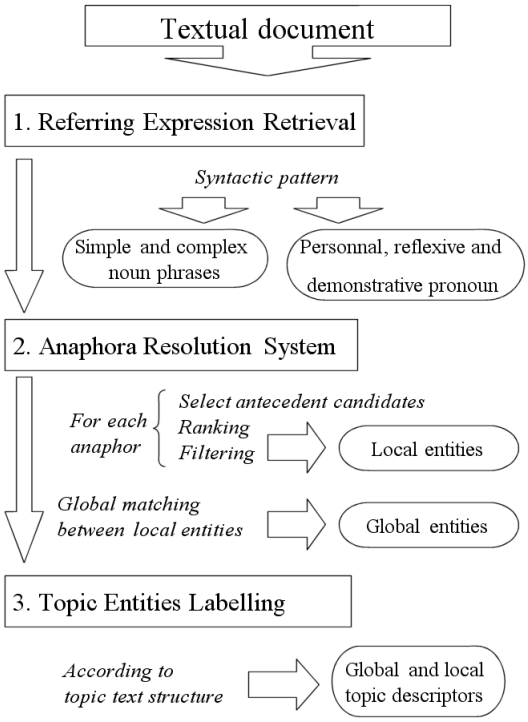

Figure 2: Identification of topic descriptors.

purpose is to reinforce the pregnancy of some entities when they are used again and not to solve all the co-references of the text. So, we prefer precision over recall in order to avoid noisy results.

The anaphora resolution process was inspired from [1, 22]. It first selects the potential antecedents for a given anaphora in the current sentence and its prior sentence. The system keeps all the noun phrases that are present or referred in these sentences. Then, candidates are ranked according to several criteria that lead to give them a weight: i) morphological criterion (gender and number agreement); ii) lexical indicators (identical head of noun phrases); iii) syntactic criteria (parallel grammatical relations, kinds of grammatical relations); iv) discursive criterion (the distance between a candidate antecedent and the anaphora in scope). Most of these criteria are computed by robust heuristics, as the relative position of the words for deciding their grammatical role for example.

The last step consists in labelling the salient entities with a topic stamp. The topic label is chosen according to a threshold that states as topical the recurrent entities and the entities which are at least one time subject, object, or which are coming along with demonstrative. These topical entities are all considered to be local topics except those occurring in several segments which are named global or hyper-global if they are repeated throughout a text.

5. TOPICAL TEXT STRUCTURE

We propose a general framework to analyze topic structure of scientific and technical texts. This framework should be refined to handle particular text segments.

5.1 Text Segmentation

In traditional text segmentation methods based on lexical recurrence [9, 10], cohesive segments are considered coherent from a thematic point of view and different because of the theme they carry. Less cohesive text parts are considered as topic shift.

Ferret [5] interprets the various cohesive parts as different degree of topic shifting: new topic detection, topic development, topic shift detection and topic shift.

In this paper we argue that lexical cohesion does not always stand for topical assignment. Indeed, an example can lexically differ from the topic it aims to instance. So we will consider segments obtained by lexical cohesion simply as coherent segments for which the role remains to specify. In addition, we argue that less cohesive segments can stand for articulation between topic segments.

5.2 Model and Basic Structures

Popular theories of discourse structure model the document as a hierarchical structure of text segments, e.g. the nucleus-satellite relation in rhetorical structure theory [19] and the dependence relation in [8].

We start from this axiom and we distinguish two kinds of segments: the topical segments and the meta-discursive ones. The former class represents topic contents and the latter acts as an articulation between topics. The articulation segments are divided into two subclasses: the announcement and the conclusion segments. An announcement segment announces explicitly or implicitly the themes or aspects which are going to be dealt with, and a conclusion segment summarizes or concludes one or several previous topic segments.

From our corpus observations, we identified two basic structures in which the various kinds of segments are related. The detection of these structures and the interpretation of their combination permit to describe the topic structure of the texts.

Contiguous topic segments stand for items of a list which should be analyzed in respect of a more global structure. In our approach, we always consider them as explicitly or implicitly introduced by a subsuming structure.

Nested segment(s) are used to specify a dependence relation. A segment is said dependent to another if it requires that the other exists in the text. We make the distinction between the recursive nesting of single segments and the nesting of several segments.

This lead us to define the following text grammar rules[2]:

- Str → Sa Str+ Sc?
- Str → St

A text structure can be parsed as a list of nested structures beginning by an announcement and potentially finishing by a conclusion.

Figure 3 shows some example of basic structures. For examples, (a) can stand for elaboration or instantiation of a single topic as well as the reprise of the focus of a segment as the theme of the following; (b) concerns various aspects of a same topic, or distinct topic developments.

[2]With *Sa* for announcement segment, *Sc* for conclusion segment, *St* for topic segment and *Str* for the structure symbol.

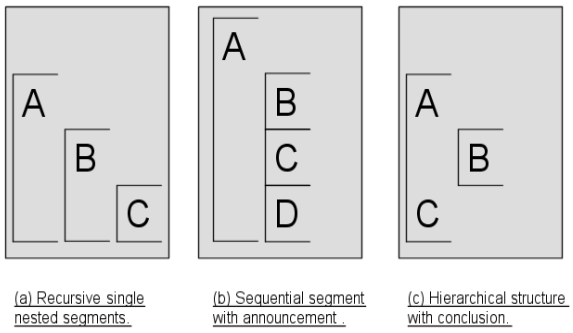

Figure 3: Various examples of basic text structure composition.

5.3 Structure Detection

Identification of the segment roles is performed by the detection of some intrinsic and extrinsic properties. We noticed that topic segments are much more lexically homogeneous than articulation segments in which global themes are more likely combined (at least two of them) and in which meta-descriptors presence is higher. Thanks to meta-descriptive lexicons and works of [21], we defined heuristics and patterns to detect announcements, conclusions and topic segments.

The structure algorithm is based on the correct assignment of announcement and conclusion segments with the topic segments. This is done by matching the same global topics between topics segments and announcement or conclusion segments respectively to the previous and following text parts. We suppose the text structure follows the text linearity so the text is back-parsed from the end to the beginning of the text.

Single nested segments are identified in the situation of two adjacent segments with the first one linked to an announcement (by having a common global theme): if the second segment has the same global theme or if it is not related to any announcements, then it is considered as a single nested segment of the previous one.

5.4 Going further...

In practice, roles of text segment are not so easily identified. It is the major problem of text segmentation methods based on lexical cohesion. Many segments are ambiguous and regularly we encounter segments concatenating both a conclusion and an announcement, or series of segments in which the first one contains the announcement of the series.

Text segmentation needs finer-grained observation. Our approach is dedicated to provide a generic model to analyze topic text structure. We argue that most of the ambiguous segments follow the same hierarchical structure that we have defined in previous sub-section. So the enhancement of the system consists of observing new ambiguous segments and setting heuristics to recognize them in order to consider respectively announcement, conclusion and topic parts. Combination of statistical and linguistic methods [7] also seems to be an interesting approach to decompose texts into segments and to assign them more accurate roles.

6. EXAMPLE AND DISCUSSION

The main issue for short text presentations is to combine both abstract and textual representations of documents. In order to limit cognitive processes and make easier the users's task, the interface should provide all and only the needed information on a same screen and provide ways of linking the different information sources.

Our corpus is composed of articles extracted from the *Computation and Language (cmp-lg)* collection, developed by the *TIPSTER SUMMAC* project (http://xxx.lanl.gov/cmp-lg), from French proceedings of TALN conference and from the French scientific news paper called *"la Recherche"*.

Currently, due to our running task for adapting the process to English, we do not have a full text description in English. The example in figure 4 is an article of *"la Recherche"*. The article deals with *"vin jaune"* or *"yellow wine"*. The author wonders which molecule is responsible of its specific flavor. In particular, the fragment presents the most straight method of wine analysis.

Our current presentation is a prototype. The left part of the screen stands for an abstract view of the text. It corresponds to the text topic structure in which global terms are mentioned. Each level presents the global terms from segments which are not embedded (nominal expressions, laying in context several global terms, are preferred). This structure offers accesses to the text content.

The right part of the screen refers to the segment under consideration. Thanks to our term labelling process, we identify in this text segment *"wine/vin"* as playing a global thematic role, which sounds as normal because the entire text is about wine. In the same way, we identify *mixture/mélange* and *element/composé* as playing local theme role. Both reinforce the role played by expressions identified as indicative label, it means *analysis/analyse* and *technique/technique*. As a note, we point out that a global theme and a meta-descriptor occur in the first sentence of the segment and that they describe correctly the text fragment since we can type it as a description of *technical analysis of wine*. Combination of meta-descriptors with topic descriptors within a same linguistic expression, position in topic segments, etc. can be used as clues to select some significant meta-descriptors among others.

It is difficult to evaluate this kind of systems. In general either each sub-system is evaluated separately or the whole system is used for a retrieval task to determine whether it facilitates searching tasks.

Our system is composed of many parts. The main dependency risk relies on genre-dependent text. Text segmentation by lexical recurrence shows some limitations with journalistic text-genre documents. Nevertheless Ferret et al. [6] have shown that using a co-occurrence network can solve this kind of problem. Concerning topic identification, Mitkov [22] shows correct results for robust pronoun resolution in scientific texts. The problem should be more difficult for other text genre. Indeed journalistic documents do not use words repetition but synonymous and metaphors. It is known that some semantic knowledge resources are required to deal with this kind of documents. We just have started to test our meta-descriptor extractor over different sets of scientific documents. Analysis requires more data, but in comparison with scientific meta-descriptor lists, a list obtained from a journalistic corpus seems to prefer some specific parts of speech like adjectives and adverbs.

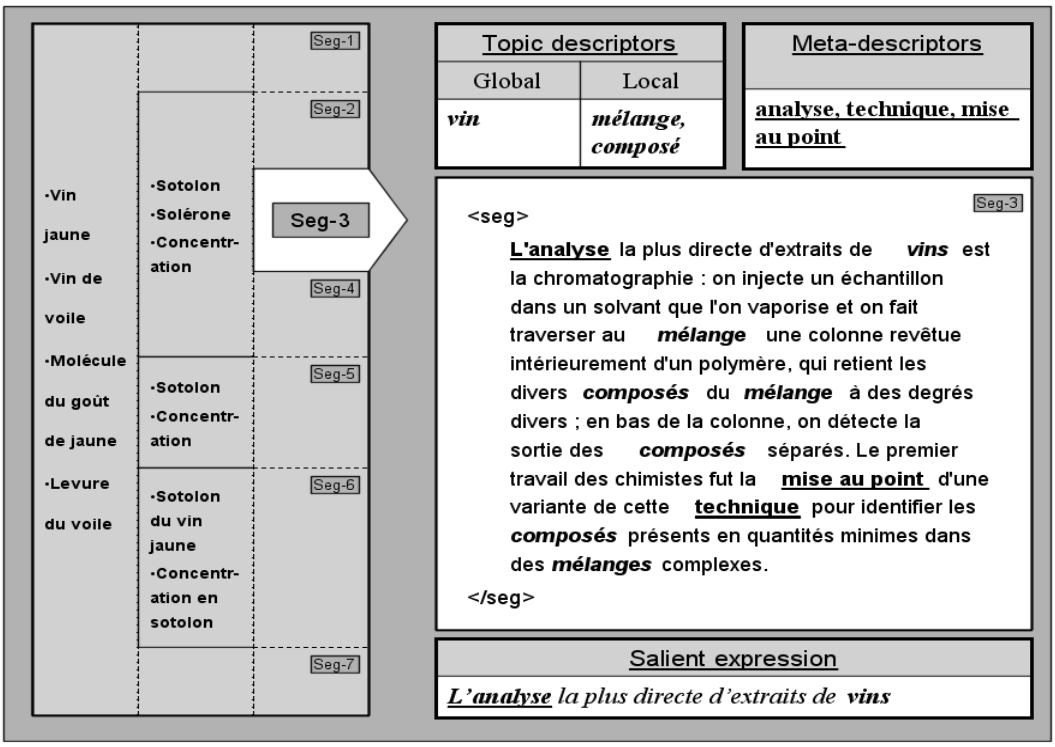

Figure 4: Example of segment description.

7. CONCLUSION AND PERSPECTIVES

Our system aims at facilitating visualization of documents and navigation intra-documents. Our approach is supported by automatic text analysis to identify topic descriptors and meta-descriptors, which are used to detect the topical structure of texts. Moreover we propose a generic framework to detect text topic structures.

Future developments will concern on one hand to enhance our text structure detection by integrating more markers to identify correctly the various segments, and on another hand to set an evaluation framework to test how significant our analysis can help a user in retrieval task.

8. REFERENCES

[1] B. Boguraev and C. Kennedy. Salience-based content characterisation of text documents. In *Proceedings of the ACL'97/EACL'97 Workshop on Intelligent Scalable Text Summarization*, pages pp. 2–9. Madrid, Spain, July 11th 1997.

[2] M. Charolles. L'encadrement du discours - univers, champs, domaines et espaces. *Cahier de recherche linguistique*, 6, 1997.

[3] H. Chen, A. Houston, R. Sewell, and B. Schatz. Internet browsing and searching: user evaluations of category map and concept space techniques. *Journal of the American Society for Information Science, Special Issue on "AI Techniques for Emerging Information Systems Applications"*, 49(7):pp. 582–603, 1998.

[4] D. R. Cutting, D. R. Karger, J. O. Pedersen, and J. W. Tukey. Scattergather: A cluster-based approach to browsing large document collections. In *SIGIR'92*, Denmark, June 1992. ACM.

[5] O. Ferret. Segmenter et structurer thmatiquement des textes par l'utilisation conjointe de collocations et de la rcurrence lexicale. In *TALN*, Nancy, 24-27 juin 2002 2002.

[6] O. Ferret, B. Grau, and N. Masson. Utilisation d'un réseau de cooccurrences lexicales pour améliorer une analyse thématique fondée sur la distribution des mots. In *ISKO*, 1997.

[7] O. Ferret, B. Grau, J.-L. Minel, and S. Porhiel. Repérage de structures thématiques dans des textes. In *TALN*, Tours, 2001.

[8] B. J. Grosz, A. K. Joshi, and S. Weinstein. Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, 21(2):pp. 203–225, 1995.

[9] M. A. K. Halliday and R. Hasan. *Cohesion in English*. London, 1976.

[10] M. Hearst. Texttiling: Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, 23(1):pp. 33–64,, March 1997.

[11] M. Hearst. User interfaces and visualization. In R. Baeta-Yates and B. Ribeiro-Neto, editors, *Modern Information Retrieval*, pages pp. 257–322. Addison-Wesley, 1999.

[12] M. A. Hearst. Improving full-text precision on short queries using simple constraints. In *Proceedings of the Symposium on Document Analysis and Information Retrieval*, April 1996.

[13] Helmut.Schmid. Probabilistic part-of-speech tagging

using decision trees. In *Proceedings of the Conference on New Methods in Language Processing*, Manchester, UK, 1994.

[14] N. Hernandez and B. Grau. Analyse thématique du discours : segmentation, structuration, description et représentation. In *CIDE'05*, Hammamet, Tunisie, 20-23 octobre. 2002.

[15] N. Hernandez and B. Grau. Extraction et typage de termes significatifs pour la description de textes. In *Congrès du chapitre franais de l'ISKO (International society for knowledge organization)*, Grenoble, 3 et 4 juillet 2003. [à paraître].

[16] C. Jacquemin and M. Jardino. Multi-dimensional and multi-scale visualizer of large xml documents. In *Proceedings of EUROGRAPHICS*, Saarbrcken, Germany, 2002.

[17] M.-Y. Kan, J. L. Klavans, and K. R. McKeown. Linear segmentation and segment relevence. In *Proceedings of 6th International Workshop of Very Large Corpora (WVLC-6)*, pages pp. 197–205, Montréal, Québec, Canada, August 1998.

[18] D. J. Lawrie and W. B. Croft. Discovering and comparing hierarchies. In *Proceedings of RIAO 2000 Conference*, pages pp. 314–330, Paris, April 12-14 2000.

[19] W. C. Mann and S. A. Thompson. Rhetorical structure theory: A theory of test organisation. Technical report isi/rs-87-190, Information Sciences Intitute, 4676 Admiralty Way, Marina del Rey, California 90290-6685,, June 1987.

[20] N. Masson. *Méthodes pour une génération variable de résumé automatique : vers un système de réduction de texte*. PhD thesis, Universit Paris XI, 1998.

[21] J.-L. Minel, J.-P. Desclés, E. Cartier, G. Crispino, S. B. Hazez, and A. Jackiewicz. Résumé automatique par filtrage sémantique d'informations dans des textes. présentation de la plate-forme filtext. *Revue Technique et Science Informatique*, 3, 2001.

[22] R. Mitkov. Robust pronoun resolution with limited knowledge. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, COLING-ACL'98*, August 1998.

[23] M.-F. Moens and R. D. Busser. Generic topic segmentation of document texts. In *Proceedings of the 24th ACM SIGIR Annual International Conference on Research and Development in Information Retrieval*, pages pp. 418–419, New York, 2001.

[24] H. Saggion and G. Lapalme. Concept identification and presentation in the context of technical text summarization. In A. for Computational Linguistics, editor, *Proceedings of the Workshop on Automatic Summarization. ANLP-NAACL2000*, Seattle, WA, USA, 2000.

[25] H. Saggion and G. Lapalme. Selective analysis for automatic abstracting: Evaluating indicativeness and acceptability. In *RIAO*, Paris, France, 2000.

[26] G. Salton, A. Singhal, C. Buckley, and M. Mitra. Automatic text decomposition using text segments and text themes. In A. Press, editor, *Proceedings of Hypertext'96.*, pages pp. 53–65, New York, 1996.

[27] S. Teufel and M. Moens. Argumentative classification of extracted sentences as a first step towards flexible abstracting. In I. Mani and M. Maybury, editors, *Advances in automatic Text Summarization*. MIT Press, 1999.

[28] J. A. Wise, J. J. Thomas, K. Pennock, D. Lantrip, M. Pottier, A. Schur, and V. Crow. Visualizing the non-visual: Spatial analysis and interaction with information from text documents. In I. C. S. Press, editor, *Proceeding of Information Visualization Symposium*, pages pp. 51–58, 1995.

[29] Y. Yaari. Segmentation of expository texts by hierarchical agglomerative clustering.