

Interfaces dynamiques de fouilles textuelles : vers une plate-forme de navigation textuelle

Javier Couto, Jean-Luc Minel

*Laboratoire LALICC, UMR 8139, CNRS, Université Paris-Sorbonne
96, bd Raspail, 75006- Paris France*

Javier.Couto@paris4.sorbonne.fr

Jean-Luc.Minel@paris4.sorbonne.fr

Résumé :

Les résultats obtenus dans le cadre de différents travaux montrent qu'il est possible d'exploiter certaines structures discursives, qui organisent le texte, afin de guider la lecture ou plutôt de suggérer des parcours de lecture. C'est ce que nous avons dénommé *navigation textuelle* et son objectif consiste à substituer à la lecture statique d'un texte une multiplicité de parcours de lecture. Notre projet vise à fournir, par l'exploitation, d'une part, des traces discursives laissées par le scripteur et d'autre part, des annotations sémantiques et de toutes connaissances jugées pertinentes par le concepteur d'un *module de visualisation et navigation* les outils qui permettront à un lecteur de déployer ses propres stratégies de recherche d'informations et de ne pas subir la linéarité du texte. Nous avons développé un premier prototype informatique afin de tester les idées proposées.

MOTS-CLES : visualisation d'information, fouille textuelle, navigation textuelle.

1. Introduction

Le fait qu'un texte soit maintenant numérisé et qu'il soit présenté au lecteur sur un écran peut être considéré comme une nouvelle mutation qui place le lecteur devant de nouvelles possibilités qui restent à explorer : « *Le texte [...] offre en effet une richesse sémiotique particulière, qui fournit de multiples objets d'interprétation et de multiples pistes d'actions [...] les lecteurs n'ont pas la même démarche envers l'objet ni la même définition de cet objet, ils ne « voient » pas la même chose* »

(Souchier & al. 2003). Pourtant, force est de constater que peu de systèmes de traitement de l'information se sont intéressés à comment exploiter cette richesse sémiotique. Par exemple, les systèmes de questions-réponses (Lin & Katz 2003) ou les systèmes de résumé automatique (Kan & al. 1999, Mani 2001, Minel & Desclés 2000, Minel 2003), construisent, à partir du texte source, un fragment textuel puis l'affichent comme une simple chaîne de caractères. Une des explications tient sans doute au fait que, le modèle du texte sous-jacent reste imprégné des contraintes imposées par une conception marquée par la prégnance technologique qui assimile texte et texte imprimé (Vandendorpe 1999). Néanmoins, différents travaux, notamment (Hearst 1999, Jacquemin & al. 2002), ont proposé de développer des interfaces qui transforment les représentations visuelles, mais ces transformations ne s'appuient pas sur l'identification de marques linguistiques, plus précisément, sur certaines structures discursives, qui organisent le texte (Charolles 1997).

Les résultats obtenus dans le cadre de différents travaux (Porhiel 2001, Jackiewicz 2002, Couto & al. 2004) montrent qu'il est possible d'exploiter ces structures discursives afin de guider la lecture ou plutôt de suggérer des parcours de lecture. C'est ce que nous avons dénommé navigation textuelle (Minel 2002, Couto & Minel 2004). L'objectif de cette navigation consiste à substituer à la lecture statique d'un texte une multiplicité de parcours de lecture que l'on peut, en partie, comparer au cheminement déambulatoire dans les hyperdocuments proposé par (Géry 2002). Ces principes de navigation se distinguent ainsi de la navigation hypertextuelle. En effet, c'est sur les marques sémiotiques et linguistiques que vont s'appuyer les opérations de navigation. La navigation proposée n'est donc pas guidée par l'auteur du texte comme dans le cas de la navigation hypertexte où les hyperliens sont placés par cet auteur

Dans les premiers systèmes informatiques de visualisation sur écran, la ligne constituait la seule unité manipulable et le défilement séquentiel la seule opération de contrôle disponible. La notion de fenêtre a permis d'introduire le contrôle spatial en deux dimensions à l'aide d'objets spécialisés que sont les barres d'ascenseur (scrolling bar), renouant ainsi avec un support qui prévalait avant l'introduction du codex. Enfin, ces dernières années les logiciels de traitement textuel ont réintroduit le format page (Word, Acrobat Reader, etc.) alors que les navigateurs utilisés pour explorer le Web cherchent au contraire à introduire de nouvelles compositions spatiales qui allient fenêtres, bandeaux, tableaux, liste, etc.

Dans les logiciels de traitement textuel la page est un construit éphémère, résultat d'un calcul qui s'applique sur une structure composée d'unités, le caractère, le mot, le paragraphe, sur lesquelles l'utilisateur doit pouvoir appliquer des traitements graphiques (taille, casse, justification, etc.). Les langages de description de ces structures (de SGML à XMLschema) ont constamment cherché à séparer les descriptions structurelles (la forme abstraite du texte) des descriptions de présentation (la forme graphique du texte) La notion de page a permis, entre autre, l'introduction d'instruments de recherche d'information ou d'aide à lecture tel que la table des matières, les index, les renvoi, etc.. Les logiciels de traitement textuel offrent potentiellement des instruments beaucoup plus puissants puisqu'ils disposent, en arrière plan, de la représentation structurelle du texte. Notamment cette représentation structurelle peut être annotée par des résultats issus de traitements

linguistiques (repérage de syntagmes saillants, de structures discursives, de relations sémantiques, etc.). L'exploitation de cette structure annotée par des logiciels de présentation permet ainsi d'envisager de nouveaux modes de lecture sur les « écrits d'écran » (Souchier 1997), ce mode de lecture restant néanmoins confiné dans un espace à deux dimensions.

L'élaboration de cette représentation structurelle nécessite de s'appuyer sur un modèle, même incomplet, de représentation d'un texte. En conséquence, les questions essentielles, qui concernent autant les linguistes que les informaticiens sont : que décrire dans un texte ? Quelles organisations textuelles faut-il décrire ? Existe-t-il des unités textuelles élémentaires, et comment celles-ci sont elles organisées ?

2. Modèles et outils de représentation d'un texte

Notre représentation d'un texte s'appuie sur les propositions de (Crispino 2003) afin de modéliser la structure logique de la manière la plus générale possible en incluant les éléments textuels qui peuvent être le mieux exploités dans notre projet. Notre vision d'un texte est celle d'une hiérarchie simple d'éléments textuels. Bien qu'il soit fréquent de trouver des textes qui présentent une hiérarchie plus complexe, nous pensons que notre option fournit, d'une part, un cadre suffisant pour la formulation des connaissances linguistiques exprimables dans notre méthode et d'autre part, elle constitue un modèle de base, qui est facilement extensible pour traiter des phénomènes plus complexes.

Nous modélisons cette hiérarchie par une structure d'arbre dont la racine représente le texte complet. Les descendants directs de cette racine représentent les sections de premier niveau dans le texte. Les descendants des sections sont les sous-sections ou les paragraphes. Les descendants de ces paragraphes sont les nœuds qui modélisent les phrases et finalement, les descendants des phrases, les feuilles de l'arbre, sont les unités lexicales. Ainsi, le parcours, dans l'ordre, des feuilles de cette hiérarchie correspond à la lecture linéaire du texte.

Pour manipuler cette structure, un langage (Crispino 2003) permet aux linguistes de spécifier des connaissances linguistiques (Ben Hazez & al. 2000, 2001). Enfin, rappelons que c'est cette structure hiérarchique qui est décorée par l'étiqueteur sémantique de la plate-forme ContextO (Minel & al. 2001). Enfin, afin de pouvoir annoter des structures discursives complexes, l'extension du langage Langtext est en cours (Crispino & al. 2004).

3. Outils de navigation

Notre conception des outils de navigation, dans le cadre de la fouille de textes, repose sur un principe général concernant le nouvel objet textuel à produire : à partir de l'objet textuel source T , le résultat de l'étiquetage du texte est un nouvel objet textuel T_f qui intègre l'objet T et une représentation décorée de T appelée T_d . Par

conséquent, un déplacement, dans la lecture du texte T_d doit toujours être répercuté, par le système de navigation, dans T . A chacun de ces objets textuels va être associée une forme sémiotique S_d qui visualise sur l'écran l'objet T_d , et des opérations de navigation O_d qui vont permettre de se déplacer dans T_d . En résumé, l'objet textuel T_f est une fonction de ces quatre éléments .

$$T_f = f(T, T_d, S_d, O_d)$$

Dans le cadre du projet Regal¹ (Ferret & al. 2001, Couto & al. 2004), développé en collaboration avec le LATTICE, le LIMSI et le CEA, certains choix de visualisation ont pu être mis à l'épreuve. Les informations à exploiter sont le résultat d'un processus d'intégration entre deux approches, l'une fondée sur un calcul de cohésion lexicale et l'autre sur le repérage de marques linguistiques. L'écran de navigation (cf. figure 1) se compose des barres standards et de quatre fenêtres, A, B, C, D.

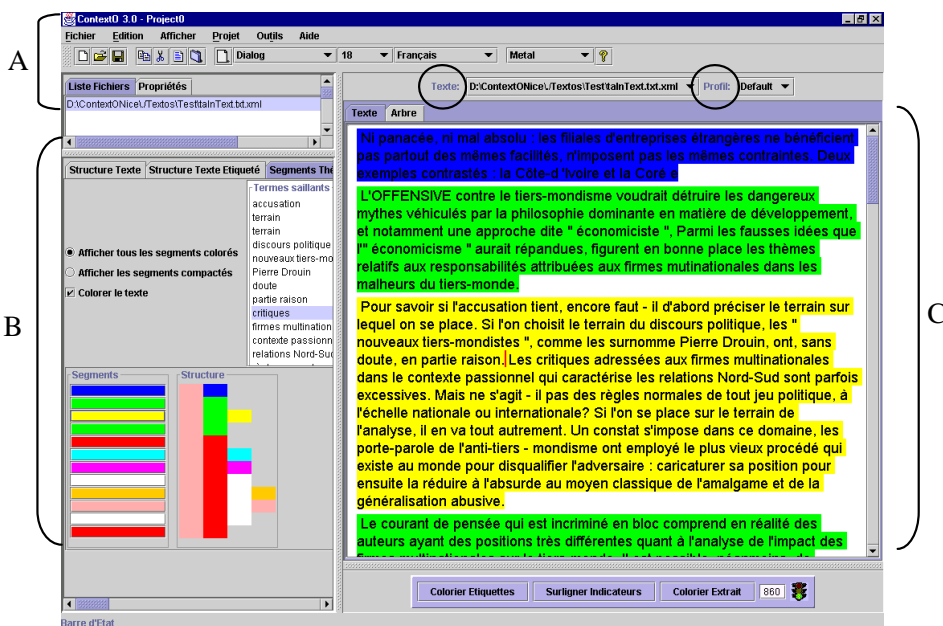


Figure 1 : Exemple d'interface dynamique spécialisée (extrait de Couto & al. 2004)

La fenêtre A affiche les différentes fonctionnalités proposées aux lecteurs. La fenêtre C affiche le texte source T dont certaines parties sont mises en relief par l'utilisation de colorisation. La fenêtre D, qui n'est pas visualisée sur la figure, affiche les différents objets T_f .

La fenêtre B offre différentes vues décorées $\{T_d, S_d\}$ du texte source $\{T\}$, accessibles par des onglets spécifiques. Ces décorations sont le résultat des calculs

¹ Le projet RÉGAL a reçu le soutien de l'ACI Cognitive (LAC038).

effectués (repérage des segments thématiques, fréquence des termes saillants, étiquette discursive attribuée à une phrase, etc.) et offre des points d'accès aux différentes parties du texte. Chaque segment thématique identifié est représenté sous la forme d'un rectangle coloré, les segments liés à un même thème étant de la même couleur. Sa taille en nombre de phrases est disponible ce qui permet à l'utilisateur d'évaluer l'importance de ce segment ou de ce thème par rapport aux autres thèmes du texte. La visualisation de la structure permet de mieux indiquer l'organisation des segments entre eux. L'inclusion d'un segment dans un autre signifie qu'il y a un lien thématique entre eux et permet de visualiser ainsi sa place par rapport aux autres thèmes. Pour chaque segment, il est possible d'afficher la liste des termes saillants avec leur fréquence. D'autres onglets permettent d'obtenir une vue hiérarchique du texte, organisée en arbre ou une vue globale des annotations attribuées aux phrases. Cette fenêtre offre donc différentes formes sémiotiques du texte : selon les thèmes, selon les marques discursives ou selon sa structure logique. Il est ainsi possible de prendre connaissance globalement du contenu du texte sans avoir à le lire et d'approfondir seulement certains points si des parties semblent intéressantes.

4. Modules de visualisation et navigation

Nous avons décrit précédemment des outils dynamiques développés dans le cadre du projet Regal, afin de gérer les interactions avec un lecteur. Si ces outils proposent certaines formes sémiotiques que l'on peut qualifier de génériques, par contre les opérations de parcours du texte sont plutôt spécifiques, c'est à dire liées à un type de structure discursive, comme, par exemple, les segments thématiques. Or, suivant le type d'application ou de l'intérêt du lecteur, les parcours de lecture peuvent être très différents.

Voici deux exemples qui illustreront notre propos. En s'inspirant des travaux de A. Jackiewicz (1998) sur le français, la phrase (1) peut être considérée comme saillante à cause de la présence de l'expression causale « *can play a very important role* », mais il est aussi important de signaler au lecteur que la phrase est sous la portée d'un cadre médiatif introduit par « *According to* ». Nous parlerons alors d'opération de visualisation déclenchée par la combinaison de l'opération de lecture, le lecteur a positionné le curseur sur cette phrase, soit en cliquant sur celle-ci, soit en la survolant. Une opération de visualisation peut ainsi être la conséquence d'une opération de navigation.

(1) « *According to usatoday.com, the jet stream contains often "jet streaks" of wind speeds faster than the surrounding regions. Apparently, as a consequence of these jet streaks, the jet stream can play a very important role in precipitation and storm formation.* »

Le deuxième exemple illustre le concept d'opération de navigation. Dans cet extrait, la première phrase introduit la notion de « *polar-front* » dont on considérera qu'il correspond au thème recherché par le lecteur. Le logiciel signale l'importance de cette phrase par une opération de visualisation, comme par exemple une colorisation,

et propose des opérations de navigation qui vont exploiter la structure discursive, un cadre organisationnel (Charolles 1997, Jackiewicz 2002), qui suit cette phrase. Ainsi ces opérations proposeront au lecteur de naviguer dans le texte en se positionnant, à sa convenance sur les phrases qui sont introduites par « First » ou « Second », etc.

(2) « *The polar-front jet is important for several reasons. **First**, because it is a region of maximum upper-tropospheric flow, it is also one of concentrated upper-air divergence and convergence. (...) **Second**, the polar-front jets, which move west to east but meander with the general upper-air waves, often "steer" the movement of major low-level air masses. (...) **Third**, jet streams are an important factor in high-altitude flight. Military and civilian jet aircraft depend heavily on reliable information about upper-air winds, which is used for planning the duration of flights and corresponding fuel consumption. **Furthermore**, jet streams generate a great deal of turbulence because of their strong wind shears ».*

(http://zebu.uoregon.edu/~js/glossary/jet_streams.html)

A partir de ce constat, nous cherchons à développer une plate-forme informatique (cf. figure 2) qui offre, pour un concepteur d'outils de navigation, des possibilités de déployer des connaissances navigationnelles. La seule contrainte est que les trajectoires de navigation doivent s'appuyer soit :

- sur l'étiquetage de segments textuels, les nœuds les feuilles de la représentation hiérarchique du texte ;
- sur l'étiquetage d'arcs entre les segments textuels qui appartiennent à cette représentation hiérarchique.

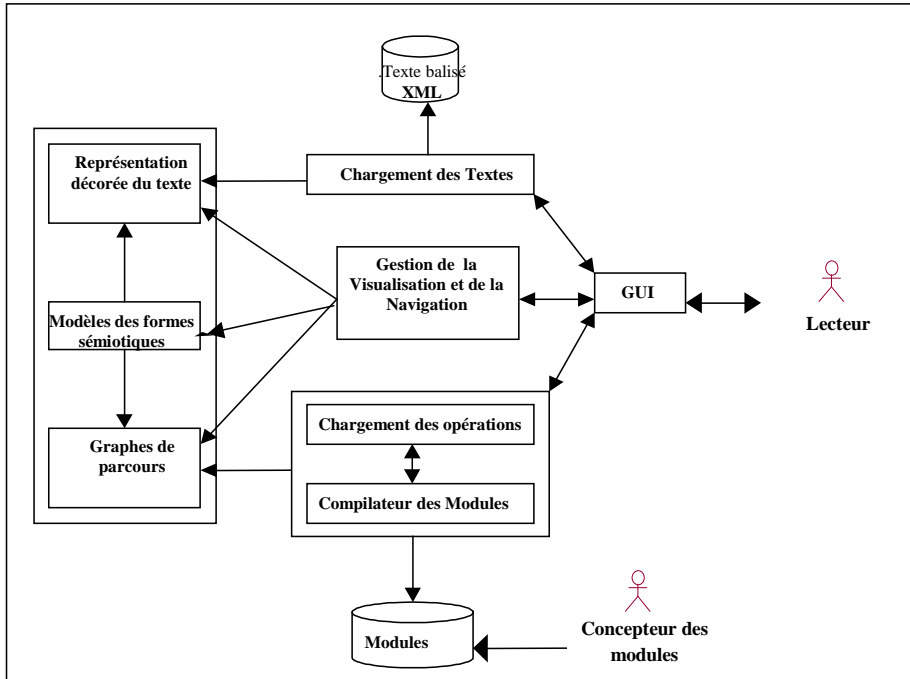


Figure 2: Architecture du prototype de la plate-forme

Un premier prototype développé en langage Java, dont l'architecture est illustrée par la figure 2, nous a permis de vérifier la validité de nos hypothèses. La plate-forme est ainsi composée de différents modules dont :

- un module qui se charge de construire à partir d'un texte balisé produit, par exemple par la plate-forme ContextO, la représentation décorée $\{T_d\}$ du texte ;
- un module de visualisation qui gère les interactions avec l'utilisateur en chargeant et en compilant à la demande une ou plusieurs cartouches de navigation. Le résultat de la compilation est un graphe de parcours qui est projeté sur la représentation décorée $\{T_d\}$. La visualisation effective est assurée par la sélection dans la base des modèles d'une ou de plusieurs formes sémiotiques $\{S_d\}$ spécifiées dans la cartouche.

Les connaissances navigationnelles sont encapsulées dans des composants logiciels appelés *modules de visualisation et navigation* (Couto 2003). Chaque *module M* identifie un ensemble de n descriptions de vue avec les paramètres et les contraintes de création, et les opérations de visualisation et de navigation possibles. Une vue V est spécifiée par une description de vue DV . La figure suivante montre la relation entre les textes étiquetés, les modules et les différentes vues.

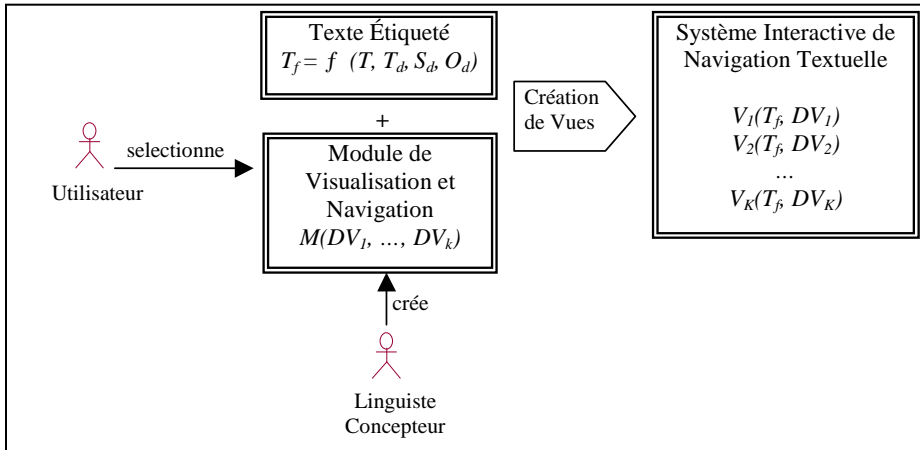


Figure 3. Principes de création de vues

Dans ce schéma, une *vue* est définie par sa *description de vue* en indiquant :

- sa représentation visuelle (texte plat, arborescente, etc.) ;
- ses paramètres (selon le type de représentation) ;
- ses contraintes (par exemple: conditions sur des unités textuelles et/ou des attributs) ;
- des opérations de navigation ;
- des opérations de visualisation.

Le contenu de ces modules est défini par un langage formel, fondé sur une syntaxe de type langage objet, qui permet d'exprimer les connaissances navigationnelles. Ces connaissances s'appuient sur le modèle du texte décrit dans un format XML et sur un ensemble de possibles opérations de visualisation et navigation.

4.1. Description des annotations textuelles

Le texte XML, selon le modèle que nous proposons, se divise en deux parties : la *tête*, où s'expriment les relations non hiérarchiques entre des éléments textuels et le *corps*, où les éléments textuels se structurent d'une façon hiérarchique. Dans le texte, chaque unité textuelle (UT) est balisée et son type est précisé, comme le montre l'exemple ci-dessous.


```

<UT Type="Paragraphe" Nro="7">
  <UT Type="Phrase" Nro="29">
    <Chaine> The polar-front jet is important for several reasons.
  </Chaine>
</UT>
  <UT Type="Phrase" Nro="30">
    <Chaine> First, because it is a region of maximum upper-tropospheric
      flow, it is also one of concentrated upper-air divergence and convergence.
    </Chaine>
  </UT>
  ...
  <UT Type="Phrase" Nro="37">
    <Attribut Nom="SemLabel"> Causalité </Attribut>
    <Chaine> Furthermore, jet streams generate a great deal of turbulence
      because of their strong wind shears. </Chaine>
  </UT>

```

Figure 4 : Exemple d'annotations du corps du texte

Dans cet exemple, les unités textuelles correspondant à une feuille dans l'arborescence textuelle possèdent un élément nommé *chaîne* qui représente la chaîne de caractères typographiques. De plus, chaque unité textuelle peut avoir un ou plusieurs attributs. Ces attributs sont typés et valués. Dans l'exemple ci-dessus, la phrase numéro 37 est annotée comme « causalité ».

Nous avons spécifié des mécanismes pour la création de nouveaux éléments textuels à partir des UT de base. Ces mécanismes consistent à exprimer des relations entre des éléments existants. Pour référer à une UT existante, nous répétons l'élément XML et cet élément est interprété comme un pointeur vers l'élément existant. C'est-à-dire que, dans la *tête*, l'élément XML :

```
<UT Type="Phrase" Nro="2" />
```

ne représente pas une phrase vide mais un pointeur à la phrase numéro 2, qui doit exister dans la partie *corps* du texte. Néanmoins, quelquefois il peut être utile de créer des relations mixtes, c'est-à-dire des relations qui se construisent à partir d'UT présentes dans le texte et de nouvelles UT. C'est le cas notamment des notes de bas de page, que nous représentons avec la relation *référence*. Dans ce cas nous avons une UT du texte se rapportant à une ou plusieurs UT qui ne font pas partie de la structure hiérarchique exprimé dans le *corps*.

Enfin, il est possible que certaines UT jouent un rôle particulier dans la relation exprimée. Nous formulons alors ces rôles sous la forme d'attributs spécifiques dont l'empan est limité à la relation elle-même. La plate-forme gère quatre types de relations : *l'ensemble*, *la séquence*, *la référence* et *le graphe*. Voici un exemple d'ensembles utilisés pour exprimer un cadre organisationnel (Charolles 1997, Jackiewicz 2002):

```

<Sequence Nom="Cadre Organisationnel">
  <Contenu>
    <Debut><UT Type="Phrase" Nro="29"/></Debut>
    <Fin><UT Type="Phrase" Nro="37"/></Fin>
  </Contenu>
  <Attributs>
    <UT Type="Phrase" Nro="29"> <Attribut Nom="Amorce" />
  </UT>

    <UT Type="Phrase" Nro="30">
      <Attribut Nom="MIL">1</Attribut> </UT>

    <UT Type="Phrase" Nro="32">
      <Attribut Nom="MIL">2</Attribut> </UT>

    <UT Type="Phrase" Nro="32">
      <Attribut Nom="MIL">3</Attribut> </UT>
  </Attributs>
</Sequence>

```

Figure 5 : Exemple d'annotations de la tête du texte

4.2. Opérations de visualisation

Rappelons (voir ci-dessus) que les opérations de visualisation et de navigation sont décrites dans les modules de navigation. L'opération de visualisation la plus importante gérée par la plate-forme est la *mise en relief*. Une opération de mise en relief est définie par :

- l'objet textuel à mettre en relief ;
- des opérations de colorisation ;
- des opérations textuelles.

L'exemple suivant (cf. figure 6) correspond à une mise en relief d'un cadre organisationnel comme celui décrit dans la figure 5.

```

<MISE_EN_RELIEF>
  <Objet> <Sequence Nom="Cadre Organisationnel" /> </Objet>
  <Coloriage>
    <Color Valeur="bleu">
      <UT Type="Phrase" /> <Attribut Nom="Amorce" />
    </Color>
    <Color Valeur="vert">
      <UT Type="Phrase" /> <Attribut Nom="MIL" /> </Condition>
    </Color>
  </Coloriage>
  <Text_Operations>
    <Ajouter>
      <UT Type="Phrase" /> <Attribut Nom="Amorce" />
      <Valeur>Amorce</Valeur>
    </Ajouter>
    <Ajouter>
      <UT Type="Phrase" /> <Attribut Nom="MIL" />
      <Valeur>MIL</Valeur>
    </Ajouter>
  </Text_Operations>
</MISE_EN_RELIEF>

```

Figure 6 : Exemple d'une opération de visualisation

Dans cet exemple, l'objet à mettre en relief est une séquence dont l'attribut *Nom* a la valeur « Cadre Organisationnel ». Il y a deux opérations de colorisation définies : la colorisation en bleu de la phrase contenant l'amorce du cadre et en vert de celles contenant les marqueurs d'intégration linéaire (MIL). Deux opérations textuelles sont définies : l'ajout du mot « Amorce » au début de la phrase contenant l'amorce et l'ajout du mot « MIL » au début de celles contenant les MIL. Par défaut, l'ajout de texte est fait entre parenthèses et en gras.

4.3. Opérations de navigation

Une opération de navigation est définie par différents éléments. Tout d'abord, par l'objet textuel sur laquelle elle s'applique. Puis par l'orientation, éventuellement indexée. Nous avons ainsi défini, *première*, *précédente[i]*, *suivante[i]* et *dernière*. Troisième élément, l'empan de texte (par exemple, le paragraphe, la section) dans lequel il faut rechercher l'objet. Et enfin, les conditions sur les attributs de l'objet, comme par exemple les valeurs des étiquettes.

```

<OP_NAVIGATION Description="Premier MIL" Type="Première">
  <Objet>
    <UT Type="Phrase" />
    <Attribut Nom="MIL" />
  </Objet>
  <Empan>
    <Sequence Nom="Cadre Organisationnel" />
  </Empan>
  <Conditions>
    <Condition>
      <UT Type="Phrase" />
      <Attribut Nom="MIL" />
    </Condition>
  </Conditions>
</OP_NAVIGATION>

```

Figure 7 : Exemple d'une opération de navigation

Dans l'exemple ci-dessus, l'utilisateur étant positionné sur une phrase qui contient un attribut de nom MIL, pourra aller, dans l'empan du cadre organisationnel, à la première phrase contenant un attribut de nom MIL. D'autres descriptions d'opérations permettraient de spécifier d'autres parcours de navigation dans un cadre organisationnel. Enfin, toutes ces *opérations de navigation* et les *opérations de visualisation* précédemment décrites, sont incorporées dans les *descriptions de vues des modules* et apparentent à la *vue* créée.

5. Conclusion

Notre projet vise à fournir, par l'exploitation, d'une part, des traces discursives laissées par le scripteur et d'autre part, des annotations sémantiques et de toutes connaissances jugées pertinentes par le concepteur d'un module de visualisation et navigation les outils qui permettront à un lecteur de déployer ses propres stratégies de recherche d'informations et de ne pas subir la linéarité du texte (Harris 1993).

Les connaissances exprimées dans ces modules doivent donc être considérées comme une aide à la lecture, voire une suggestion de lecture. Étant donné que la conception des modules s'appuie sur certaines informations déjà présentes dans les documents la qualité de la visualisation et de la navigation repose sur la compétence du linguiste mais aussi sur la richesse du langage de représentation.

Nous avons développé un premier prototype informatique afin de tester les idées proposées. Nous cherchons maintenant à expérimenter et à mesurer, pour des cas d'utilisations différents, l'intérêt de cette approche, notamment en termes de qualité de lecture. Ce constat rend le processus d'évaluation difficile car les critères

d'évaluation sont multiples. Deux aspects sont en effet à évaluer : le langage offert aux concepteurs et les connaissances des modules développés.

6. Références bibliographiques

- Ben Hazez, S., J.-P. Desclés, J.-L. Minel. (2001). Modèle d'exploration contextuelle pour l'analyse sémantique des textes, *TALN 2001*, Tours, p.73-82.
- Ben Hazez S., J.-L. Minel. (2000). Designing Tasks of Identification of Complex Patterns Used for Text Filtering, *RIAO'2000*, Paris, p. 1558-1567.
- Charolles M. (1997). L'encadrement du discours : univers, champs, domaines et espaces, *Cahier de Recherche Linguistique, LANDISCO*, Université Nancy 2, p. 1-73.
- Couto J. (2003). Développement d'une plate-forme de navigation textuelle, *Journée Doctorale de LaLICC*, Octobre 2003.
- Couto J., Minel J.-L., S (2004). (2004). Outils dynamiques de fouille textuelles, *RIAO 2004*, Avignon, p. 420-430.
- Couto J., Ferret O., Grau B., Hernandez N., Jackiewicz A., Minel J.-L., Porhiel S (2004). RÉGAL, un système pour la visualisation sélective de documents, *Revue d'Intelligence Artificielle*, 32 pages, à paraître.
- Crispino G. (2003). Une plate-forme informatique de l'Exploration Contextuelle : modélisation, architecture et réalisation (ContextO). Application au filtrage sémantique de textes. Thèse de Doctorat. Université Paris-Sorbonne.
- Crispino G, Jackiewicz, A., J.L. Minel. (2004). Spécification et implantation informatique d'un langage de description des structures discursives., *TALN 2004*, Fès, p. 131-136.
- Ferret O., Grau B., Minel J.-L., Porhiel S. (2001). Repérage de structures thématiques dans des textes, *TALN 2001*, Tours, p. 163-172.
- Gery M. (2002). Un modèle d'hyperdocument en contexte pour la recherche d'information structurée sur le Web. *Revue des Sciences et Technologies de l'Information*, 7/2002, Hermès, Paris, p. 11-44.
- Harris R. (1993). *La sémiologie de l'écriture*. CNRS Editions, Paris, 377 p.
- Hearst M. (1999). User Interfaces and Visualization. In *Modern Information Retrieval*. R. Baeta-Yates, B. Ribeiro-Neto (eds), Addison-Wesley, p. 257-322.
- Jackiewicz, A. L'expression de la causalité dans les textes. Contribution au filtrage sémantique par une méthode informatique d'exploration contextuelle. Thèse de Doctorat. Université Paris-Sorbonne, février 1998.
- Jackiewicz A. (2002). Repérage et délimitation des cadres organisationnels pour la segmentation automatique des textes, *CIFT'02*, pp. 95-107, Hammamet, Tunisie.
- Jacquemin C., Jardino M. (2002). *Multi-dimensional and Multi-scale Visualizer of Large XML Documents*, Proceedings of EUROGRAPHICS, Saarbrücken, Germany.
- Kan M.Y., Mckeown K.R. (1999). Information extraction and Summarization : Domain Independence through Focus Types. *Computer Science Technical Report, Columbia University*.
- Lin J., katz B. (2003). Question Answering Techniques for the Word Wide Web. *EACL 2003, Budapest*.
- Mani I. (2001). *Automatic Summarization*, John Benjamins Publishing Company, Amsterdam.
- Minel J.-L., J.-P. Desclés. (2000) *Résumé Automatique et Filtrage des textes*, in *Ingénierie des langues*, (sous la direction de J.-M. Pierrel) Paris, Editions Hermès, p. 253-270.

- Minel J.-L., E. Cartier, G. Crispino, J.-P. Desclés, S. Ben Hazez, A. Jackiewicz. (2001). Résumé automatique par filtrage sémantique d'informations dans des textes, Présentation de la plate-forme FilText, *Technique et Science Informatiques*, n° 3, Hermès, Paris, p. 369-396.
- Minel J.-L. (2002). *Filtrage sémantique de textes. Problèmes, conception et réalisation d'une plate-forme informatique*. Habilitation à diriger des recherches, Université Paris-Sorbonne.
- Minel J.-L. (2003). *Filtrage sémantique. Du résumé à la fouille de textes*. 200 p., Hermès, Paris.
- Porhiel S. (2001). Linguistic expressions as a tool to extract thematic information, *Corpus Linguistic 2001*, Lancaster University.
- Souchier E. (1997). Lire et Ecrire : *Editer des manuscrits aux écrans autour de l'oeuvre de Raymond Queneau*. Habilitation à diriger des recherches, Université Paris7 Denis Diderot.
- Souchier E., Jeanneret Y, Le Marec J. (ed). (2003). *Lire, écrire , récrire*. Etudes et Recherche de la Bibliothèque du Centre Pompidou.
- Vandendorpe, C.(1999). *Du papyrus à l'hypertexte*. 271 p., Editions la Découverte, Paris.