Text Linguistics and Navigation. Questions about Text

Javier Couto, Jean-Luc Minel

Text Linguistics and Navigation:

Questions about Text

Javier Couto* — Jean-Luc Minel**

* *Universidad de la República – Facultad de Ingeniería*
INCO
J. Herrera y Reissig 565, Montevideo, Uruguay
jcouto@fing.edu.uy

** *MoDyCo, UMR7114, CNRS*
Université Paris X
200 Avenue de la République, France
Jean-Luc.Minel@u-paris10.fr

ABSTRACT. *In this paper we address the problem of accessing text information by text navigation. We present an approach to text navigation conceived as a cognitive process exploiting linguistic information present in texts. We claim that the navigational knowledge involved in this process can be modeled in a declarative way with the Sextant language. Since this language refers exhaustively to specific linguistic phenomena, we define a customized text representation. These different components have been implemented in the text navigation system NaviTexte. NaviLire, an application of NaviTexte is described.*

KEYWORDS *Assisted navigation of texts, navigation knowledge management, text model.*

1. Introduction

For the past thirty years, text linguistic researchers have worked on describing linguistic marks of textual coherence in order to bring out principles of text structuring [Lundquist 1980]. A set of concepts and models of textual interpretation has been worked out, including for example, anaphors, connectors, mental spaces, etc. In particular, these studies have shown that even for languages apparently close like French and Danish, texts are not organized in the same way [Lundquist 2006]. Consequently, text linguistics has important implications in foreign language teaching, especially from a contrastive point of view, when language pairs are analyzed through texts used in authentic communication situations. It seems that the teaching of text linguistics contributes to sharpen the attention of students towards the building of well-formed texts and to stimulate their own production of texts. Consequently, a tool that allows the student (reader) to perceive textual units that contribute to and maintain text coherence and to navigate between them in a text, can be supposed to be an outstanding didactic tool for teaching reading of foreign language texts, as well as producing written texts in the foreign language.

From the point of view of many Natural Language Processing (NLP) tools, a user seeks to accomplish a specific task. These tools anticipate a set of possible interactions with the user, often fixed, to assist her/his work. This is the case, for example, of several automatic summarization, question-answering and information retrieval systems, among others. Nevertheless, some information needs cannot be satisfied, a priori, by a standard automatic summary, or cannot being expressed in terms of a precise question or a query in a given language. It is in this use scenario, of vague but real information needs, where **text navigation** appears as an interesting alternative to traditional systems. Rather than resolve a specific task, a text navigation system offers to the user a suite of search and mining tools to find the needed information.

The notion of text navigation has evolved through time. Nevertheless, this term usually refers to hypertext systems, which offer the possibility to activate hyperlinks, moving the reading point from a text unit source to another one, the target, this change being intra or intertextual. The classic hypertext conception presents some limitations. First, the hyperlink activation is not assisted. In other words, imprecise, poor or no information is provided to the reader before s/he activates the link. Second, the reader does not know where the movement will be carried out in the text (before or after the reading point or outside the text), which generates the lost in hyperspace problem [Elm & al. 1985, Edwards & al. 1989]. Finally, hyperlinks are embedded in the hypertext. Therefore, there is no clearly distinction between text constituents and navigation knowledge. In addition, by not explicitly modeling this knowledge, it is not reusable.

Different solutions have been proposed to address the problems mentioned. Some researchers [Danielson 2002] have tried to mitigate the lost in hyperspace

problem offering global maps where the reading point is clearly identified, by showing it in context. Adaptive hypertext [Mathe 1994, Brusilovsky 1996] relying on user model, proposes to modify the way the text is shown on the screen. For example, [Zellweger et al. 1998] provides additional information at a link source to assist readers in link activation. Dynamic hypertext [Bodner 1999] computes the value of hyperlinks using several criteria such as text similarity or predefined relations. In this approach, a hyperlink is defined as a query returning a text node. With the raise of the Semantic Web, the idea of augmenting nodes by exploiting ontologies has been explored, and interesting results and applications has been achieved [Domingue 2004, Bechhofer 2006] .

In some way, our conception of text navigation is related to the notion of computed query, but rather than taking into account criteria depending on the reader, the target is computed by exploiting linguistic information in texts. Moreover, the queries are not placed in texts but they are encapsulated as knowledge by a specific language (Sextant), which allows isolating the navigational knowledge to create knowledge bases. This language may use external resources such as ontologies, but the way the resources are used must be explicitly modeled not automatically inferred from them). Both texts and queries (navigational knowledge) are interpreted by NaviTexte, which manages the interactions with a reader.

The remainder of this paper is organized as follows. In the next section, we discuss our approach to text navigation. The third section describes text representation and the fourth one a navigational knowledge modeling language called Sextant. The fifth section details the text navigation system NaviTexte. The sixth section describes the NaviLire application. At last, conclusions are presented.

1.1. Defining text navigation

Our conception of text navigation lies in the hypothesis that navigating through texts is the expression of a cognitive process related to specific knowledge [Minel 2003, Couto & al. 2006]. More precisely: we claim that a reader moves through texts applying some knowledge to exploit linguistic information present in texts (discursive markers). Moreover, we claim that this knowledge may be articulated in a declarative way relying on information in texts, coded, on the one hand, by its structure, and, on the other hand, by specific annotations.

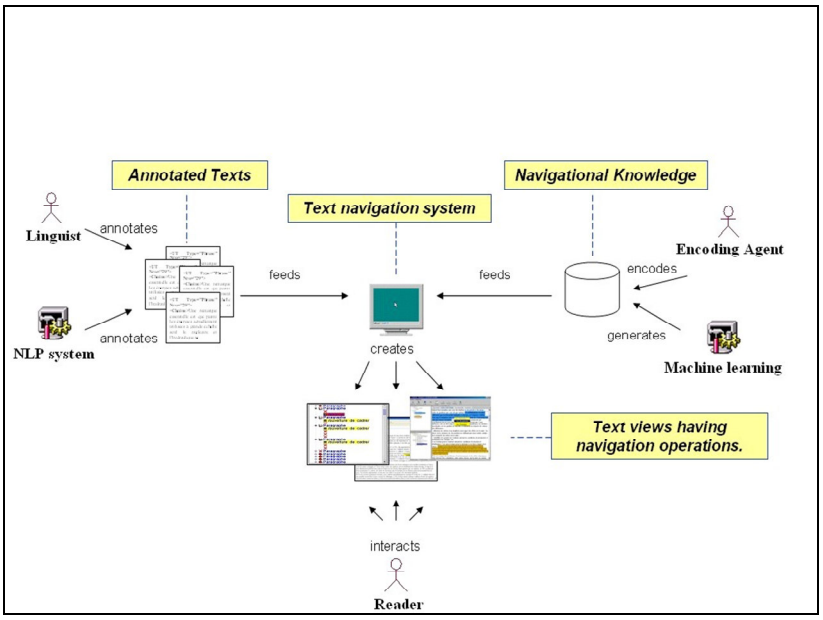

Figure 1. *Elements of text navigation*

The main difference between classic hypertext and our conception of navigation lies on the status of texts and on the definition of navigational knowledge. In the case of hypertext, the visualization of a text is unique and navigational knowledge is encoded (embedded) in the text. In our approach, there are several ways to visualize a text [Couto 2006], each way called text view, and for each view, different navigational knowledge may be defined. Several views may coexist in a text navigation system. As a consequence, the navigation is not guided by the author, compared to hypertext navigation where s/he determines the links, but it is the result of an interpretation process made by the reader, by choosing pertinent views and relying on text structure and annotations.

Our conception of navigation (*cf.* Fig.1) relies on four elements: i) a text representation allowing linguistic specific phenomena and annotations; ii) a language to model navigational knowledge; iii) an agent (an individual or a software) able to encode such knowledge; iv) a system, called NaviTexte, to interpret and apply knowledge to a specific text.

2. Text Representation

Text representation is a recurrent issue in NLP. Computer text processing inevitably needs a text representation to work with. The kind of processing determines usually the most appropriate representation according to some criteria: cost/performance ratio, flexibility, exhaustivity, simplicity, etc. Structured text

representations, widely used nowadays, generally adopt a hierarchical approach where syntactic, semantic, discursive, and page format aspects usually coexists. The Text Encoding Initiative [TEI] is emblematic of this kind of representations, that we believe not appropriate to our approach. On the one hand, we do not want to mix up formatting aspects with text constituents ones. On the other hand, some authors [Webber 2003, Wolf 2005] have criticized the choice of tree representation to model some discourse phenomena. For example, in [Wolf 2005] the authors have shown that graph structures are required to represent, in some cases, discourse coherence. Others examples, as anaphoric relations, show that there exists discourse phenomena not modelable with hierarchical structures.

2.1. *A More Comprehensive Representation*

The limitations of existing representations have motivated the definition of a customized text representation [Couto 2006], inspired by [Crispino 2003, TEI]. Its four main goals are: i) to not restrict the text units type to a predefined set (section, paragraph, phrase...); ii) to offer at the same time a hierarchical organization of text units and an organization allowing the expression of non-hierarchical relations; iii) to consider the titles as text units not as text units' attributes); iv) that all text units hierarchical, non hierarchical and title units may have an unlimited number of annotations of any kind, able to propagate to other text units according to annotation heritage criteria.

2.2. *The Hierarchical Relations*

Our text representation is based on typed units (TU), whose type may be freely defined. Evidently, the types to use by the encoder of navigational knowledge should be coherent to the types defined in texts. The description of a text is made up of two parts: the head and the body. In the body, TU are hierarchical arranged and each TU has three attributes that establish its key: type, number and, optionally, level. Each TU may have an unlimited number of attributes called annotation. Only units without subordinates in the hierarchy have an attribute called string, which represents its lexical string.

2.3. *The Non-hierarchical Relations*

To represent non hierarchical relations, we can define new elements in the head, using predefined constructors applied to existing TU in the body. Four constructors are available: *Set*, *Sequence*, *Reference* and *Graph*. A *Set* is a collection of TU for which there exists an equivalence relation from the point of view of the annotator. For example, TU with different POS values can express the same topic (like verb and noun phrase) or different noun phrases refer the same named entity. A *Sequence* is an ordered collection of TU for which a relation of syntactic or semantic cohesion exists. For example, in the body it is not possible to represent, as a unique structure,

discontinuous enumeration, like "First, (…) Second," which can be achieved with a *Sequence*.

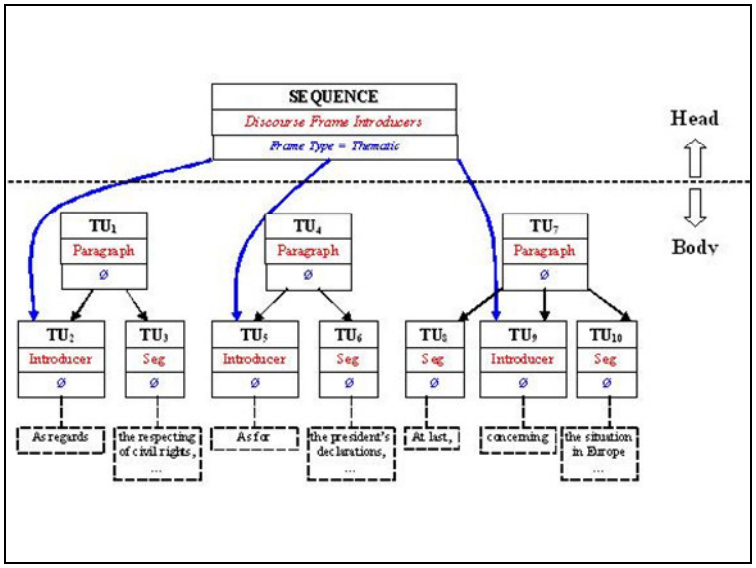

Figure 2. *Example of a Sequence of TU*

Figure 2 presents a diagram that shows the creation of a *Sequence* for discourse frame introducers [Charolles 1997]. A *Reference* defines an oriented relation between two TU. The representation of the association between a pronominal anaphora and its referent or the nucleus-satellite relations in the RST [Mann & al. 1987] are typical examples of utilization. At last, a *Graph* is used to build multiple relations between TU. This structure is very useful to represent complex discourse structures like coherence tracks, different levels of indirection in indirect discourse , the manifestation of feelings of a novel character [Mathieu 2005] or a thematic index like those presented at the end of books.

We would like to state that units defined by the constructors may have its own annotations and are treated as units by the modeling language Sextant. This means that we can directly manipulate them, for example to move through different constructed units or trough the elements of a specific one.

3. Modeling Navigational Knowledge: the Sextant Language

To allow the unambiguous isolation of navigational knowledge we need a formal modeling language. We want to model the knowledge applied by a reader to move through texts, claiming that this knowledge exploits linguistic information present in texts. We do not say that this is the only way a reader may move through texts, but

we say that this is the kind of way that we are going to model. Thus, our language must rely on the text representation presented in the precedent section.

3.1. *Knowledge Modules ant Text views*

A text view may be a full view or a partial view focused in some specific phenomena present in the text (for example a view of all discourse referents). The constituent elements of a view are formalized in a view description, which contains the type of view, its parameters (depending on the type), the creation constraints conditions to verify by the TU of the view) and the navigation operations (see next section). At present, four types of view have been defined: plaintext, tree, graph and temporality. The three firsts types are described in [Couto 2006]. The last one graphically represents temporality in texts and a complete description may be found in [Battistelli 2007]. Several view descriptions may be gathered by the encoder in an entity called navigational knowledge module. The creation of a view may be conceptualized as the application of a view description to a specific text. Thus, the application of a module implies the creation of a set of text views.

3.2. *Expressing operations of navigation*

The notion of computed query mentioned in section 1 is formalized in Sextant as a navigation operation, which links a source TU to a target TU. In classic hypertext systems one must explicitly connect the specific source to the specific target. For example, if a reader wants, for all definitions in a scientific paper, to move from one to the following one, several hyperlinks must be defined. In our approach we specify the source and the target using conditions. As a result, we can abstract, for example, the navigational knowledge that states "go from one definition to the following one", being "definition" one of the TU annotations.

We can specify several conditions for the source and the target. We say that a navigation operation is available for a TU if this TU verifies the source conditions. A text navigation system should find the TU that verifies the target conditions. As several TU in the text may verify them, we need a way of disambiguation. This is done by the orientation parameter, which specifies the direction of the target search by using one of these options: *first*, *last*, *forward(i)*, *backward(i). First* and *last* indicate that the search of the target is absolute: the TU to select will be the first (respectively the last) TU that verify the conditions. *Forward*(i) and *backward(i)* indicate that the search is carried out relatively to the source (before or after) and indexed by the integer i. For example, "forward(3)" is interpreted as the third TU, after the source, of all the TU whose attributes verify the target conditions.

3.3. *The conditions language*

The conditions language is an important component of Sextant ant it is composed by basic conditions, TU elements existence conditions, hierarchical conditions and non-hierarchical conditions.

Basic conditions concern TU's attributes and annotations. For this kind of condition we use a notation close to the pattern notion. We define an operator called TU, having five operands that correspond to the following properties: type, number, level, annotations and string. With the three first operands and the fifth one, we denote constraints of equality, inequality, order, prefix, suffix and substring occurrence. The fourth operand is used to indicate the existence or non-existence of annotations, whether it is an annotation name, a value or a name-value pair. For TU elements existence conditions, we define operators without operands to verify if a TU has annotations, string, title, parent and children. All conditions may be combined using the classic logic operators OR, AND and NOT. Figure 3 presents a generic example of an operation navigation.

IF (condition UT_{source})
THEN : DO SELECT CRITERIA (Orientation, Ordre)
WHERE {(condition UT_{cible})
AND
(Relation (UT_{source}, UT_{cible})
} ;
: DO SHOW (Libellé de l'Opération) ;

Figure 3. *Generic navigation opération.*

4. NaviTexte: a Text Navigation System

Several adaptive navigation systems have been proposed [Brusilovsky & al. 1994, 1996, Bodner 1999]. While they are goal specific (learning, tutoring, reading, etc.), NaviTexte [Couto 2006] is a generic text navigation system implementing our approach. This means that, depending on texts and knowledge modules, NaviTexte may be used, for example, as a learning, tutoring or reading system. Another important difference is that NaviTexte gives the user the liberty to navigate through the text following its own interests (the system propose - the reader chooses), while the mentioned systems try to maintain a user stuck to a given route (the user chooses - the system propose).

NaviTexte consists of sub-systems dealing with: text representation, navigational knowledge, visual representation and user interaction. The first one builds a text representation in memory from a text annotated manually or by dedicated software Cunningham & al. 2002, Bilhaut & al. 2003]. The second sub-system loads and compiles the knowledge modules. The result of this compilation is a graph of potential navigation courses that in practice is calculated as needed and stored in

optimization data structures. The third sub-system calculates and displays different text views and the fourth one manages the user interaction.

5. NaviLire, an Application of NaviTexte

Building an application with NaviTexte requires a set of texts and navigational knowledge modules. Both text representation and Sextant language have XML implementations with dedicated editors to use in case of a manual text annotation and a human knowledge encoder, respectively (*cf.* Fig.1). So far four applications have been developed: alternative automatic summarization [Couto & al. 2007], the NaviLire project [Lundquist & al. 2006], re-reading Madame Bovary [Mathieu 2005] and temporality in texts [Battistelli & al. 2007]. We present NaviLire to illustrate NaviTexte potential.

In the reading process, a reader has to deal with two basics types of cognitive problems. First, identifying discursive referents in a text and choosing correct relations between noun phrases that refer to them. In other words, the reader has to decide between a co-reference relation, in which there is only one referent, and a referential disjunction, in which there are several referents. This cognitive competence is crucial for the building up of a coherent mental representation of the text and hence central in the learning process: "learning from text requires that the learner constructs a coherent mental representation of the text" [Kintsch 2003:307]. Second, identifying the function and orientation intended by the sender. This orientation is generally marked from the beginning of the text and consequently acts as an "interpretation program" [Lundquist 1980]. Identifying this orientation, which is essentially provided by the predications, is also crucial for a correct deciphering of the semantic and pragmatic coherence.

As a general rule, textual coherence can be divided into three types, viz. the referential, the predicative and the pragmatic coherence, [Lundquist 1980]. In actual text navigation, however, the specific units to be focused on depend on the characteristics of the text in question. Figure 4 shows an example of identification of "Thematic coherence" in a French text. The student must identify discursive marks and a navigation operation allows him to check if he does not forget the preceding one.

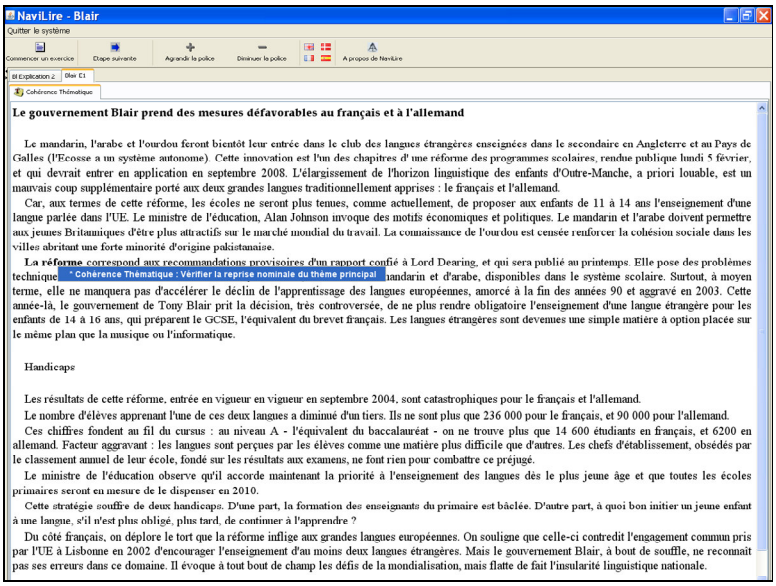

Figure 4. *Example of navigation operation in Navilire*

6. Conclusions and future works

We have presented our approach to text navigation conceived as a cognitive process that exploits linguistic information present in texts. We have defined it and explained the main differences with the classic hypertext navigation approach. The four elements needed to implement our approach are described: a text representation, the navigation knowledge modeling language Sextant, the knowledge encoding agents (via applications) and the NaviTexte system. One application of NaviTexte, NaviLire has been presented, showing the versatility of our approach. The quantitative results of our experimentation with Danish students learning French confirm the improvement obtained by using text navigation [Lundquist & al. 2006].

A long term consequence of modeling navigational knowledge is the creation of knowledge bases exchangeable and reusable. Current collaborations are reusing the knowledge coming from the NaviLire project into others e-learning projects in different languages (English, Latin and Spanish).

We think that our approach may have a significant impact on the way text is being read when its amount or nature does not allow sequential reading the Web. We work at present into several improvements of our approach. Related to some future application navigation in corpora, we are extending our concepts to a multi-text scenario. Related to last works in Web Wise, we plan to couple our approach to Semantic Web approaches to exploit existing annotations. Moreover, to achieve scalability in some kind of applications, typically open text ones, the application of

machine learning techniques to the acquisition of navigational knowledge is being studied.

Acknowledgments

The project NaviTexte is funded with an Ecos-Sud (U05H01) grant.

7. References

Battistelli D., Minel J.-L., Schwerr S., « Représentation des expressions calendaires : vers une application à la lecture des biographies», *TAL 47/2*, 25 p, 2007..

Berge C., *Théorie des Graphes*, Paris, Dunod, 1958.

Bechhofer S., Yesilada Y. and Horan B., COHSE: Knowledge-Driven Hyperlinks *Semantic Web Challenge at the International Semantic Web Conference*,. 2006

Bilhaut F., « The Linguastream Platfom », *Proceedings of the 19th Spanish Society for Natural Language Processing Conference (SEPLN)*, Alcalá de Henares, Spain, p. 339-340, 2003.

Bodner R., Chignell M., « Dynamic hypertext: querying and linking », *ACM Computing Surveys,* vol 31, n° 4,p. 120-132, 1999.

Brusilovsky P., « Adaptive Hypermedia: An attempt to analyse and generalise », *UM'94 Fourth International Conference on User Modelling*, p. 80-91, 1994.

Brusilovsky P., « Methods and techniques of adaptive hypermedia », *User Modeling and User-Adapted Interaction,*vol 6*, 2-3,* p. 87-129, 1996.

Charolles M., « L'encadrement du discours - Univers, champs, domaines et espace », *Cahier de recherche linguistique, LANDISCO*,vol 6, Université Nancy 2, p. 1-73, 1997.

Couto J., Une plate-forme informatique de Navigation Textuelle : modélisation, architecture, réalisation et applications de NaviTexte. Thèse de doctorat, Université Paris-Sorbonne, 2006.

Couto J., Lundquist L., Minel J.-L., « Naviguer pour apprendre», *EIAH 2005*, Montpellier, p. 45-56, 2005.

Couto J., Minel J.-L, « SEXTANT, un langage de modélisation des connaissances pour la navigation textuel», *ISDD'06*, Caen, p. 80-90, 2006.

Cunningham H., Maynard D., Bontcheva K. & al., « GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications, *ACL'02*, ACM Press, Philadelphie, Pennsylvanie, p. 168-175, 2002.

Danielson D.R., « Web navigation and the behavorial effects of constantly visible maps », *Interacting with Computers*, 14, 2002, p. 601-618, 2002.

Domingue J. B., Dzbor M., Motta E. 2004. Magpie: Browsing and Navigating on the Semantic Web, In *Proceedings of the Conference on Intelligent User Interfaces*, Portugal, 2004.

Edwards D.M., Hardman L., « Lost in hyperspace: cognitive mapping and navigation in a hypertext environment», in R. McAleese (Ed.), *Hypertext: Theory and Practice*. Oxford, England: Intellect Books, p. 105-125, 1989.

Elm W.C., Woods D.D., « Getting lost: A Case Study in Interface Design. », *Proceeding of the Human Factors Society*, p. 927-931, 1985.

Kintsch W., *Comprehension. A Paradigm for Cognition*, Cambridge, Cambridge University Press, 1998/2003, 1998.

Lundquist L., *L'analyse textuelle. Méthode, Exercices*, Copenhagen, Nordisk Forlag, 1980.

Lundquist L., *Tekstkompetence på fremmedsprog*, Copenhagen, Forlaget samfundslitteratur, 2006.

Lundquist L., J.L. Minel, Couto J., « NaviLire, Teaching French by Navigating in Texts», *IPMU'2006*, Paris, 2006.

Mathe N., Chen J., « A User-Centered Approach to Adaptive Hypertext based on an Information Relevance Model », *4th International Conference on User Modeling (UM'94)*, Hyannis, MA,. p. 107-114, 1994

Mathieu, Y. Y., « Annotations of Emotions and Feelings in Texts », In Conference on *Affective Computing and intelligent Interaction (ACII2005)*, Beijing, Springer Lecture Notes in Computer Science, p. 350-357, 2005**.**

Minel J.-L., *Filtrage sémantique. Du résumé à la fouille de textes*. Paris Hermès, 2003.

Text Encoding Initiative, http://www.tei-c.org

Thompson S., Mann W., «Rhetorical structure theory, a framework for the analysis of texts», *IPRA Papers in Pragmatics*, p. 79-105, 1988.

Webber B., Knott A., Stone M., Joshi A., « Anaphora and Discourse Structure », *Computational Linguistics, vol 29, n°4*, p. 545-588, 2003.

Wolf G., Gibson A., « Representing Discourse Coherence: A Corpus-Based Study », *Computational Linguistics, vol. 31, n0. 2*, p. 249-288, 2005.

Zellweger P. T., Chang B. and Mackinlay J. D., Fluid links for informed and incremental link transitions.In Proceedings of the Ninth ACM Conference on Hypertext and Hypermedia : Links, Objects, Time and Space, *HYPERTEXT '98*. ACM, New York, NY, 50- 57, 1998.