Chapitre 5

Analyses et interactions

Dans ce chapitre, nous présentons les analyses automatiques effectuées à partir des dispositifs LUCIA et les interactions qui en découlent. Nous débutons par une description des techniques de projection des informations lexicales sur des textes (5.1). Cette projection permet la mise en place d'interfaces de parcours rapide d'ensemble de documents et d'assistance à leur exploration (5.2). Ces interfaces de visualisation et d'interaction sont adaptées aux tâches. Nous avons été amené à étudier les aspects génériques et spécifiques de ces tâches pour mieux les élaborer. La projection des données des dispositifs sur des textes permet la mise en place de processus automatiques. Les interfaces et les calculs sont utilisées pour assister l'usager dans le cadre de l'étude de la métaphore et pour une veille documentaire. Ces travaux font l'objet de deux parties distinctes qui présentent les résultats obtenus et leurs perspectives (5.3 et 5.4). Finalement, dans une dernière partie, nous posons la question de l'évaluation de notre système centré sur l'utilisateur dans l'état actuel de ce travail (5.5).

5.1 Projections des informations lexicales 182

5.2 Visualisation et interaction 190
5.2.1 Techniques de visualisation interactive 190
5.2.2 Interactions génériques et spécifiques 196
5.2.3 Facteurs à prendre en considération 207

5.3 Étude de la métaphore 210
5.3.1 Première expérience 211
5.3.2 Observations et résultats 213
5.3.3 Conclusions et perspectives pour l'étude de la métaphore 223

5.4 Veille documentaire 226
5.4.1 LUCIASearch 227
5.4.2 Exemple d'utilisation 232
5.4.3 Conclusions et perspectives sur le projet de veille documentaire 247

5.5 Évaluation 248

5.1 Projections des informations lexicales

Dans cette partie, nous présentons le travail d'analyse automatique préparatoire aux différentes tâches. Les informations soumises par l'utilisateur sont projetées sur des documents. En l'absence d'analyse grammaticale ou syntaxique, la recherche des entités lexicales des dispositifs dans les textes relève d'une technique simple d'appariement qui peut amener certaines imprécisions ou ambiguïtés. Pour chaque entité lexicale repérée, les informations qui lui sont associées dans les structures LUCIA sont projetées sur le corpus par une annotation. Cette projection a pour but le repérage automatique ultérieur des occurrences et récurrences d'attributs et valeurs d'attributs. Ces phénomènes sont exploités différemment en fonction de la tâche pour laquelle est utilisée le système. Les détails de cette exploitation sont donc détaillés dans les parties 5.2 (relative à l'utilisation de ces données pour la mise en place des interfaces de lecture et de visualisation), 5.3 (relative à l'analyse de la métaphore) et 5.4 (relative à l'utilisation de LUCIA pour une tâche de veille documentaire).

Dans le chapitre précédent, nous avons vu que les données construites en interaction avec le logiciel LUCIABuilder étaient stockées dans différents fichiers au format XML. Il était donc logique, d'un point de vue technologique, de continuer d'exploiter la technologie JAVA et le langage XML pour explorer les textes inconnus. La phase préparatoire à l'exploitation des données sur corpus inconnu nécessite trois étapes (figure 77).

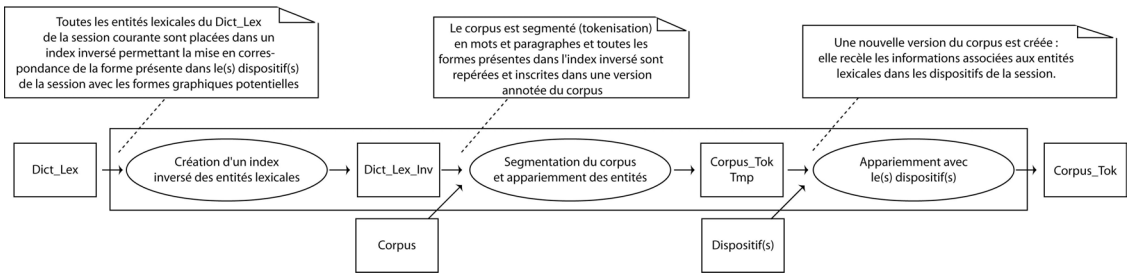


Figure 77 – Projection des données d'une session LUCIA sur un corpus

Trois étapes successives sont nécessaires pour obtenir, à partir d'un corpus inconnu, une version annotée de ce même corpus à l'aide des informations fournies par l'utilisateur.

La première étape consiste à créer un index inversé des entités lexicales et des formes possibles qui leurs sont associées dans le `Dict_Lex` de la session (figure 78).

Extrait du Dict Lex

```
<lex id="bdb5638">
      <lemme cat="N" >bourrasque</lemme>
      <flexion genre="F" nb="S">bourrasque</flexion>
      <flexion genre="F" nb="P">bourrasques</flexion>
</lex>
```

Extrait du Dict Lex inversé (Dict Lex Inv)

```
bourrasque;bdb5638;
bourrasques;bdb5638;
```

Figure 78 – Extraits d'un `Dict_Lex` **et** `Dict_Lex` **inversé correspondant.**

Dans la figure 78, nous voyons que le `Dict_Lex` inversé (`Dict_Lex_Inv`) n'est pas au format XML. En effet, comme le montre le schéma de la figure 77, il ne s'agit que d'un fichier temporaire pour l'annotation d'un corpus inconnu. Ce fichier n'est pas généré automatiquement à chaque exploration de corpus si les dispositifs de la session n'ont pas été modifiés. Il permet simplement d'accélérer l'étape suivante en ne parcourant pas le `Dict_Lex` dans son ensemble. La technique employée permet d'obtenir un `Dict_Lex_Inv` de taille inférieure au `Dict_Lex` qui lui correspond.

La seconde étape consiste à créer un corpus annoté à partir d'un corpus de textes inconnus et du `Dict_Lex_Inv` qui vient d'être créé. Cette étape nécessite une segmentation des textes du corpus effectuée en paragraphes et graphies (des détails seront donnés ultérieurement en fonction des tâches). Pour chaque texte du corpus, on crée en sortie une version XML du texte segmenté où les entités lexicales présentes dans le `Dict_Lex_Inv` sont repérées à l'aide d'une balise `lexApp` (pour entité lexicale appariée) avec la référence adéquate présente dans l'index inversé comme sur le modèle suivant (figure 79). Comme nous le verrons dans la partie (5.4) relative à la veille documentaire, des documents composites (pouvant présenter des images, des animations, etc.) peuvent être analysés en fonction de ces mêmes principes avec des manipulations supplémentaires.

Texte du corpus brut (Extrait de l'article 1 – corpus « Le Monde sur Cd-Rom »):

```
La bourrasque monétaire pourrait quand même brouiller une telle appréciation.
```

Document apparié :

```
<l>
<t>La</t>
<t>bourrasque<lexApp ref="bdb5638"/></t>
<t>monétaire<lexApp ref="bdm30036"/></t> <t>pourrait</t> <t>quand</t>
<t>même</t> <t>brouiller</t> <t>une</t> <t>telle</t>
<t>appréciation</t><t>.</t>
</l>
```

Figure 79 – Segmentation et appariement d'entités lexicales.

D'un point de vue technique, la segmentation s'effectue avec un *StreamTokenizer* – segmenteur sur flux de données – sur le modèle de celui utilisé dans MEMLABOR (chapitre 4 - partie 4.2.3).

Ce segmenteur minimal permet de distinguer les éléments séparés par des espaces, des apostrophes et des signes de ponctuations. Des exceptions peuvent être ajoutées au code comme pour par exemple *aujourd'hui*. Les exceptions gérées actuellement par le programme n'ont été testées que pour le français.

Si le `Dict_Lex_Inv` contient plusieurs entrées pour une même forme (par exemple des homographes), autant de balises de type `lexApp` sont placées dans le segment correspondant. Le corpus ainsi modifié est placé dans un fichier avec une extension `tok.tmp` pour « corpus temporaire tokenisé ». Les entités lexicales complexes ne sont pas repérées en tant que telles au cours de cette étape (figure 80).

Texte du corpus brut (Extrait de l'article 1 – corpus « Le Monde sur Cd-Rom »):

```
Quant à Paris, l'humeur était massacrante, en début de semaine, sous les co-
lonnes du palais Brongniart.
```

Document apparié :

```
<t>Quant</t> <t>à</t> <t>Paris</t><t>,</t> <t>l'</t><t>humeur</t>
<t>était</t> <t>massacrante</t><t>,</t> <t>en</t>
<t>début<lexApp ref="bdd13660"/></t>
<t>de<lexApp ref="bdd11881"/></t>
<t>semaine</t><t>,</t> <t>sous</t> <t>les</t> <t>colonnes</t>
<t>du<lexApp ref="bdd13323"/><lexApp ref="bdd13324"/></t>
<t>palais<lexApp ref="bdp33078"/></t>
<t>Brongniart<lexApp ref="neo6"/></t><t>.
```

Figure 80 - Segmentation et appariement d'entités lexicales parties d'entités complexes.

La figure 80 montre que les éléments d'entités lexicales complexes sont tous annotés en fonction des informations du `Dict_Lex_Inv` (*début*, *de*, *palais*, *Brongniart*). Les entités complexes des dispositifs (en l'occurrence *palais Brongniart*) ne sont pas repérées en tant que telles à ce stade de la projection des données des dispositifs.

La troisième étape consiste à projeter l'ensemble des données des dispositifs sur le corpus pour matérialiser la correspondance entre les graphies repérées et les informations qui leur ont été associées dans la session. Pour cela, document par document, chaque segment correspondant à une graphie appariée est agrémenté de balises supplémentaires `tabApp` (pour table appariée) comme sur le modèle suivant (figure 81).

```
<l>   <t>La</t>
<t>bourrasque
<lexApp ref="bdb5638"/>
<tabApp ref="disp_La_météo_tab11" vals=" attr11val0"/>
</t>
<t>monétaire
<lexApp ref="bdm30036"/>
<tabApp ref="disp_La_Bourse_tab3" vals=" attr3val0"/>
</t> <t>pourrait</t> <t>quand</t> <t>même</t> <t>brouiller</t> <t>une</t>
<t>telle</t> <t>appréciation</t><t>.</t> </l>
```

Figure 81 – Mise en relation des graphies du corpus avec les tables des dispositifs

Les identifiants utilisés au cours de cette annotation sont générés automatiquement par LUCIABuilder. Dans l'exemple proposé en figure 81, la session courante comporte deux dispositifs distincts (repérés par `disp_La_météo` et `disp_La_Bourse`). Il s'agit des dispositifs en rapport avec la bourse et l'économie d'une part, et la météorologie d'autre part, présentés respectivement p.159 et p.189.

Comme pour l'étape précédente, si une entité est présente dans plusieurs dispositifs de la session, autant de balises `tabApp` que nécessaires sont ajoutées à l'annotation. L'exemple montre une annotation simple, ce qui signifie que les attributs correspondants aux entités lexicales repérées sont uniquement ceux de la table dans laquelle ils apparaissent (attribut `vals` de la balise `tabApp`). Il s'agit des attributs de catégorie, spécifiques aux entités lexicales de la table (ou des tables) à laquelle ils correspondent. Dans le chapitre précédent, nous avons vu que la présentation en table du modèle n'était qu'une représentation possible et que de simples associations entités/attibuts-valeurs étaient identiques d'un point de vue computationnel. Comme nous allons le voir en particulier dans la partie 5.2, il est intéressant de distinguer les attributs participant à la caractérisation de la catégorie représentée par la table dans laquelle apparaît une entité (les attributs de catégorie) des attributs associés à cette entité par le truchement des liens de sous-catégorisation (les attributs hérités). Il est ainsi possible de procéder à une annotation analogue à celle proposée en figure 81 enrichie cette fois des attributs/valeurs hérités. La figure 82 montre les modifications induites par cet enrichissement.

```
<l>
   <t>La</t>
   <t>bourrasque
      <lexApp ref="bdb5638"/>
      <tabApp ref="disp_La_météo_tab11" vals="attr11val0" herit=""/>
   </t>
   <t>monétaire
      <lexApp ref="bdm30036"/>
      <tabApp ref="disp_La_Bourse_reg3" vals="attr3val0" herit="attr2val0
      attr1val1 "/>
   </t>
   <t>pourrait</t> <t>quand</t> <t>même</t> <t>brouiller</t> <t>une</t>
   <t>telle</t> <t>appréciation</t><t>.</t>
</l>
```

Figure 82 – Mise en relation des graphies du corpus avec les attributs de catégorie et les attributs hérités correspondants.

Texte du corpus brut (Extrait de l'article 557 – corpus « Le Monde sur Cd-Rom »):

(…) il y a deux ans, n'en menait pas large en début de séance.

Document apparié :

```
<t>il</t> <t>y</t> <t>a</t> <t>deux</t> <t>ans</t><t>,</t> <t>n'</t><t>en</t>
<t>menait</t> <t>pas</t> <t>large</t> <t>en</t>
<t>début<lexApp ref="bdd13660"/><tabApp ref="disp_La_Bourse_reg9"
vals=" attr9val1" clexPos="1" clexSize="3" herit="attr2val3 attr1val1 "/>
</t>
<t>de<lexApp ref="bdd11881"/><tabApp ref="disp_La_Bourse_reg9"
vals=" attr9val1" clexPos="2" clexSize="3" herit="attr2val3 attr1val1 "/>
</t>
<t>séance<lexApp ref="bds45019"/><tabApp ref="disp_La_Bourse_reg9"
vals=" attr9val1" clexPos="3" clexSize="3" herit="attr2val3 attr1val1 "/>
</t><t>.</t>
```

Figure 83 - Mise en relation des graphies du corpus avec les attributs de catégorie et les attributs hérités correspondants : cas des entités complexes.

La figure 83 montre la technique d'annotation employée pour les entités complexes, celles qui sont composées de plusieurs mots. À la balise `tabApp`, sont ajoutées deux balises `clexPos` et `clexSize` qui correspondent respectivement à la position de l'entité en question dans l'entité complexe et au nombre total d'entités composant l'entité complexe. L'heuristique considérée est celle qui tend à favoriser les chaînes syntagmatiques les plus longues, celles qui correspondent à des entités lexicales complexes, même si elles sont composées au moins d'une entité simple présente dans les dispositifs de la session. Ainsi, *séance* et *début de séance* ont été distinguées dans le dispositif en rapport avec la bourse : la figure 83 montre que dans l'extrait annoté, c'est *début de séance* qui a été retenu.

Certaines ambiguïtés peuvent apparaître dès la constitution du `Dict_Lex_Inv`. Par exemple, des graphies telles que *analyses* sont associées aussi bien au pluriel du substantif *analyse* qu'à la forme de la deuxième personne du présent du verbe *analyser*. Souvent, ces cas concernent plusieurs entités lexicales présentes une même ligne d'une même table (nous avons déjà abordé le fait qu'il est possible de tendre vers une approche morphémique pour les lignes des tables p.73). Ils ne posent donc pas de problème lors de la projection des informations sur les textes et les phases de calcul. Pour d'autres cas, en l'absence d'analyse grammaticale, les deux formes sont annotées dans les textes et amènent donc à des rapprochements erronés par rapport aux informations des dispositifs. Ces erreurs peuvent être repérées par l'utilisation des interfaces de visualisation. Si elles sont sources de problèmes conséquents pour la tâche, on peut y remédier soit en modifiant la structuration des dispositifs de la session, soit en modifiant les informations contenues dans le `Dict_Lex` (nous avons vu à travers la présentation de LUCIABuilder que cette fonctionnalité était permise par le programme). Si la deuxième solution permet de réduire le nombre d'erreurs, elle ne les règle pas toutes. Par exemple, supprimer *analyses* des formes possibles du verbe *analyser* ne permet plus de faire correspondre ces deux formes. C'est le processus itératif d'utilisation du système qui permet d'évaluer qu'elle est la forme la plus redondante et donc la plus intéressante à prendre en considération.

Les ambiguïtés peuvent concerner deux entités lexicales décrites dans deux éléments LUCIA différents (ligne, table, dispositif). Il est possible par exemple de catégoriser *courbe* parmi les objets permettant l'analyse aussi bien de phénomènes boursiers que de phénomènes météorologiques. Dans ces cas limites, l'usager est devant les deux mêmes choix précédents : soit il supprime l'entité de l'un ou l'autre des dispositifs de la session, soit il fait perdurer l'ambiguïté si elle n'amène pas des erreurs suffisamment gênantes lors des analyses automatiques. C'est une nouvelle fois le processus itératif qui permet d'assister le choix de la meilleur solution pour ces problèmes. Nous avions lors de premières expériences associé *temps* à un dispositif en rapport avec « la météorologie et le temps qu'il fait ». L'analyse des premières projections à l'aide des outils de visualisation (c.f. 5.2) nous ont alors montré que la majorité des utilisations de temps avait trait au *temps qui passe* plutôt qu'au *temps qu'il fait*. Cette entité n'apparaissait donc pas pertinente pour le dispositif en question et a été supprimée. D'une manière générale, les choix qui relèvent de la *sémantique légère* amènent à rencontrer des ambiguïtés sémantiques et syntaxiques qui pourraient être levées avec des quantités de données très importantes et/ou des analyses spécifiques. Deux raisons nous ont amenés finalement à ne pas prendre en considération ces phénomènes : l'interaction avec l'utilisateur et le calcul sur la récurrence. L'interaction permet de montrer les erreurs du calcul à l'utilisateur qui peut choisir d'y remédier si celles-ci gênent véritablement le bon déroulement de la tâche. Au moins deux façons d'y remédier sont possibles : restructurer les données dans les dispositifs ou revoir les informations des `dict_lex`. Le calcul sur la récurrence permet de prendre principalement en considération les phénomènes redondants, les récurrences d'attributs et valeurs/attributs, les récurrences d'entités lexicales dans les textes. Si une en-

tité est très présente et a été associée à un domaine où elle n'apparaît finalement pas à sa place, elle peut en être ôtée. Si une entité est peu présente et que les rapprochements des processus automatiques sont en partie erronés, alors ces erreurs peuvent perdurer et être considérée comme telles au cas par cas lors de l'utilisation.

Une fois les textes annotés avec les données issues des dispositifs, ces informations sont exploitées différemment selon la tâche en cours. Dans la partie suivante, nous verrons qu'elles sont utilisées pour la mise en place d'une interface de visualisation et d'interaction spécifique aux tâches pour lesquelles le système est employé.

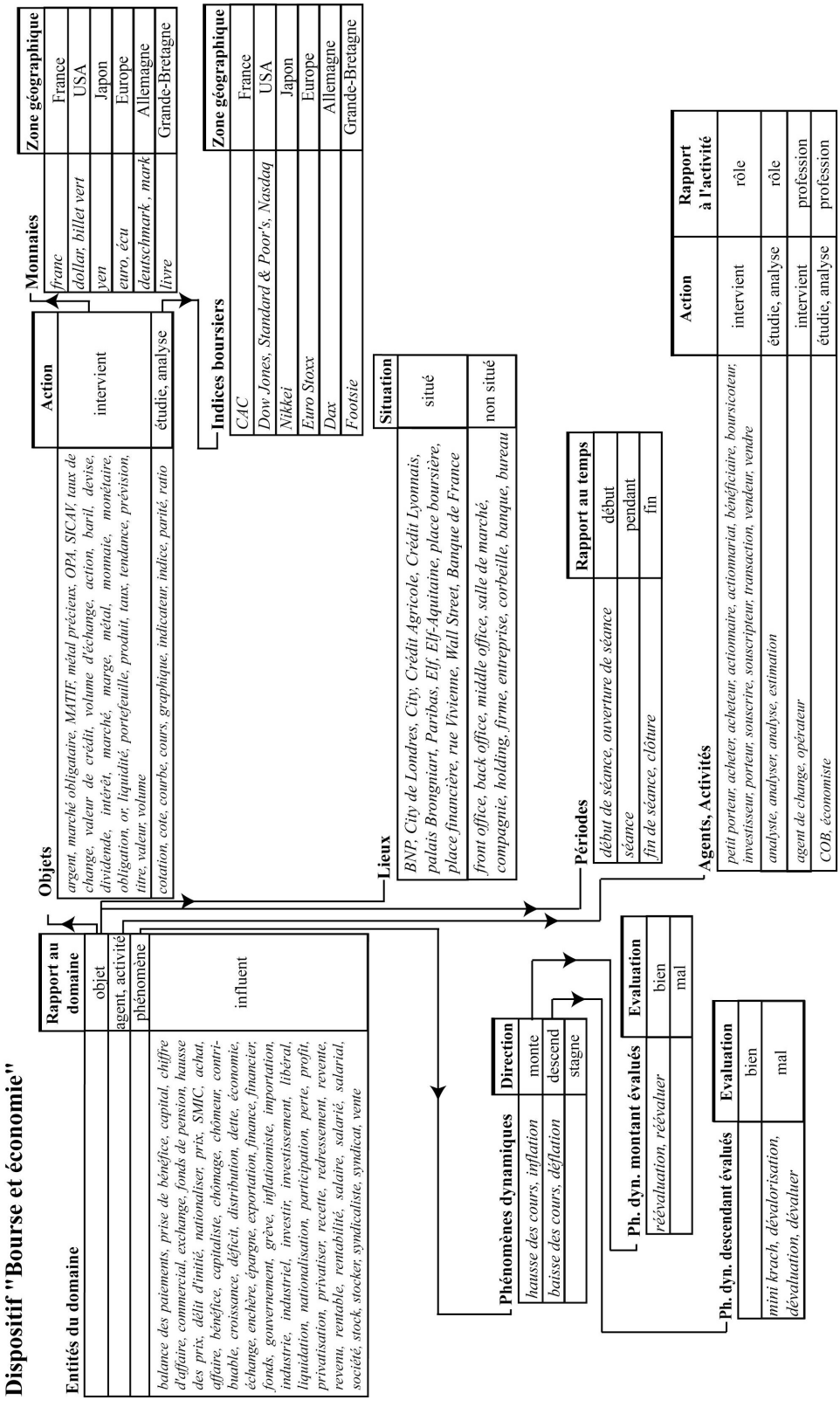


Figure 84 – Dispositif « Bourse et économie ».

5.2 Visualisation et interaction

Dans cette partie nous présentons les moyens utilisés pour présenter les résultats des analyses à l'utilisateur, lui permettre de les exploiter au mieux et assister l'évaluation de leur adéquation avec ses attentes par rapport à la tâche en cours. L'évaluation en question concerne également celle des ressources dans le cycle itératif d'utilisation du système.

Cette partie débute par un rapide état de l'art des techniques de visualisation et d'interaction pour des données textuelles (5.2.1). Nous proposons ensuite une présentation comparée des deux applications mettant en œuvre les traitements automatiques qui seront présentés spécifiquement dans les parties suivantes de ce chapitre. Il est important de noter que si les principes mis en jeu pour l'étude de la métaphore sont destinés à des experts du modèle, nos solutions quant à la veille documentaire peuvent être utilisées par un usager novice. C'est également le cas pour la plupart des autres tâches que le système peut assister (c.f. chapitre 4 - partie 4.1.4). Dans les deux cas, les applications doivent être en mesure de fournir des résultats facilement interprétables et rapidement exploitables. Une étude de ces contraintes communes permet de cerner les représentations visuelles et les interactions à mettre en œuvre dans les outils qui instrumentent LUCIA [Ferrari et Perlerin, 2004]. Nous devons ainsi préciser les besoins relatifs à la navigation dans une collection de documents et d'autres corrélés à la représentation à différentes échelles d'un même document pour un repérage rapide de parties intéressantes ou une analyse approfondie (5.2.2). Finalement, nous analysons le caractère générique de ces besoins et leur dépendance éventuelle vis-à-vis du modèle, de la tâche et de l'utilisateur pour mieux évaluer la généricité de nos propositions (5.2.3). Les travaux exposés dans cette partie ont fait l'objet de plusieurs publications [Perlerin et Ferrari, 2003*] et [Ferrari et Perlerin, 2004*] et ont donné lieu à plusieurs projets de DESS[117] dont [Taillepied, 2004] et [Hubert, 2003].

5.2.1 Techniques de visualisation interactive

La visualisation de documents numériques dépasse largement la problématique du TAL. Par exemple, une majorité des systèmes d'exploitation actuels proposent par défaut la métaphore d'un bureau avec une arborescence de dossiers et sous-dossiers pour accéder aux documents électroniques sur les machines. Ce type de présentation s'avère souvent difficile à manipuler pour des tâches nécessitant l'exploration de grandes quantités de données. Les modules de recherche sont souvent sollicités pour retrouver un document à partir du nom du fichier correspondant ou simplement en fonction d'une chaîne de caractères que l'on sait présente dans le document recherché (par exemple les fonctions `find` ou `locate` du système Linux). Ici les manques des biais de visualisation sont palliés par des

[117] DESS RADI, Réseaux, Applications Documentaires et Images, département d'informatique de l'université de Caen.

moyens d'interaction minimaux. Cependant de nombreuses autres solutions de visualisation et d'interaction avec des ensembles de documents existent.

Parmi les travaux les plus représentatifs de la visualisation et l'interaction avec des documents, on trouve le *Document Lens* [Mackinlay et Robertson, 1993]. Le système propose une vision en perspective déformante de documents textuels voisins et permet de situer un document dans un ensemble comme on peut le faire avec un support papier (à gauche sur la figure 85). Certaines solutions nécessitent en outre des appareillages complexes. C'est le cas par exemple du *Stereoscopic Field Analyser* [Ebert *et al.*, 1996] qui s'inspire de la réalité virtuelle. À l'aide de lunettes et de capteurs électroniques, on peut évoluer dans un espace documentaire représenté en trois dimensions (à droite sur la figure 85)

Figure 85 - Systèmes *Document Lens*, et *SFA*
d'après [Mackinlay et Robertson, 1993*] et [Ebert *et al.*, 1996*].

Lorsque les documents peuvent être placés dans des structures hiérarchiques (d'un thème général à un ensemble de sous-thèmes, d'un chapitre à ses paragraphes, d'un dossier à ses sous-dossiers, etc.), la structure arborescente de la hiérarchie peut être exploitée de multiples façons pour la visualisation de l'ensemble. Avec de telles configurations, on peut bien entendu n'envisager que la représentation en arbre à l'aide par exemple de simples *Java Swing JTree components.* On peut également élaborer des solutions plus complexes du type *Space Tree* [Grosjean *et al.*, 2002] où l'arbre représenté se déplie dynamiquement avec des fonctions de recherche et de filtrage (à gauche sur la figure 86). Les *Tree-Maps* [Johnson et Schneiderman, 1991] s'inspirent des techniques de représentations des graphes pour offrir la possibilité de transformer une arborescence sous la forme de rectangles contigus (à droite sur la figure 86). Le principe est d'affecter à chaque nœud de l'arbre correspondant une valeur numérique représentative de la taille ou de la complexité du sous-arbre enraciné à ce nœud. D'autres techniques du même type exploitent des représentations en trois dimensions avec des pyramides [Beaudoin *et al.*, 1996], des plans, des cubes, etc.

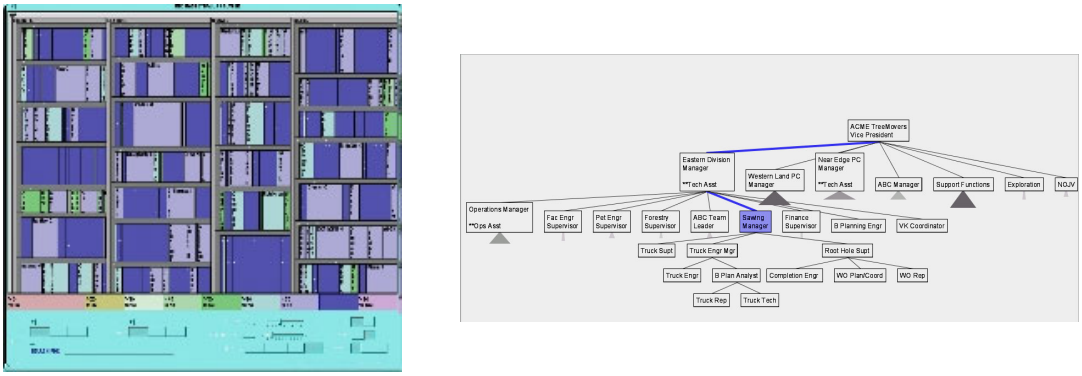


Figure 86 – Systèmes *Tree-Maps et Space Tree*
***d'après* [Johnson et Schneiderman, 1991*] et [Grosjean *et al.*, 2002*].**

Dans le cadre de la recherche documentaire, l'exemple le plus connu reste les *TileBars* de Hearst [Hearst, 1995]. À la suite d'une requête, un ensemble de rectangles correspondant chacun à un document jugé pertinent par le système est soumis à l'utilisateur. Dans ces rectangles quadrillés, chaque ligne correspond à un mot-clef de la requête et chaque colonne est grisée proportionnellement à la fréquence du mot-clef au sein du segment de document qui lui est associé (figure 87).

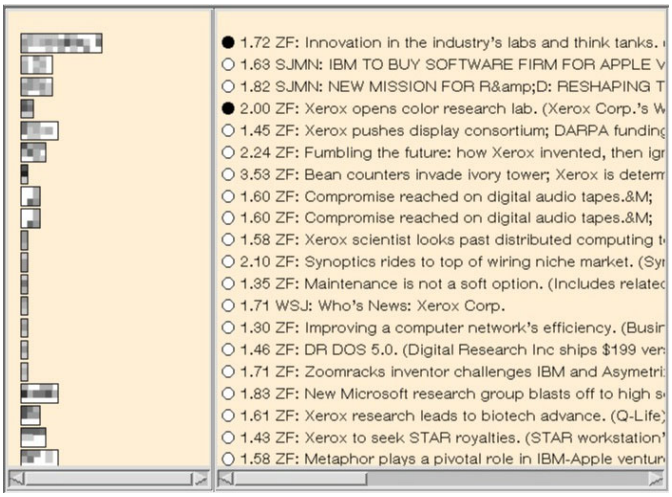


Figure 87 – *TileBars* d'après [Hearst, 1995*].

Le principe de la cartographie est utilisé pour représenter des ensembles de documents et mettre en évidence des proximités et des liens entre documents. Le métamoteur de recherche Kartoo[118] présente ainsi les résultats d'une requête sous la forme d'une carte reliant entre eux les sites en fonction des mots les plus fréquents qui leur sont communs (à droite figure 88). Un principe analogue est exploité par le métamoteur MapStan[119] (à gauche figure 88).

[118] http://www.kartoo.com

[119] http://search.mapstan.net/

Dans [Shneiderman, 1996], l'auteur décrit et compare plusieurs méthodes de visualisation de corpus de documents textuels à l'aide d'expérimentations avec des utilisateurs. Les conclusions de l'étude montrent que la représentation en deux dimensions nécessite des efforts moins importants que celle en trois dimensions mais l'absence de tâches véritables à évaluer limite les autres résultats de l'étude. D'une manière générale, l'évaluation de telles méthodes ne peut s'effectuer au final qu'en confrontant les applications à des utilisateurs comme dans [Cribbin et Chen, 2001] et [Cockburn et McKenzie, 2002]. Nous pensons également qu'une véritable tâche impliquant ces utilisateurs doit entrer en jeu comme l'a montré, dans un autre champ d'étude, l'expérience décrite dans le chapitre 3..

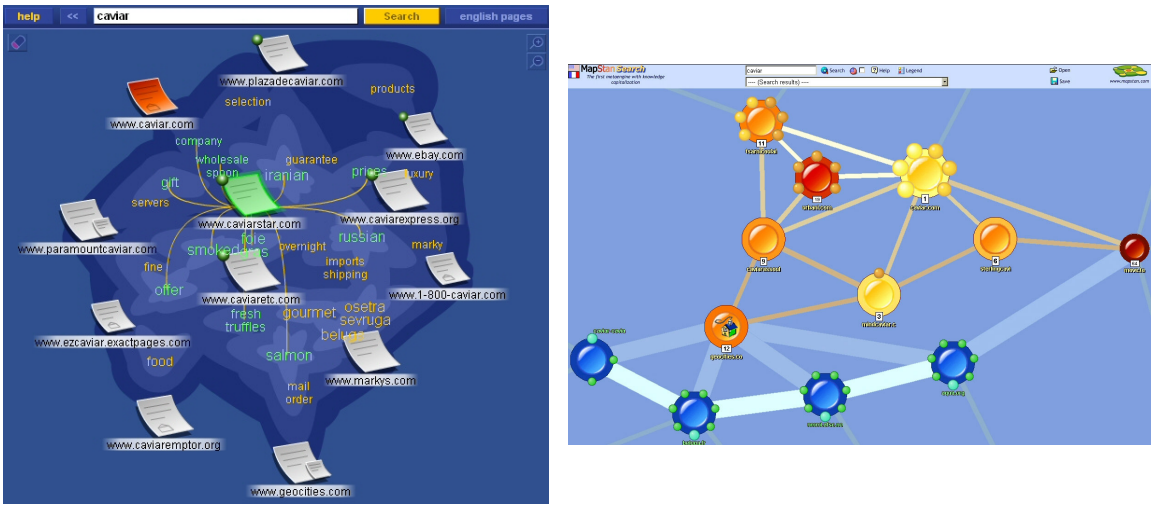


Figure 88 –Aperçu des résultats de *Kartoo* et *Mapstan* pour la requête « caviar ».

Au niveau du document ou du texte, les solutions sont nombreuses pour par exemple permettre la visualisation de relations de similarité entre parties ou fragments. Dans [Salton et *al.*, 1995*], on propose ainsi de projeter la représentation 2D d'un texte sur le périmètre d'un cercle et d'en relier les passages calculés comme similaires au moyen de segments (à gauche sur la figure 89). Jacquemin et Jardino [Jacquemin et Jardino, 2002*] s'inspirent de ces travaux avec 3D-XV un logiciel interactif de visualisation de documents volumineux encodés dans le format de DTD `docbook` (format XML) par une technique de projection et de coloration permettant la mise en valeur de passages relevant d'une thématique particulière (à droite sur la figure 89). L'interaction passe par le choix de détails de la représentation, des zooms et une présentation conjointe de la représentation en trois dimensions et du texte correspondant avec des parties coloriées. La coloration d'une partie du texte figure la présence majoritaire d'un thème donné.

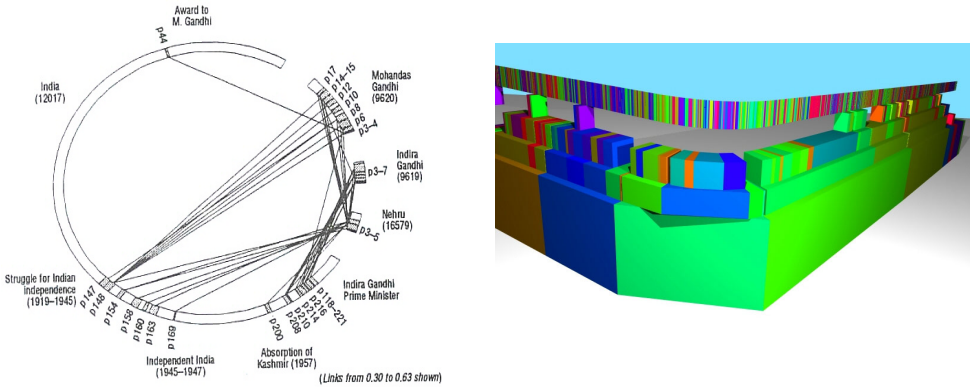


Figure 89 – Représentation de similarité entre paragraphes d'un texte d'après [Salton *et al.*, 1995] et 3D-XV d'après [Jacquemin et Jardino, 2002]

L'intérêt grandissant pour les travaux sur corpus a fait croître les besoins de techniques de visualisations spécifiques pour le TAL. Les nombreux logiciels dédiés à l'analyse de données textuelles, comme par exemple les logiciels HYPERBASE d'Etienne Brunet[120], LEXICO[3] de l'équipe CLA2T de l'université Paris III[121] ou encore LEXICA de la société Le Sphinx[122] proposent ainsi tous des biais de visualisation de résultats d'analyses sur corpus dépassant les simples listes textuelles. On trouve dans ces programmes informatiques des projections en Analyse de Composantes Principales (ACP), des courbes, des graphiques en secteurs, etc. Certains de ces logiciels utilisent également la couleur pour mettre en évidence des termes dans les documents pour faciliter l'étude de leurs distributions (c'est le cas par exemple de TROPES de la société Acetic[123] ou ALCESTE de la société Image[124]). Beaucoup de ces systèmes permettent une interaction avec l'utilisateur pour par exemple changer des paramètres de configuration comme les couleurs, ou un déplacement de *focus* sur les données visualisées (navigation, zoom, …). Notons enfin que LUCIA est à l'origine du projet ProxyDocs de Roy [Roy, 2004] qui utilise une représentation cartographique d'ensemble de documents classés thématiquement et dont les thèmes sont décrits par les techniques de LUCIA.

Les travaux en visualisation sont basés sur l'imagination des concepteurs et sur des considérations relevant de l'ergonomie, de la psychologie cognitive. Les solutions proposées ont longtemps été très gourmandes en ressources logicielles et très onéreuses en temps et coût humain. L'apparition et la démocratisation de techniques légères comme le dessin et les animations vectorielles (en Flash ou avec

[120] http://ancilla.unice.fr/~brunet/pub/hyperbase.html

[121] http://www.cavi.univ-paris3.fr/ilpga/ilpga/tal/lexicowww/lexico3.htm

[122] http://www.lesphinx-developpement.fr/

[123] http://www.acetic.fr/tropesfr.htm

[124] http://www.image.cict.fr/Index_Alceste.htm

SVG), les fonctionnalités graphiques de langage comme Java ou les transformations XML à partir de feuilles XLST ont considérablement amélioré cet état de fait.

Les techniques de visualisation d'ensembles documentaires sont un sujet d'intérêt de nombreux chercheurs et firmes productrices de logiciels car la représentation des données est un facteur essentiel du point de vue de la personnalisation des services. Dans une perspective humaine, la présentation des données et les interactions proposées affectent l'utilisation d'une solution logicielle. Il est maintenant admis qu'une véritable personnalisation implique que l'usager puisse non seulement adapter ses données (ce qui est permis par l'utilisation de LUCIABuilder) mais aussi avoir accès à des biais de visualisation et d'interaction spécifiques avec les résultats automatiques [Vassiliou *et al.*, 2002]. La spécificité des représentations doit donc être préalablement étudiée relativement à l'utilisateur et à la tâche. Nous verrons dans les parties suivantes que dans un cadre de TAL, la dépendance des représentations doit être également étudiée relativement au modèle utilisé et au type de matériau manipulé.

La visualisation, en tant que représentation personnalisée des données implique à parts égales les facteurs humains et technologiques [Pednault, 2000]. Du point de vue technologique, des structures de données adéquates sont nécessaires. Les moyens techniques doivent correspondre au type de public concerné ; il semble, par exemple, illusoire de vouloir proposer un appareillage de type lunettes 3D pour visualiser les résultats d'un moteur de recherche de l'Internet. Du point de vue humain, le système doit avoir été conçu relativement aux souhaits qui régissent la tâche courante. Dans le domaine des services commerciaux de l'Internet, Pednault [*ibid.*] propose un ensemble de recommandations pour qu'une représentation personnalisée de données soit efficace. Nous les adaptons ici à notre problématique :

- Il faut être simple et flexible. Il faut représenter seulement ce qui est nécessaire à l'usager pour résoudre ou assister sa tâche. Il faut éviter les présentations trop restrictives ou trop précises qui peuvent être caduques ou inutilisables. Nous voyons ici poindre la notion de palier (ou échelle) de présentation.
- La représentation doit être riche et fluide. La fluidité fait ici référence à la valeur ajoutée apportée par l'interface. La richesse de la représentation est corrélée à l'interprétabilité des informations qui y sont présentées. Le rapport entre les données fournies et celles qui apparaissent dans les résultats doit être facilement mis en relation par l'utilisateur. Il faut permettre une compréhension optimale des raisonnements effectués à partir des données (nous parlons ici des « rapprochements » effectués par nos logiciels).

Dans les parties suivantes, nous présentons les solutions logicielles que nous avons élaborées en fonction de l'étude des facteurs à prendre en considération pour chacune des applications (étude d'un fait de langue et veille documentaire) et des tâches à satisfaire (visualisation d'ensemble de documents, assistance à l'exploration d'un document, etc.).

5.2.2 Interactions génériques et spécifiques

Nous avons vu en détail dans le chapitre précédent comment pouvait être utilisé LUCIABuilder dans deux cas : l'étude d'un fait de langue et une veille documentaire. Ces deux cadres d'utilisation s'adressent à des types d'utilisateurs distincts. Ils présentent cependant une tâche centrale commune : la constitution des ressources. Pour présenter les résultats d'analyse en fonction des spécificités des tâches, nous proposons de préciser quelles sont celles qui sont génériques aux deux applications.

Dans le cadre d'une veille documentaire, l'efficacité du système ne se limite pas à la qualité des résultats automatiques. La qualité de l'interaction entre l'utilisateur et ces résultats y contribue en grande partie puisque c'est l'usager qui a au final le dernier mot. Les besoins d'interaction sont donc centrés autour des deux situations suivantes :

- la navigation dans une collection de documents analysés pour un repérage rapide des documents les plus pertinents (ou les plus susceptibles de l'être) ;
- l'identification et la lecture assistée des zones de texte pertinentes dans ces documents.

Dans le cadre du projet d'étude de la métaphore conceptuelle, destiné à des spécialistes de la langue et du modèle, les besoins d'interactions sont également centrés autour de la double situation précédente :

- la navigation dans la collection d'articles pour repérer ceux pouvant contenir des emplois de la métaphore étudiée ;
- l'identification et l'analyse détaillées des zones de texte pertinentes dans les articles.

En faisant le parallèle entre ces deux applications, nous montrerons dans la suite de cette section les principes communs et les spécificités de ces deux phases d'interaction. La spécification des aspects génériques aux deux tâches nous permet de dégager deux paliers de présentation : celui de l'ensemble de documents et celui du document lui-même. Pour le premier, nous parlerons de vue macroscopique et pour le second de vue microscopique.

5.2.2.1 Parcours du corpus : vue macroscopique

Dans les deux applications, une collection de documents existe, et l'utilisateur doit naviguer dans cette collection pour y repérer les documents qui l'intéressent le plus. Cet aspect apporte donc une dimension générique aux interfaces proposées. La présentation de tous les documents est une fonctionnalité générique, mais dans cette collection, la représentation de chaque document doit différer selon l'application car les critères de sélection d'un document au sein de l'ensemble ne sont pas identiques.

Pour l'étude sur la métaphore

Pour l'étude de la métaphore conventionnelle *météorologie boursière,* deux dispositifs sont construits : l'un correspondant à la bourse, l'autre à la météo. Chacune des entités lexicales de ces dispositifs est repérée automatiquement au sein des documents du corpus. Pour étudier le phénomène, la première tâche est de naviguer dans le corpus et d'y repérer les documents susceptibles de receler un emploi métaphorique attendu, c'est-à-dire des documents où au moins une entité lexicale en rapport avec la météorologie apparaît. Pour faciliter ce repérage, nous utilisons une interface regroupant l'ensemble des documents traités par les modules de projection des informations de la session (figure 90). Chaque document y est représenté sous la forme d'un graphique en histogrammes. Chaque barre d'un histogramme correspond à une table d'un dispositif. Elle hérite de la couleur associée à cette table dans LUCIABUILDER. Sa hauteur est proportionnelle au nombre d'entités lexicales du document qui apparaissent dans cette table. Un lien hypertexte permet enfin d'accéder au document représenté par le graphique. Lorsque des barres de la couleur dominante du domaine source sont repérées, une observation poussée de l'histogramme permet d'évaluer plus finement le lexique employé dans le document : le passage de la souris sur l'une des barres déclenche l'affichage du nom de la table associée et du nombre d'entités lexicales décrites dans cette table et repérées dans le document (figure 90). Le diagramme contient aussi un rappel des noms des dispositifs, le nombre de lexies trouvées pour chacun d'entre eux, ainsi qu'un lien vers le document analysé.

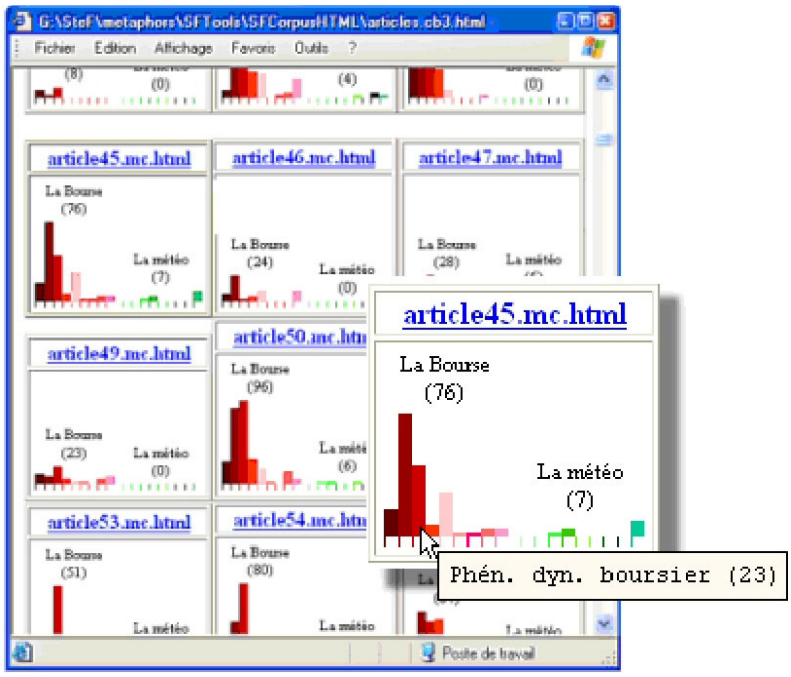


Figure 90 - Visualisation du corpus analysé pour le projet IsoMeta

L'ensemble des histogrammes est représenté au sein d'une même page HTML qui offre ainsi une vision d'ensemble du corpus traité, une vue macroscopique sur les documents du corpus. Cette

présentation est très simpliste : les documents sont ordonnés selon leur place dans la collection initiale. On peut cependant ne faire apparaître que les documents qui présentent au moins une entité lexicale présente dans le dispositif en rapport avec la météorologie pour faciliter le repérage à ce stade de l'étude. Un mécanisme de filtrage et d'ordonnancement est également proposé en fonction des propositions de la partie 5.4. Ce qui est véritablement intéressant ici est la manière dont la représentation d'un document dans cette collection est adaptée à l'utilisateur et à la tâche. Les informations présentes dans les histogrammes sont très précises au regard du modèle de structuration lexicale utilisé. Elles nécessitent une bonne appréhension de ce modèle de la part de l'utilisateur. Le choix de couleurs dominantes opposées pour les deux domaines facilitent le repérage rapide des articles intéressants (la météorologie apparaît en nuances de vert et la bourse en nuances de rouge dans la figure 90). Un tel choix n'est pas inhérent au modèle, il est conseillé lorsqu'il y a plusieurs dispositifs. De façon identique, il est préconisé d'utiliser le dégradé d'une même couleur en présence d'un seul dispositif. Notons enfin que la taille des barres n'est pas relativisée car la collection est constituée de textes de tailles comparables. Cet aspect introduit une dépendance de nos représentations au matériau traité, indépendante du modèle et de l'utilisateur.

Cette interface de visualisation s'appuie sur les langages HTML et CSS et les feuilles de transformation XSLT à partir du corpus au format XML. Elle fonctionne avec IExplorer5.0 et JDK1.4 (et les versions ultérieures de ces gratuiciels). Le format des documents manipulables par nos modules est le format TXT (ou HTML en utilisant la transformation HTMLtoTXT proposée dans MEMLABOR – c.f. 4.2.3).

Pour la veille documentaire

Dans le cadre de la veille documentaire, les analyses peuvent être effectuées par le module LUCIASearch pour filtrer et ordonnancer les résultats obtenus de moteurs de recherche (ce module est présenté en 5.4.1). Au niveau des représentations proposées et par rapport à la tâche précédente, des différences majeures apparaissent :

- l'intérêt n'est plus simplement de repérer des entités d'un dispositif dans un document mais de pouvoir rapidement en évaluer l'importance relative ;
- le corpus est beaucoup moins homogène en genre, en taille, etc. ;
- les descriptions des ressources peuvent être moins complexes que précédemment (l'utilisateur n'est plus nécessaire un expert de la langue ou du modèle).

Pour faciliter la navigation dans les listes de résultats, nous proposons en conséquence une représentation schématique du document qui conserve son aspect visuel global, et intègre une coloration des parties de texte correspondant aux thèmes attendus par l'utilisateur. Les parties colorées des zones de textes sont proportionnelles aux nombres de récurrences d'attributs que l'on y trouve. C'est la cou-

leur de la table majoritairement représentée comme support de récurrences qui est utilisée. En cas d'égalité, c'est la couleur de la table majoritaire où apparaît l'attribut mis en jeu dans la récurrence qui est représentée. À l'aide de ce schéma, nous proposons de visualiser de façon schématique la proportion des récurrences relativement aux parties des documents et non au document dans son ensemble - nous nous inspirons en cela des *TileBars* (c.f. p.192). Les récurrences sont ici considérées par rapport au thème auquel elles peuvent correspondre, par rapport aux attributs des tables des dispositifs. La représentation schématique a pour but de conserver l'aspect visuel du document, elle intègre les images et la mise en forme d'origine. L'aspect visuel global des documents est important à présenter car il permet le repérage rapide des invariants aspectuels d'un site ou d'un genre textuel dans un format numérique. Par exemple, un article de journal sur l'Internet est souvent présenté dans un document où apparaissent un menu de navigation (celui permettant de naviguer dans le numéro ou les archives du journal en question), des photos (pour illustrer l'article) voir des noms de rubriques, des publicités à des places définies (en haut et/ou à droite) etc. (figure 91)

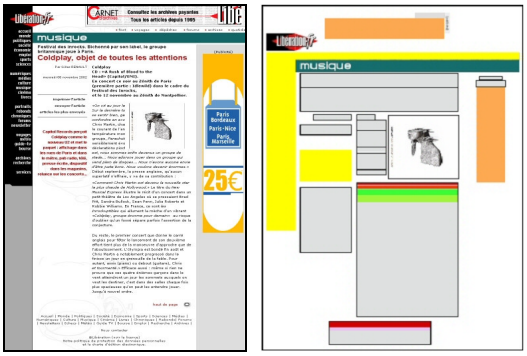


Figure 91 - Visualisation d'un document pour la veille documentaire:

à gauche un article du journal Libération (format HTML), à droite la représentation schématique SVG produite automatiquement avec coloriage de certaines parties du texte.

Les représentations SVG des documents sont obtenues d'un module en Java appelé HTMLtoSVG. Elle exploite le format SVG pour permettre certaines interactions intéressantes comme le zoom intégré au langage. L'utilisation de SVG nous permettra également par la suite d'intégrer des hyperliens vers les parties intéressantes des documents.

Après avoir corrigé les éventuelles erreurs du code HTML (avec JTidy c.f. note 93, p.131), les structures remarquables du document sont répertoriées et analysées. Il s'agit de repérer les zones de texte qui seront analysées et celles qui ne seront qu'affichées dans la représentation. Les zones qui ne seront pas analysées sont celles qui correspondent :

- aux images et aux animations (type Flash, Shokewave…) qui illustrent le document ;
- aux index ou tables des matières qui permettent la navigation intra-site ou vers d'autres sites ;
- aux bandeaux publicitaires.

Le repérage des zones de textes, des images et des animations s'effectue par l'analyse (*parsing*) du code HTML : la syntaxe des balises d'insertion de ces éléments est dûment définie pour ce langage. Pour les zones de navigation et les publicités, des techniques spécifiques ont été élaborées sur les principes suivants.

Les zones de navigation sont généralement placées à l'intérieur de tables qui contiennent majoritairement des hyperliens. Leur repérage relève d'un calcul sur le nombre de liens présents dans la table. Plusieurs méthodes ont été expérimentées :

- Si le nombre de liens est supérieur au nombre d'éléments qui ne présentent pas de liens, la table est considérée comme une zone de navigation. Cette méthode s'avère fiable mais la satisfaction que l'on obtient des résultats décroît fortement lorsque les zones analysées présentent par exemple plusieurs liens de petites tailles (un simple mot par exemple) et un seul grand paragraphe de texte dans la même zone (de l'ordre de plusieurs lignes).
- Si la zone analysée est un nœud ayant beaucoup de fils dans le *graphe web* incluant le document, alors elle correspond à une zone de navigation. Cette technique est fiable mais elle suppose la mise en place de la création du *graphe web* du document et donc du site où il apparaît. Les calculs nécessaires sont donc très importants.
- Si le nombre de liens de la table par rapport au nombre total d'éléments qui y sont présents est supérieur à un certain seuil, la table est considérée comme une zone de navigation. L'inconvénient de cette méthode est bien entendu la détermination de la valeur du seuil. Cependant cette solution est simple et s'avère fiable pour beaucoup des sites analysés. C'est celle que nous avons mise en place pour la création des représentations des documents en SVG. Du point de vue technique, on parcourt le contenu de la zone à analyser et l'on calcule la grandeur cumulée en nombre de caractères des textes des hyperliens. Si les hyperliens sont des images, c'est la taille des images qui est calculée. Si le nombre de caractères ou la hauteur cumulée des images de la zone est supérieure à la moitié du nombre de caractères ou de la hauteur totale de la zone, alors cette zone est considérée comme une zone de navigation.

Dans les représentations SVG, les zones de navigation sont coloriées par défaut en jaune. La couleur est bien sûr modifiable pour ne pas entrer en conflit avec les couleurs utilisées dans les dispositifs.

Le repérage des zones publicitaires est délicat. Celles-ci peuvent être de tailles différentes, à des endroits épars et encodées de façons multiples. Elles peuvent correspondre à de simples images, à des *frames*, à des *scripts* d'animation, etc. Il existe déjà un certain nombre de solutions pour la détection de publicités dans des documents de l'Internet. Par exemple, le *BannerFilter*[125] d'Andy Lyttle

[125] http://freshmeat.net/projects/bannerfilter/

compare la taille de la zone analysée avec des tailles de zones qui correspondent couramment à des publicités (120x600, 350x80 pixels, etc.). Cette solution est approximative car les e-commerçants qui connaissent l'astuce proposent de plus en plus des zones publicitaires de tailles variables. De plus, cette méthode entraîne une perte possible d'images ou de parties de texte lorsque le site utilise des zones de ces tailles pour du texte. Une solution plus efficace consiste à analyser le texte des hyperliens, le texte des éventuels messages contextuels associés aux images de la zone (balise `alt`) et les adresses des hyperliens. Les zones publicitaires contiennent en effet souvent des mots-clefs du type *ads*, *pub*, *advertisement*, *banner*, etc. Elle permettent également souvent une navigation via des services de comptage de clics comme *doubleclick.net*, *makemoney.com*, *hyperbanner.net*, etc. Le repérage de ces chaînes de caractères amène à caractériser les publicités. Cette méthode est satisfaisante mais présente quelques désavantages. D'une part, la liste des mots clefs ou des services possibles doit être régulièrement mise à jour ; certaines zones peuvent parfois échapper à l'analyse. En outre, l'utilisation de simples mots-clefs présente toujours les mêmes problèmes : en l'absence de règles exhaustives ou d'une interprétation humaine, la présence d'un mot-clef n'est jamais un critère définitif pour un repérage quelconque. Ainsi, certains répertoires de l'Internet se nomment *pub* pour signifier qu'ils sont publiques et des liens vers des documents présents dans ces répertoires peuvent amener des erreurs de repérage. Dans les représentations SVG, les zones publicitaires sont coloriées par défaut en orange. Cette couleur est elle aussi modifiable par l'utilisateur.

Une fois les zones remarquables du document répertoriées, on peut procéder à l'analyse des zones de texte et procéder à la création des représentations ainsi que du rapport d'analyse correspondant. Les représentations SVG des documents d'un même ensemble peuvent être présentées conjointement à leur rapport d'exploration au sein d'un même document HTML comme sur le modèle de la figure 92. Les résultats rassemblés dans cette figure ont été obtenus de LUCIASearch (c.f. 5.4.1 p.227) par l'interrogation des moteurs de recherche Yahoo et Lycos à partir d'un dispositif en rapport avec la bourse. La page HTML présente une vision macroscopique de l'ensemble des documents pressentis comme pertinents par le logiciel en fonction des données fournies par l'utilisateur. Il s'agit d'un prototype : l'intégration de HTMLtoSVG à LUCIASearch n'a pas été totalement finalisée. Nous pouvons apprécier ici la valeur ajoutée qu'apporte la représentation schématique des documents coloriés en fonction du dispositif en question et du repérage des zones remarquables. Le premier document retenu présente principalement deux zones de textes où apparaissent des thèmes décrits dans le dispositif tandis que le second voit apparaître ces thèmes dans au moins trois parties de la zone textuelle principale ainsi qu'une zone à droite de la page. Cette dernière correspond en fait à une table HTML qui recèle le texte *Études Économiques et Financières* dans une taille de police de 6 dans le document d'origine. La taille de la police est corrélée avec la taille de la zone correspondante ce qui explique son importance dans la représentation. Il s'agit en fait d'une limite du module HTMLtoSVG. Cependant, cette limite peut être appréhendée en tant que telle par l'examen rapide du schéma puisque la zone en

question n'est pas directement intégrée au corps principal du document et il semble peut probable qu'une partie importante du texte présent dans le document soit placée dans une zone adjacente.

Dans la figure 92, les informations proposées sont relatives à la requête effectuée à partir de LUCIASearch (table « Infos recherche ») et aux résultats obtenus (table « Rapport d'exploration »). À chaque document jugé potentiellement intéressant, on a ajouté les informations obtenues du moteur consulté (extrait du document, taille pour Yahoo, URL) ainsi que celles relevant de l'exploration du texte par les modules d'analyse (ici en fonction des tables du dispositif utilisé).

Figure 92 – Insertion des représentations SVG dans une page de résultats de LUCIASearch.

5.2.2.2 Parcours des textes et parties de texte : vue microscopique

Une fois un document jugé potentiellement intéressant à partir des vues macroscopiques, il faut pouvoir parcourir efficacement ce document. Dans cette partie, nous changeons donc de palier de représentation.

Pour le projet ISOMETA, présenter le document pour en permettre la lecture n'est pas suffisant en soi, les emplois métaphoriques doivent être rapidement repérés, et donc mis en évidence au sein du document. Ce projet est destiné à des utilisateurs experts à la fois de la langue qui est leur objet d'étude et du modèle. Pour l'analyse d'un fait de langue comme la métaphore, les informations à leur proposer sont nombreuses et complexes, et les représentations visuelles que nous avons élaborées reflètent cette dimension. Pour la vision macroscopique du corpus, nous proposons, comme nous l'avons vu précédemment, une représentation en histogrammes qui a pour objectif de permettre de saisir en un regard la présence d'entités lexicales associées par l'utilisateur au domaine source de la métaphore étudiée. Lorsque l'utilisateur décide d'explorer un document pressenti comme intéressant depuis cette représentation, il a besoin d'y repérer les emplois métaphoriques potentiels. Nous exploitons la couleur pour le guider en surlignant les entités lexicales des dispositifs (figure 93). Un parcours rapide du document (à l'aide de la barre de navigation verticale du navigateur) permet de se rendre facilement vers les zones de textes où apparaissent les entités lexicales d'un dispositif donné (ce principe est le même que celui proposé dans THEMEEDITOR – c.f. 4.2.4 p.141). L'utilisateur peut donc repérer rapidement les unités lexicales intéressantes pour sa tâche et observer localement le phénomène étudié. Le passage de la souris sur les unités surlignées déclenche ici aussi l'affichage d'informations complémentaires pour aider à l'interprétation des résultats. Il s'agit du nom de la table (ou des tables) dans laquelle l'entité est présente ainsi que les attributs et valeurs d'attributs de catégorie correspondants. Dans l'exemple de la figure 93, l'entité *thermomètre* appartient à la table « Outils de mesure » et correspond dans cette catégorie à [Axe : température].

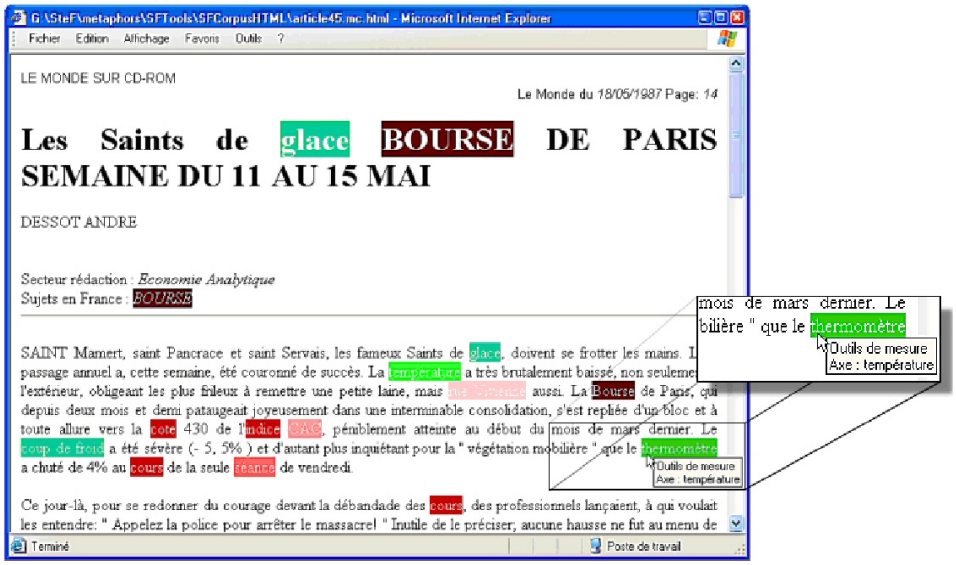


Figure 93 – Visualisation d'un document analysé dans le cadre du projet ISOMETA.

D'un point de vue technique, la représentation colorée est obtenue automatiquement depuis la version XML du document analysé. La page HTML de la vue microscopique est le résultat d'une transformation par une feuille XLST. Cette page est visible dans les principaux navigateurs pour l'Internet. Cette représentation s'avère également satisfaisante pour une tâche de veille documentaire où l'approche thématique est plus probante.

L'interface proposée pour le parcours des documents dans le cadre de l'étude la métaphore ne rend pas compte visuellement des récurrences d'attributs. Elle permet une approche thématique qui n'est pas suffisante pour intégrer toutes les informations utiles. En effet, si les informations que l'on peut obtenir d'un passage de la souris sur une partie de texte coloriée correspondent aux attributs de catégorie correspondant à l'entité lexicale, la couleur utilisée est celle associée à la catégorie. Dans le cas où il y a plusieurs attributs, ceux-ci n'apparaissent donc pas en tant que tels dans l'interface. De plus, si les attributs hérités sont bien associés aux entités par les analyses automatiques, ils ne sont pas visibles non plus.

Pour l'étude de la métaphore, une interface supplémentaire a donc été développée. Elle exploite un affichage en trois dimensions afin de pouvoir cumuler des informations visuelles sur les unités lexicales pertinentes. Il s'agit d'un programme nommé 3D- LUCIAVisualizer (figure 94) qui a été réalisé au cours d'un projet de DESS [Taillepied, 2004*].

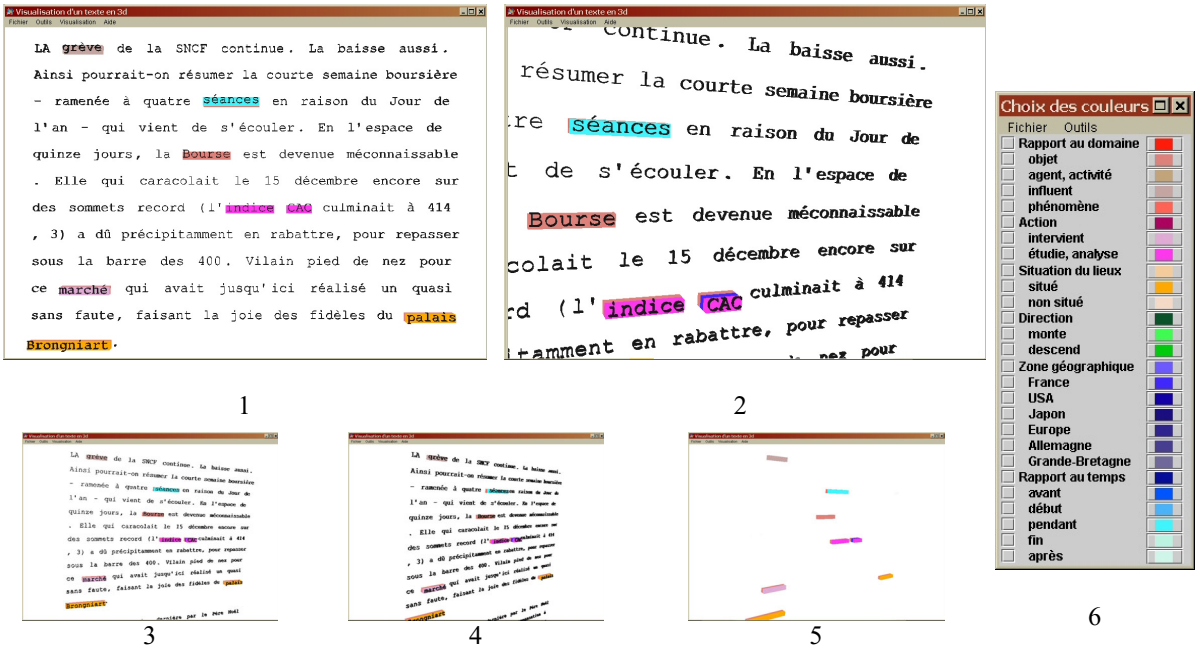
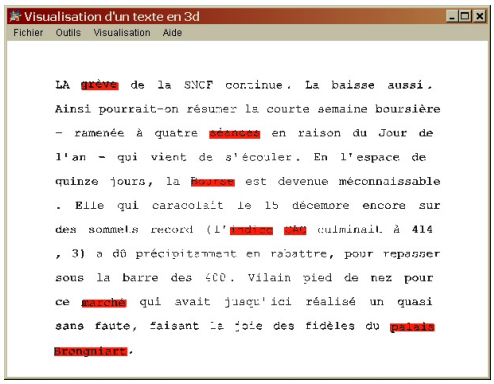


Figure 94 – Copies d'écran de 3D-LUCIAVizualiser

3D-LUCIAVisualizer a été crée à la suite de la réalisation d'un premier prototype de visualisation des récurrences d'attributs/valeurs. Le premier prototype réalisé en Javascript permettait de faire apparaître et disparaître des rectangles colorés superposés aux entités lexicales d'un texte. Les rectangles figuraient la présence d'un attribut associé à une entité lexicale dans un des dispositifs de la session. Au-delà de deux attributs et donc deux rectangles colorés, il était très difficile d'apprécier quoi que ce soit sur le texte : toutes les informations intéressantes pour la tâche ne pouvaient apparaître conjointement. Nous nous sommes alors tournés vers la troisième dimension pour pouvoir apprécier l'association d'une entité avec de multiples attributs et valeurs d'attributs (rappelons qu'il est intéressant de pouvoir évaluer la récurrence d'attributs hérités d'autres tables pour une entité donnée). 3D-LUCIAVisualizer permet ainsi à l'utilisateur de choisir une couleur par attribut ou par valeur d'attribut. Il peut également associer directement un dégradé de couleur pour les attributs d'un même dispositif. Les documents au format XML préalablement analysés sont visibles dans l'interface principale du programme. On peut choisir à loisir les attributs et valeurs d'attributs à faire apparaître ou disparaître. On peut également, comme sur la partie 5 de la figure 94, n'exposer que les boîtes colorées correspondant aux attributs ou aux valeurs d'attributs, faire pivoter en trois dimensions le texte (3 et 4 - figure 94) et zoomer sur les zones intéressantes du texte (2 - figure 94). L'utilisation de cette proposition logicielle nous a amené aux mêmes conclusions que celles proposées dans [Shneiderman, 1996*] (déjà évoqués p.193) : l'utilisation de la 3D montre rapidement ses limites en terme de maniabilité et d'effort à fournir pour repérer les informations à l'écran. Les rotations et les zooms proposés dans 3D-LUCIAVisualizer s'effectuent à l'aide du clavier ; de l'avis des expérimentateurs (moi-même, Stéphane Ferrari et Nicolas Taillepied) l'utilisation prolongée du logiciel est la cause d'une apparition ra-

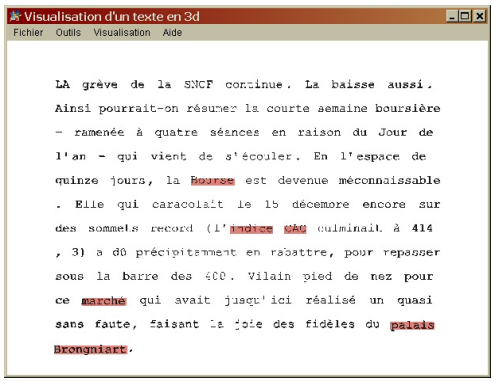
pide des symptômes de la naupathie. Cependant, moyennant un apprentissage minimal et une utilisation modérée des rotations, cette proposition s'avère efficace pour le repérage visuel des récurrences en fonction des données présentées dans une session et donc l'évaluation à la fois des résultats obtenus des logiciels et des ressources proposées initialement. Cette interface permet de mettre en valeur les récurrences d'attributs aussi bien de catégorie (qui sont liés à un thème) qu'hérités (qui structurent le domaine). Par exemple, différentes configurations de la représentation du texte proposé dans la figure 94 permettent le repérage de récurrences et d'en apprécier la répartition dans des parties du texte comme sur les exemples suivants[126] (figure 97).



7 entités lexicales supportent une récurrence de l'attribut [Rapport au domaine] dans le dispositif de la Bourse : *grève*, *séance*, *Bourse*, *indice*, *CAC*, *marché* et *palais Brongniart*.

Le thème de la Bourse semble donc présent et réparti régulièrement dans le paragraphe visualisé. La lecture du texte permet de valider ce fait.

Figure 95 - Visualisation d'isotopies potentielles avec 3D- LUCIAVisualizer (1)

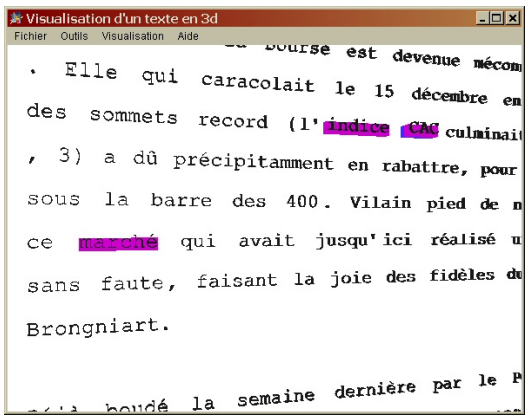


6 entités lexicales du dispositif de la Bourse sont supports d'une récurrence de l'attribut [Rapport au domaine : objet] : *Bourse*, *indice*, *CAC*, *marché*, *Palais Brongniart*.

Le sous-thème « des objets du domaine » semble donc présent et bien réparti dans ce paragraphe.

Figure 96 - Visualisation d'isotopies potentielles avec 3D- LUCIAVisualizer (2)

[126] Les informations sont facilement repérables en faisant apparaître et disparaître une à une les valeurs intéressantes ou en utilisant une simple rotation selon l'axe vertical.



Les entités *indice*, *CAC* et *marché* du dispositif de la Bourse sont supports d'une récurrence de l'attribut [Action]. *CAC* est corrélé à [Zone géographique : France]. Ce dernier attribut n'étant pas redondant, il ne participe à aucune récurrence en fonction du dispositif utilisé.

Figure 97 – Visualisation d'isotopies potentielles avec 3D- LUCIAVisualizer (3)

5.2.3 Facteurs à prendre en considération

De manière formelle, les interfaces que nous proposons répondent aux exigences classiques de ce type d'application :

- elles sont interactives (les couleurs et les groupes thématiques peuvent être modifiés par l'utilisateur, le passage de la souris sur certaines zones permet l'affichage d'informations ciblées…) ;
- elles proposent l'abstraction de certaines données pour rendre leur contenu plus explicite par l'utilisation des graphiques en histogrammes, de représentations schématiques et la possibilité de modifier la quantité d'information affichée ;
- elles proposent la mise en valeur de données pertinentes pour la tâche par un moyen graphique facilement repérable et issu des propositions de l'utilisateur pour une exploration rapide des documents.

Dans les parties précédentes, nous avons présenté les solutions que nous avons élaborées pour visualiser les résultats automatiques obtenus automatiquement à partir des ressources fournies par l'utilisateur et interagir avec eux. Nous avons montré l'intérêt des représentations graphiques qui, plus que les rapports d'exploration textuels, permettent d'évaluer les « rapprochements » qui participent à l'interprétation d'un texte. Si Rastier affirme que la difficulté avec l'ordinateur est de ne pas crouler sous les informations, les interfaces proposées doivent, comme un retour des choses, justement faire la part entre l'information utile et superflue. L'utilisateur peut alors confirmer ou infirmer ses propositions initiales. Pour les deux applications présentées (ISOMETA et la veille documentaire), il est possible de distinguer parmi les besoins d'interaction des aspects génériques réutilisables.

Dans les deux applications, le système n'est pas en mesure de décider pour l'utilisateur de la pertinence des documents. Il permet simplement de les ordonner selon le nombre, la densité et le type

des récurrences trouvées et de présenter ce premier résultat agrémenté éventuellement d'un rapport d'exploration textuel. Le parcours de l'ensemble ainsi constitué est une tâche interactive dont la généricité repose sur l'utilisation d'un corpus, d'une collection de documents et sur la sélection de documents pertinents. Pour cette tâche commune, la généricité est cependant réduite. La disposition de l'ensemble des documents et les fonctionnalités de navigation dans cet ensemble (incluant la possibilité de sélectionner pour observation détaillée un document particulier) sont autant d'aspects génériques. Cependant, la représentation d'un élément dans l'ensemble ne participe plus de la généricité de l'interaction. Chaque document doit être représenté pour permettre à l'utilisateur d'apprécier rapidement sa pertinence, tant absolue que relative à la collection. Qu'il s'agisse de repérer les emplois métaphoriques ou les zones du document en rapport avec un des thèmes d'une veille documentaire, la phase d'exploration d'un texte présente elle-aussi des aspects génériques et des aspects spécifiques. La généricité concerne les fonctionnalités de navigation et de lecture d'un document pour l'assistance. La spécificité des interactions à mettre en œuvre dans cette phase concerne la présentation des résultats d'analyse, dépendante des informations utiles à la tâche. C'est donc la tâche pour laquelle est utilisé le système qui apparaît comme premier facteur à prendre en considération pour l'élaboration des interfaces.

Le modèle sur lequel se fonde une application de TAL constitue un autre facteur de dépendance potentielle pour les interactions et les représentations visuelles. Dans nos travaux, ce facteur est fixe et ne peut être modifié. Pour les deux applications, il est possible d'afficher les résultats d'analyse conjointement au texte d'origine ; une partie de ces résultats est superposée au document affiché pour la lecture. Cette possibilité n'est pas systématique, elle tient en partie à l'existence d'analyses locales, qu'il devient donc possible de situer dans le document. D'autres approches, globales, synthétiques, produisent des résultats qui ne peuvent être mis en relation avec un élément particulier du texte. Il convient donc de noter que le modèle influence les moyens d'interaction, même si nos propres travaux ne permettent pas son analyse. L'approche du global vers le local est commune aux deux applications : c'est en fonction des configurations des récurrences au sein des textes que l'on peut évaluer la validité d'une association locale au niveau d'une entité lexicale. Ceci constitue un autre aspect générique aux deux applications.

L'utilisateur constitue un second facteur à prendre en considération que l'aspect individu-centré de notre approche rend prégnant. C'est à lui que revient la phase finale d'interprétation du matériau textuel. Les deux applications étudiées ont l'avantage de faire intervenir des utilisateurs dont les niveaux d'expertise de la langue, du modèle et des outils qui le mettent en œuvre peuvent être très différents. Or, le niveau d'expertise de l'utilisateur est en rapport direct avec la manière dont les résultats d'analyse sont à présenter. Dans notre approche, son influence commence dès la constitution des ressources lexicales et se poursuit jusqu'à la phase finale d'observation des résultats.

Les applications étudiées tendent à montrer que ce n'est pas tant la tâche qui guide le choix des interactions et des représentations visuelles, que la manière selon laquelle l'utilisateur peut appréhender cette tâche. Lorsque l'utilisateur est supposé expert au départ, toute la richesse du modèle peut se transposer aux interactions et aux représentations qui lui sont proposées. Mais lorsqu'il n'est familier ni du modèle ni des outils, les interactions et les représentations visuelles utilisées doivent avoir autant l'objectif de l'aider dans sa tâche que celui de le former. Il est donc indispensable d'offrir à l'utilisateur la possibilité de s'approprier pleinement toutes ces notions pour, à terme, améliorer son efficacité dans la réalisation de la tâche concernée. Le modèle LUCIA prévoyant un retour sur les ressources à l'issue d'un cycle d'utilisation, la familiarisation avec le modèle passe aussi par un lien fixe entre les ressources et la manière dont il peut y être fait référence lors de la présentation de résultats d'analyse. C'est pourquoi l'association de couleurs aux tables est faite de manière permanente et non pas uniquement lors de chaque phase d'analyse. C'est également le cas pour les attributs et valeurs d'attributs dans 3D-LUCIAVisualizer : les couleurs associées sont conservées dans des fichiers de configuration et peuvent être réutilisées à loisir. Prévoir la manière dont il sera fait référence aux ressources lors de la présentation interactive des résultats à l'utilisateur et utiliser dès leur constitution des méthodes similaires pour présenter ces ressources améliore selon nous la lisibilité des modèles et leur prise en main. Les interactions et les représentations visuelles véhiculent une grande partie de l'information que le logiciel communique à l'utilisateur, leur cohérence avec les notions des modèles mis en œuvre aide donc l'utilisateur à se les approprier.

L'étude des facteurs à prendre en considération pour la réalisation d'interface pour le TAL nous a permis de mieux caractériser l'influence des modèles linguistiques utilisés, du type d'utilisateur, des tâches et du matériau traité sur les interactions nécessaires dans les applications de TAL [Ferrari et Perlerin, 2004]. L'instrumentation informatique de la linguistique est enrichie par les interactions et les moyens de visualiser les informations. Il nous semble donc pertinent de revisiter les propositions de standards et plates-formes d'ingénierie linguistique sous cet angle, en y intégrant les aspects interactionnels trop souvent délaissés. Il convient cependant de prendre conscience des limites de la généricité des interfaces d'interaction et de visualisation et de concevoir au cas par cas des solutions adaptées.

Dans les parties suivantes, nous présentons les processus d'analyse spécifiques et les résultats obtenus dans les deux champs d'applications choisis : l'étude de la métaphore (5.3) et la veille documentaire (5.4).

5.3 Étude de la métaphore

Dans cette partie, nous présentons les résultats obtenus pour l'étude de la métaphore *météorologie boursière* dans le cadre du projet ISOMETA (pour Isotopies et Métaphore). ISOMETA est un projet qui réunit trois chercheurs : Stéphane Ferrari, Pierre Beust et moi-même. Son but est de produire des aides à l'interprétation des métaphores et des indices de détection d'emplois métaphoriques grâce à l'utilisation de LUCIA.

La méthode expérimentale adoptée est la suivante : à partir de l'étude d'un ensemble de documents sélectionnés arbitrairement dans le corpus extrait de *Le Monde sur CD-ROM* qui traite de l'économie et de la bourse (décrit p.118), deux dispositifs en rapport avec les domaines source et cible de la métaphore « météorologie boursière » ont été construits [Perlerin *et al.*, 2002]. Ces données ont ensuite été projetées automatiquement sur le corpus dans son ensemble en fonction des techniques exposées dans les parties précédentes. Les résultats de cette projection ont été analysés manuellement. Nous étions assistés en cela par nos logiciels de visualisation interactive des résultats. Plusieurs itérations du processus expérimental ont été nécessaires. Les premières nous ont amenés à modifier les ressources initialement construites et à apporter des modifications au protocole de construction de départ (5.3.1). Ces modifications ont été intégrées au modèle tel qu'il a été décrit dans les chapitres précédents mais elles sont intéressantes à présenter comme faisant partie d'un cycle d'utilisation du système car elles montrent ainsi les évolutions induites par l'analyse des résultats des logiciels. C'est un argument supplémentaire en faveur d'un processus de recherche et de développement en aller-retour entre des outils (des *logiciels d'étude*) et des corpus (des *corpus d'étude*)[127]. Ces modifications concernent entre autres la distinction entre les attributs partagés et les attributs propres. Cette première expérience a été également en partie décrite dans [Perlerin *et al.*, 2002*]. Cette distinction a permis d'obtenir des résultats pour la caractérisation et l'aide à l'interprétation d'emplois métaphoriques. Elle n'a pas cependant permis de proposer des processus entièrement automatiques pour distinguer des emplois métaphoriques d'emplois non métaphoriques d'entités lexicales du domaine source (5.3.2). Les conclusions et perspectives de cette étude sont présentées dans une dernière section (5.3.3).

Dans toute cette partie, les tables ont été construites par les chercheurs impliqués dans le projet, i.e. par deux spécialistes du modèle (Beust et Perlerin) ainsi qu'un spécialiste du traitement automatique du fait de langue étudié (Ferrari).

[127] Les premiers étant conditionnés par les seconds comme nous l'avons vu en 5.2.

5.3.1 Première expérience

À la suite d'une étude statistique d'une partie du corpus (188 premiers articles – ce qui représentait approximativement un tiers des 565 présents pour un échantillon d'étude i.e. 134 220 tokens), nous avons pu créer des groupes d'entités lexicales en distinguant d'abord celles qui pouvaient avoir trait aux domaines source et cible étudiés et en regroupant ensuite celles qui pouvaient être regroupées dans un thème dans THEMEEDITOR (XXX). Ces 188 premiers articles ont été choisis arbitrairement pour constituer le corpus d'observation. L'ensemble du corpus a ensuite été utilisé pour les analyses. Dans les premiers dispositifs, les attributs ont été choisis pour leur capacité à catégoriser, décrire et distinguer les entités d'un même groupe en fonction des observations qui avaient été faites du corpus. Ceux qui ont été choisis étaient apparus dans un premier temps comme les plus pertinents en fonction de nos expertises. Un consensus parmi les auteurs s'était ainsi dégagé mais certains manques sont apparus à l'issue de la projection de cette première présentation sur le corpus. Les outils de visualisation et d'interaction développés donnaient techniquement entière satisfaction : ils permettaient d'identifier les textes où les deux domaines étaient co-présents et d'observer les attributs/valeurs mis en jeu dans les parties intéressantes de ces textes [Perlerin *et al.*, 2002*]. Cependant, les informations projetées sur les textes du corpus n'offraient pas la possibilité de proposer une caractérisation des emplois métaphoriques ou de fournir une assistance satisfaisante à leur interprétation. Un cycle d'utilisation peut amener à évaluer l'adéquation des données initialement proposées avec une tâche donnée. La figure 98 présente une des tables initialement créées pour le domaine de la météorologie.

Descripteurs de température	**Type**	**Caractère**
surchauffe, torride, étouffant	chaud	fort
chaud, réchauffement, réchauffer	chaud	faible
froid, glacial, polaire	froid	fort
frileux, frilosité, refroidir, petite laine, coup de froid	froid	faible

Figure 98 – Table LUCIA établie pour l'amorce du processus itératif d'après [Perlerin *et al.*, 2002*].

Cette table a été construite en se conformant au protocole de construction des dispositifs que nous avions élaboré initialement assistés des différentes propositions logicielles exposées dans les chapitres précédents. La seule différence avec celui proposé p.174 (c.f. 4.5) était qu'il ne proposait pas encore la distinction entre attributs propres et attributs partagés.

*L'**accalmie** intervenue ensuite sur le front **monétaire**, grâce aux interventions des **banques centrales** et aux déclarations apaisantes des officiels américains, a cependant **réchauffé** l'ambiance à la **corbeille**.*

(12) Extrait de l'article 42 du corpus « Le monde sur CD-ROM ».

Ainsi, pour l'exemple (12), la présence d'isotopies supportées d'une part par *accalmie* et *réchauffé* pour le domaine de la météorologie et par *monétaire*, *banques centrales* et *corbeille* d'autre part, pouvaient être montrées par la visualisation des résultats et les rapports d'exploration obtenus des processus automatiques. Mais les attributs utilisés ne permettaient pas de mettre en place des principes d'aide à l'interprétation des emplois d'*accalmie* et *réchauffé* dans le cotexte de l'article. On pouvait simplement repérer la présence de [Type : chaud] et [Caractère : faible] pour *réchauffé* et les attributs/valeurs correspondant aux entités mises en valeur dans l'exemple (12) décrites dans les dispositifs de l'expérience. En d'autres termes, les éléments de signification potentiels associés aux entités lexicales permettaient de repérer des récurrences en rapport avec chaque domaine décrit mais pas en relation avec les deux domaines. Il était impossible de trouver des récurrences notables d'attributs et valeurs d'attributs des deux domaines dans les textes pour des emplois de la métaphore étudiée. Le manque qui est apparu concernait l'absence d'attributs utilisés pour décrire les deux domaines. Il concernait donc le protocole de construction des dispositifs : c'est cette première utilisation du modèle dans le cadre du projet ISOMETA qui nous a amené à conseiller l'utilisation d'attributs partagés, utilisables pour catégoriser des entités lexicales de domaines distincts.

Ce constat nous a ainsi amené à distinguer les attributs partagés entre plusieurs dispositifs des attributs utilisés que pour un seul domaine (que nous appelons attributs propres). Cette proposition a eu également une répercussion sur nos propositions logicielles. Il était dorénavant apparu indispensable de distinguer les listes d'attributs des dispositifs où ils apparaissaient pour pouvoir facilement en réutiliser d'un dispositif à l'autre et matérialiser la distinction entre les deux types d'attribut (avant cela, les attributs étaient intégrés au fichier du dispositif dans lesquels ils apparaissaient). Nous avons ainsi élaboré le mécanisme des `dictattr`.

Définitions :

- Un attribut propre n'est utilisé pour une session donnée que dans un seul dispositif.
- Un attribut partagé est utilisé pour une session donnée dans au moins deux dispositifs.

Nous voyons que c'est toujours relativement à une tâche et une situation d'utilisation du modèle que l'on peut attribuer des propriétés aux attributs (comme dans la partie 4.4 p.160 du chapitre 4). Dans les faits, comme nous le verrons plus loin, il apparaît que la notion d'attribut partagé est trop restrictive. Si le caractère partagé est relatif à une session, la création et la manipulation d'attributs au cours de l'utilisation du système amène à considérer de façon spontanée des attributs comme potentiellement partageables. Le caractère partageable relève de l'impression de l'usager et bien entendu du ou

des domaines qu'il aborde dans sa session mais aussi de l'observation de l'apparition des attributs dans des textes. Ainsi, pour la création de `dictattr` distincts ou pour l'analyse des métaphores, nous avons été amenés à préférer l'opposition attributs partageables / attributs propres à l'opposition attributs partagés / attributs propres. Nous en verrons la raison principale et l'incidence par l'exemple dans la partie 5.3.2. Le principe de catégorisation proposé dans LUCIA limite à la fois les entités lexicales et les attributs à mettre en jeu. L'utilisation d'attributs partagés n'est pas systématique. C'est une contrainte inhérente à la tâche d'étude du fait de langue dans le cadre d'utilisation du modèle.

Si dans une même chaîne syntagmatique des entités lexicales impliquent la présence et donc la récurrence possible d'attributs propres à deux domaines différents, il est possible dans certains cas d'établir l'existence d'emplois métaphoriques. La présence d'un attribut partagé permet éventuellement de trouver la trace d'un point commun entre les deux domaines tandis que des valeurs différentes de ces attributs permettront d'y trouver une différence. La partie suivante propose des exemples qui mettent en œuvre ces principes.

5.3.2 Observations et résultats

À la suite de la première expérience, nous avons donc revu les dispositifs à partir de la distinction entre attributs propres et attributs partageables. Il s'agissait d'une modification qualitative des données initiales vis-à-vis de la tâche en cours. Les deux dispositifs ont donc été modifiés à l'issue d'un deuxième cycle d'utilisation du système. Celui en rapport avec la météorologie a déjà été présenté (p.159). La figure 84 (p.189) présente celui en rapport avec la bourse et l'économie dans son état actuel. Selon les définitions proposées précédemment, les attributs partagés entre les deux dispositifs sont :

- [Rapport au domaine : objet *vs.* agent, activité *vs*. influent *vs*. phénomène] ;
- [Action : étudie, analyse *vs*. intervient] ;
- [Rapport à l'activité : rôle *vs*. profession] ;
- [Direction : monte *vs*. descend] ;
- [Évaluation : bien *vs*. mal].

Les attributs propres au dispositif en rapport avec la météorologie sont :

- [Axe : agitation *vs*. couverture nuageuse *vs*. température *vs*. pression] ;
- [Etat : liquide *vs*. solide *vs*. gazeux] ;
- [Force : violent *vs*. très violent].

Les attributs propres au dispositif en rapport avec la bourse sont :

- [Zone géographique : France *vs.* USA *vs.* Japon *vs.* Europe *vs.* Allemagne *vs.* Grande-Bretagne] ;
- [Situation : situé *vs.* non situé] ;
- [Rapport au temps : début *vs.* pendant vs. fin].

L'observation de ces attributs amène à considérer spontanément les attributs [Etat : liquide *vs*. solide *vs*. gazeux], [Force : violent *vs*. très violent], [Situation : situé *vs*. non situé] et [Rapport au temps : début *vs*. pendant *vs*. fin] comme possiblement utilisables pour structurer et décrire des entités lexicales d'autres domaines que celui pour lequel ils ont été utilisés. On peut donc les considérer comme partageables. Par exemple [Etat] pourrait être utilisé dans le domaine de la physique, [Force] pour décrire et distinguer des entités en rapport avec des domaines aussi divers que les genres cinématographiques ou certains sports. Le cas de l'attribut [Zone géographique : France *vs*. USA *vs*. Japon *vs*. Europe *vs*. Allemagne *vs*. Grande-Bretagne] est différent car l'opposition proposée met en jeu à la fois des pays et un continent - une entité internationale qui apparaît pertinente pour le domaine de la bourse et l'économie : l'Europe. Si cette opposition s'avère acceptable pour distinguer les entités *franc*, *dollar*, *yen*, *euro*, *deutschemark* et *livre* dans le dispositif de la figure 84 c'est avant tout parce que la période correspondant au corpus d'observation utilisé rendait courante l'utilisation cooccurrente de ces termes (remarquons qu'*écu* a également été retenu alors qu'à l'heure actuelle ce terme n'est plus utilisé pour désigner une monnaie européenne : c'est une réalité diachronique du corpus dont les documents ont été rédigés entre 1987 et 1989). Nous voyons ici que si [Zone géographique] peut être considéré comme un attribut partageable, sa validité en tant que telle est dépendante de la dimension temporelle, et du domaine et de la période du corpus considérée. De manière synchronique avec le corpus, il est considéré comme partageable.

Il est important de noter que l'attribut [Rapport au domaine] a un statut particulier au sein des deux dispositifs. Comme nous pouvons le constater, les deux tables qui l'utilisent dans les deux dispositifs sont constituées majoritairement de lignes vides. D'une manière générale, la table « Entités du domaine » a une forte teneur ontologique. Elle permet de stabiliser le dispositif en étant le point de départ des sous-catégorisations que le composent. Elle permet également de décrire soit des entités très communes pour le domaine (comme *météo* pour le dispositif en rapport avec la météorologie avec [Rapport au domaine : objet]) soit des entités qu'il est finalement difficile de considérer comme ayant trait au domaine (comme *grève* et *salaire* pour le dispositif en rapport avec la bourse et l'économie avec [Rapport au domaine : influent]) – nous avons déjà constaté ce fait lors de l'expérience d'écrite dans le chapitre 3 (partie 3.4 p.104). Les éventuelles récurrences de l'attribut [Rapport au domaine] ne sauraient donner d'indications véritablement spécifiques sur les entités lexicales des textes. Elles n'ont d'ailleurs pas été considérées lors du repérage des récurrences intéressantes pour la métaphore. Un tel

attribut est considéré comme très générique par rapport au domaine et comme trop général pour que ses récurrences soient exploitables pour cette tâche.

La contrainte de construction des dispositifs à l'aide d'attributs partageables permet d'analyser des récurrences trans-dispositifs et ainsi retrouver des résultats connus sur la métaphore mais jamais formalisés comme tels :

- l'absence d'entité lexicale dans un domaine cible d'une métaphore conceptuelle pour certaines valeurs sémantiques peut être comblée par une lexicalisation dans le domaine source (5.3.2.1) ;
- le lien métaphorique entre un domaine source et un domaine cible peut être de nature analogique (5.3.2.2) ;
- le lien métaphorique entre un domaine source et un domaine cible peut amener des éléments de significations nouveaux à partir du domaine source. (5.3.2.3).

5.3.2.1 Entités lexicales absentes du domaine cible

L'absence d'entité lexicale dans un domaine cible d'une métaphore conceptuelle pour certaines valeurs sémantiques peut être comblée par une lexicalisation dans le domaine source. Cette absence donne lieu à des emplois que l'on peut considérer comme métaphoriques. Relativement à la définition du domaine par un dispositif LUCIA, on peut apprécier ce fait dans l'étude des exemples suivants (13), (14), (15) et (16).

*…La **bourrasque** **monétaire** pourrait quand même brouiller une telle appréciation. Le **franc** n'a pas cessé de perdre du terrain face au **mark**, qui atteignait, le 2 janvier, le cours record de 3, 312 F, malgré le soutien de la **monnaie** française par la **Banque de France**…*

(13) Extrait de l'article 1 du corpus « Le monde sur CD-ROM ».

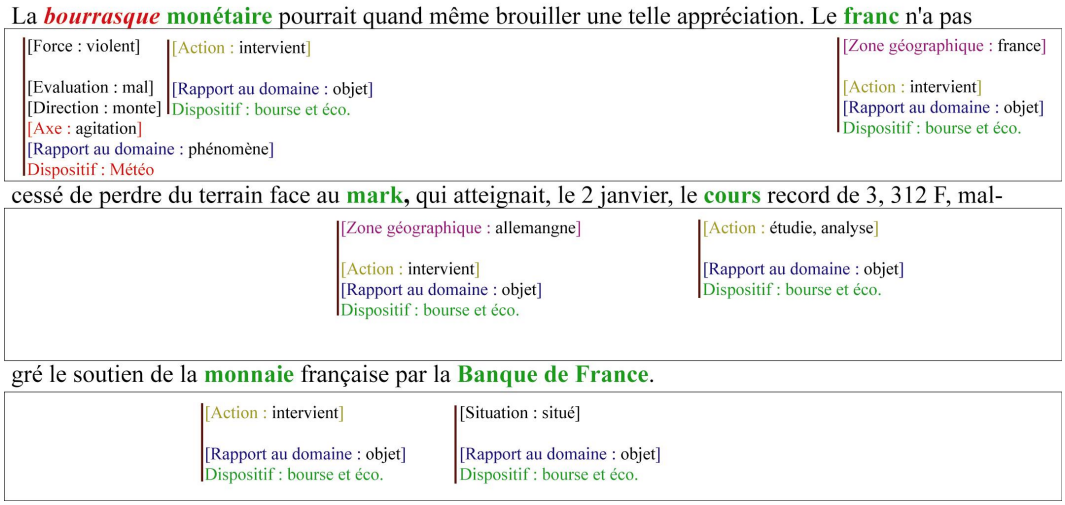


Figure 99 – Attributs et valeurs d'attributs pour l'exemple (13).

Malgré le caractère partagé des attributs [Direction] et [Évaluation], aucune entité lexicale n'a pu être associée à ces attributs avec les valeurs [Direction :monte] et [Évaluation : mal] dans le dispositif en rapport avec la bourse et l'économie. Nous constatons en examinant les attributs de catégories (présentés sur une première ligne dans la figure 99) et les attributs hérités par les entités lexicales présentes dans l'exemple (13) (présentés dans les lignes suivantes dans la figure 99), qu'un certain nombre d'isotopies potentielles apparaissent : une récurrence propre au domaine de la bourse et de l'économie (supportée par *monétaire*, *franc*, *mark*, *cours*, *monnaie* et *Banque de France*), la récurrence des attributs [Action] (supportée par *bourrasque*, *monétaire*, *franc*, *mark*, *cours*, *monnaie* et *Banque de France*), etc. Remarquons que *monnaie française* et *F* (pour franc) ne sont pas repérés comme support d'une isotopie : l'examen d'un tel texte peut amener à ajouter ces entités aux dispositifs utilisés : il s'agit d'une mise à jour quantitative du dispositif. *Monnaie française* et *F* pourraient être ajoutés à la ligne de la table « Monnaie » et décrits par [Zone géographique : France]. L'examen des isotopies qui apparaissent dans le texte de l'exemple (13) permet de pressentir la présence d'un emploi métaphorique de *bourrasque* car le terme n'est pas le support d'un récurrence en rapport avec son domaine. Comme nous le verrons par l'examen des autres exemples, cet indice n'est cependant pas systématique. En revanche, la non-récurrence des attributs partageables [Force] [Évaluation], [Direction] et [Rapport au domaine : phénomène] amène à les considérer comme valables pour assister l'interprétation de ce terme dans le présent contexte. Une réécriture de l'entité lexicale en contexte pourrait être possible. Cette réécriture peut avoir valeur d'assistance à l'interprétation si elle met en jeu les signes, signifiants ou éléments de signification utilisés dans les dispositifs. On peut voir que *bourrasque* est utilisée dans l'exemple (13) au même titre qu'un *phénomène* évalué *mal* qui *monte violemment* dans le domaine de la bourse et l'économie. Rappelons qu'aucune entité lexicale associable au domaine de la bourse et de l'économie n'a été repérée dans le corpus comme pouvant mettre en jeu ces

éléments de signification : un phénomène de la bourse qui monte et qui est évalué négativement – nous n'en avons pas trouvé en dehors d'emplois métaphoriques tels que celui présenté en (13).

*…COURAGE, fuyons... Tel était le slogan en vogue ces jours derniers sous les lambris du **palais Brongniart**, où la **bourrasque monétaire** a fait s'envoler nombre d'**investisseurs** et autant d'espoirs de nouveaux records…*

(14) Extrait de l'article 10 du corpus « Le monde sur CD-ROM ».

L'examen de l'exemple (14) est identique au précédent à la présence près de *s'envoler* qui donne lieu à une isotopie supportée par *bourrasque* et *envoler* qui ne peut être repérée avec les ressources actuelles en terme de récurrence d'attributs/valeurs. Il s'agit de l'une des limites actuelles de l'approche : nous y reviendrons en 5.3.3.

*…Il n'empêche que les **intérêts** pétroliers suscitent encore des convoitises. Abu-Dhabi a profité de la crise pour entrer à hauteur de 5 % dans le **capital** de Total, un associé et ami il est vrai. Le Koweït a racheté, lui, 16% des **actions** BP, **privatisée** en pleine **bourrasque**…*

(15) Extrait de l'article 215 du corpus « Le monde sur CD-ROM » [128].

L'utilisation de *bourrasque* dans des articles relatifs à l'économie est multiple. Remarquons d'ailleurs que les autres entités correspondant aux valeurs d'attributs de catégorie et hérités sont dans le dispositif en rapport avec la météorologie : *rafale*, *tourbillon*, *orage*, *orageux* et *tempête*. Certaines d'entre elles pourraient être substituées à *bourrasque* dans les exemples précédents sans que l'interprétation que l'on puisse faire de ces textes changent profondément (*orageux* ne peut pas faire l'objet d'une telle substitution car il y a une différence de statut grammatical avec les autres entités). Il s'agit là d'une des perspectives du projet ISOMETA (voir 5.3.3).

*… Pour l'instant, le film d'octobre 1989 n'a pas cherché à surpasser dans le grandiose celui d'octobre 1987, encore moins celui d'octobre 1929. Dix jours après la grande **bourrasque**, les effets en ont déjà presque été effacés sur la plupart des grandes **places financières**.*

*La suite du scénario ? Les **météorologues financiers** restent prudents. " On peut avoir à tout moment un nouveau **krach boursier** ", n'a pas hésité à prédire, samedi 21 octobre, M. Maurice Allais, le prix Nobel d'**économie** 1988, à Nice, à l'occasion de l'université annuelle du Club de l'Horloge. Les acteurs **financiers** eux-mêmes sont plongés dans une grande incertitude…*

(16) Extrait de l'article 565 du corpus « Le monde sur CD-ROM »

L'exemple (16) ci-dessus interroge nos propositions quant à la portée des isotopies, c'est-à-dire, dans le cadre d'ISOMETA, les empans de texte à prendre en considération pour pouvoir détecter des indices de la présence d'emplois métaphoriques. En effet, dans les deux paragraphes consécutifs,

[128] Nous remarquerons qu'en fonction du protocole expérimental exposé en début de cette partie, les articles numérotés au-delà de 188 ne faisaient pas partie du sous-corpus ayant servi à l'élaboration des dispositifs.

on constate la présence de deux termes du dispositif en rapport avec la météorologie : *bourrasque* et *météorologues*. L'examen de leurs places dans les dispositifs amène à constater qu'ils ne partagent que l'attribut [Rapport au domaine] dont nous avons déjà souligné le statut particulier. Leur co-présence ne nous permet pas de conclure sur l'utilisation métaphorique des deux termes, cependant que l'examen du texte par un spécialiste, assisté éventuellement par les outils de visualisation et d'interaction créés à cette fin, permettrait de constater un filage de la métaphore météorologique tout le long de l'article. Le filage d'une métaphore consiste précisément à prolonger des récurrences en rapport avec le domaine source au-delà des limites des phrases et des paragraphes. Les principes proposés ici pour assister l'interprétation et la détection des métaphores s'avèrent possible à mettre en place au niveau de la phrase ou du paragraphe mais sur un empan plus grand, le savoir d'un expert est nécessairement requis.

Une récurrence en rapport avec la prédiction et l'incertitude peut être repérée dans l'exemple (16). Elle est supportée par *météorologue*, *prédire* et *incertitude*. Il s'agit d'un cas identique à celui examiné avec *bourrasque* et *s'envoler* dans l'exemple (14). D'une manière générale, de telles configurations ont été soulignées deux fois dans les exemples proposés dans cette partie.

…La semaine écoulée est à cet effet significative. Les 1, 7% de progression n'ont pas été acquis de manière constante. Lundi (+ 0, 14%) et mardi (- 0, 07%) étaient des journées à troubler les plus confiants. Rien ne s'y passait. Le ***calme plat*** *s'installant, certains pensaient que, privé des* ***affaires*** *qui avaient donné du piment à la cote jusqu'alors, le* ***marché*** *"retombait comme un soufflé "…*

(17) Extrait de l'article 365 du corpus « Le monde sur CD-ROM »

…APRES trois semaines de ***calme plat****, la* ***Bourse*** *s'est subitement réveillée et l'ennui s'est très rapidement dissipé. L'activité a repris tant et si bien que l'****indice CAC*** *a pulvérisé son précédent record historique de 470, 4 établi le 24 avril dernier avant que la place parisienne ne sombre dans la léthargie. Vendredi, ce* ***baromètre*** *des valeurs françaises atteignait un nouveau sommet à 478, 5 révélant ainsi une progression de 4% en cinq* ***séances****…*

(18) Extrait de l'article 491 du corpus « Le monde sur CD-ROM »

L'absence d'entité lexicale dans un domaine cible d'une métaphore conceptuelle pour certaines valeurs sémantiques comblées par une lexicalisation décrite de manière similaire dans le domaine source est apparue également pour d'autres exemples. Les exemples (17) et (18) montrent les mêmes phénomènes avec *calme plat* associé à [Évaluation : mal] [Direction : stagne] dans le dispositif en rapport avec la météorologie.

5.3.2.2 Analogie du lien métaphorique

Le lien métaphorique entre un domaine source et un domaine cible peut être de nature analogique. L'examen des exemples suivants permet de voir comment à partir des ressources des deux dispositifs, cette analogie peut être formalisée.

*…Mieux même, les **places financières** semblent avoir retrouvé une nouvelle santé. Des records à la hausse tombent à Tokyo, à **Wall Street** et ailleurs. Le **Dow Jones** par exemple, le **thermomètre** de la **Bourse** de New-York, qui avait chuté de 508 points lors du "lundi noir" déjà rangé dans les rayons de l'histoire **financière**, a repris 102, 9 points mardi, puis 188, 7 points mercredi…*

(19) Extrait de l'article 126 du corpus « Le monde sur CD-ROM »

Dans l'exemple (19), on peut considérer que *thermomètre* est source d'une métaphore *in præsentia* dont la cible est exprimée par *Dow Jones*. Cette métaphore trouve une explication à travers la récurrence de l'attribut [Rôle] avec la valeur [étudie, analyse]. Nous reproduisons ici les extraits des dispositifs correspondant pour mieux apprécier le phénomène.

Objets	Action
couverture nuageuse, agitation, pression, température, ciel	intervient
degré	étudie, analyse

Outils de mesure	Axe
anémomètre	agitation
pluviomètre	couverture nuageuse
thermomètre	température
mercure, baromètre	pression

Figure 100 – Extrait du dispositif en rapport avec la météorologie présenté p.159.

Objets	Action
argent, marché obligataire, MATIF, métal précieux, OPA, SICAV, taux de change, valeur de crédit, volume d'échange, action, baril, devise, dividende, intérêt, marché, marge, métal, monnaie, monétaire, obligation, or, liquidité, portefeuille, produit, taux, tendance, prévision, titre, valeur, volume	intervient
cotation, cote, courbe, cours, graphique, indicateur, indice, parité, ratio	étudie, analyse

Indices boursiers	Zone géographique
CAC	France
Dow Jones, Standard & Poor's, Nasdaq	USA
Nikkei	Japon
Euro Stoxx	Europe
Dax	Allemagne
Footsie	Grande-Bretagne

Figure 101 – Extrait du dispositif en rapport avec la bourse et l'économie présenté p.189.

Les éléments potentiels de significations mis en jeu par les entités *Dow Jones* et *thermomètre* regroupent les attributs de la catégorie ainsi que ceux hérités. Les ensembles des éléments de significa-

tion des deux entités partagent donc l'attribut [Rôle]. Cet héritage d'attributs permet de trouver une récurrence de [Rôle] dans l'exemple (19). Il s'agit de l'équivalent d'une isotopie trans-dispositif, une récurrence supportée par deux entités lexicales de deux dispositifs distincts. La présentation des attributs et leur redondance trouvée par les analyses automatiques permettent de suspecter un emploi métaphorique et de proposer une aide à son interprétation : *thermomètre* est utilisé comme ayant trait à un objet qui sert à l'étude et l'analyse au même titre qu'un *graphique* ou une *courbe*. Une telle réécriture utilisant les attributs et valeurs d'attributs mis en jeu permet de rendre compte de la nature analogique du lien métaphorique. Cette analogie concerne la structuration des deux domaines en fonction des contraintes du modèle. On retrouve le même phénomène avec *thermomètre* et *indicateur*, *indice* dans les exemples (20), (21) et (22). Dans les exemples (21) et (22), la présence de l'attribut spécifiant les indices boursiers en fonction de leur zone géographique supportée par *CAC* et *Nikkei* renforce le phénomène. Cette présence pourrait donner lieu au repérage de récurrences avec *valeurs françaises* pour l'exemple (21) et *Tokyo* et *Kabuto-cho* pour l'exemple (22) si après observation, ces lexies étaient ajoutées au dispositif en rapport avec la bourse.

Tout s'est passé comme prévu. En fin de matinée, ***l'indicateur*** *instantané de* ***tendance*** *enregistrait une avance de 1, 5%. A 12 h 30, au début de la* ***séance*** *principale, des agents des renseignements généraux (RG) patrouillaient sous les lambris pour s'assurer que le dispositif était bien en place. Jusqu'à la clôture à 14 h 30, le* ***thermomètre*** *du* ***marché*** *ne cessa de monter pour s'élever de 3, 4%.*

(20) Extrait de l'article 197 du corpus « Le monde sur CD-ROM »

L'indice CAC*, le* ***thermomètre*** *des* ***valeurs*** *françaises, est revenu au voisinage de la barre des 300 points enfoncée le 28 octobre 1987.*

(21) Extrait de l'article 261 du corpus « Le monde sur CD-ROM »

Mardi, pourtant, et à la suite de ***Wall Street****, les principales places retrouvaient l'optimisme, les* ***cours*** *ouvrant pour la plupart à la hausse. Mardi, la* ***Bourse*** *de Tokyo a regagné une grande partie des pertes de la veille.* ***L'indice Nikkei****, le* ***thermomètre*** *du Kabuto-cho, avait reculé de 1, 8% lundi. Il était en hausse de 1, 5% mardi.*

(22) Extrait de l'article 557 du corpus « Le monde sur CD-ROM »

New-York, c'est la crainte d'un réveil de ***l'inflation*** *qui a déprimé les* ***cours*** *des* ***obligations*** *et poussé à la hausse leurs* ***rendements****, celui de l'emprunt à trente ans du Trésor, véritable* ***thermomètre*** *pour les* ***investisseurs*** *qui bondissait de 8, 50% à plus de 8, 75%.*

(23) Extrait de l'article 285 du corpus « Le monde sur CD-ROM »

L'exemple (23) se distingue par la présence de *emprunt à trente ans du Trésor* qui non seulement n'est pas présent dans le dispositif en rapport avec la bourse mais serait certainement apparu en

corrélation avec [Action : intervient] et non [Action : rôle] sur le modèle des entités déjà présentes dans la table « Objets ». La présence de *cours* pourrait amener à retrouver le lien d'analogie dans l'emploi métaphorique mais il ne s'agit pas véritablement d'un élément *in præsentia* pour la métaphore en question. Il est alors difficile de conclure sur une telle configuration.

Jeudi, un ***rayon de soleil*** *daigna même filtrer à travers les verrières, et le* ***marché****, virtuellement à l'arrêt à l'ouverture matinale (+ 0, 03%) enregistrait en fin de journée une avance de 0, 74%. A la veille du week-end,* ***l'indice CAC****-40, dans le rouge durant la première partie de la journée (- 0, 14%), revenait ensuite dans le vert et y restait (+ 0, 42%). Bref, d'une semaine à l'autre, le* ***mercure*** *est remonté de presque un degré au* ***thermomètre*** *de la* ***Rue Vivienne****.*

(24) Extrait de l'article 574 du corpus « Le monde sur CD-ROM »

La ***température*** *a très brutalement baissé, non seulement à l'extérieur, obligeant les plus* frileux *à remettre une* petite laine*, mais* ***rue Vivienne*** *aussi. La* ***Bourse de Paris****, qui depuis deux mois et demi pataugeait joyeusement dans une interminable consolidation, s'est repliée d'un bloc et à toute allure vers la* ***cote*** *430 de* ***l'indice CAC****, péniblement atteinte au début du mois de mars dernier. Le* ***coup de froid*** *a été sévère (- 5, 5%) et d'autant plus inquiétant pour la " végétation mobilière " que le* ***thermomètre*** *a chuté de 4% au cours de la seule* ***séance*** *de vendredi.*

(25) Extrait de l'article 45 du corpus « Le monde sur CD-ROM »

Dans les exemples (24) et (25), nous sommes en présence d'une alternance de plusieurs entités associées aux deux domaines – remarquons que *frileux* et *petite laine* pourraient éventuellement être intégrées au dispositif de la météorologie à la suite de l'étude de cet exemple. Nous sommes ici en présence du même phénomène que précédemment pour *thermomètre* (présence d'*indice* et *CAC*). Les autres entités de la météo n'apportent pas de support supplémentaire à la récurrence de [Rôle]. On constate ici encore que la métaphore météorologique est filée. La fiabilité des résultats pour la distinction avec un emploi qui serait non métaphorique de *thermomètre* n'a pas été évaluée, il semble que les seules indications obtenues de l'analyse en terme de récurrence d'attributs des textes ne permettent pas de systématiser les principes exposés dans cette partie (c.f. 5.3.3).

5.3.2.3 Nature créatrice du lien métaphorique

Le lien métaphorique entre un domaine source et un domaine cible peut amener des éléments de significations nouveaux à partir du domaine source. Pour apprécier la formalisation de ce fait à l'aide des concepts de LUCIA, nous reproduisons dans les figures suivantes des extraits des deux dispositifs (figure 102 et figure 103).

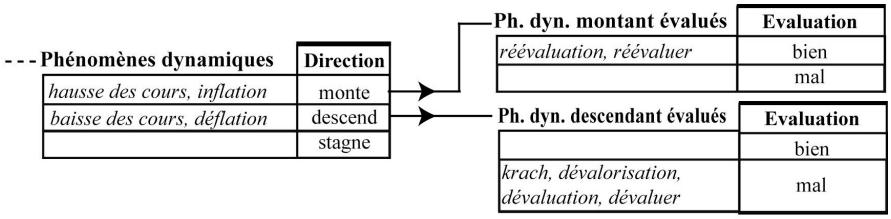


Figure 102 – Extrait du dispositif en rapport avec la bourse et l'économie présenté p.189.

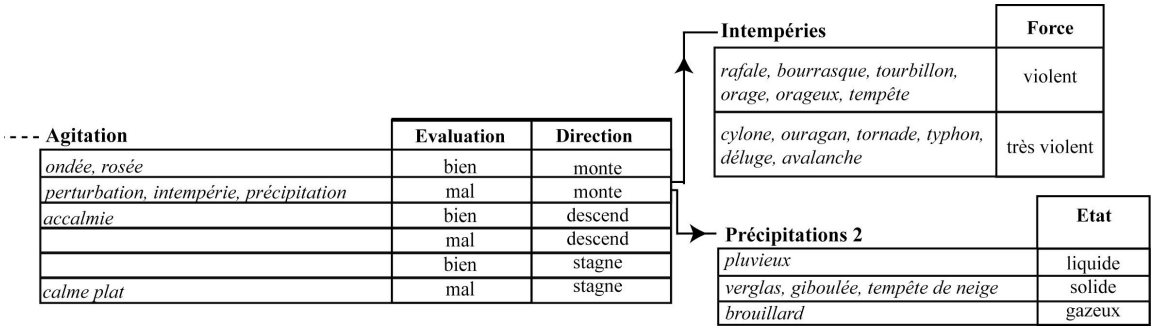


Figure 103 – Extrait du dispositif en rapport avec la météorologie présenté p.159.

*Ce **krach** était dû, en très grande partie, à la chute vertigineuse et incontrôlée du **dollar**, signe que la **tempête** affecte dorénavant les **marchés financiers**.*

(26) Extrait de l'article 153 du corpus « Le monde sur CD-ROM »

L'exemple (26) met en présence les entités *tempête* et *krach. Tempête* est source d'une métaphore *in absentia* dont la cible n'est donc pas lexicalisée dans le cotexte. Deux récurrences des attributs [Direction] et [Évaluation] peuvent cependant être repérées. Celles-ci constituent un faisceau de récurrences mettant en valeur la table des « Phénomènes dynamiques » du domaine de la bourse. Notons également que du point de vue de la description de la cible (et non d'une récurrence) l'attribut [Rapport au domaine] prend une valeur identique pour les deux entités. Les deux lexies sélectionnent la même valeur [Évaluation : mal], cette appréciation apparaît donc particulièrement pertinente dans une interprétation possible de cette métaphore et pour une réécriture. En revanche, les entités lexicales sélectionnent deux valeurs distinctes de l'attribut [Direction] : [Direction : monte] pour *tempête* et [Direction : descend] pour *krach*. Cet attribut est donc moins saillant que l'attribut [Évaluation] pour la réécriture que l'on peut proposer. L'attribut [Force] a été utilisé pour spécialiser certaines « Intempéries », il n'entre pas en jeu dans le faisceau des récurrences repérées par les processus. On considère alors qu'il est transporté depuis le domaine source vers le domaine cible. Les attributs redondants dans la chaîne syntagmatique analysée [Direction] et [Évaluation], constituent les premiers éléments d'une interprétation de la métaphore. Ici, *tempête* est considéré comme ayant trait à *phénomène dynamique* évalué comme *mal*. Nous l'avons précisé en début de cette partie, l'attribut [Force] est un attribut partageable, dans ces circonstances, il est considéré comme caractérisant l'aspect créatif

de cette métaphore. Nous sommes donc tentés de caractériser le *phénomène dynamique boursier* en question comme étant non seulement *évalué mal* (ou mauvais) mais également *violent*. Le transport de [Force] est permis parce-qu'il est partageable.

5.3.3 Conclusions et perspectives pour l'étude de la métaphore

Le projet ISOMETA a permis de caractériser la dynamique des attributs et valeurs d'attributs mis en jeu dans certains emplois métaphoriques et de montrer la capacité de LUCIA à rendre compte d'opérations de transports et d'actualisation d'attributs et valeurs d'attributs en contexte. Nous avons montré à ce propos que l'on pouvait retrouver la nature analogique autant que la nature créatrice de la relation entre la source et la cible d'une métaphore en étudiant les configurations des attributs/valeurs pour chacun de ces phénomènes. Pour rendre compte de l'aspect analogique, les techniques de structuration proposées pour les domaines sont suffisantes. Pour l'aspect créatif, la méthodologie n'est pas encore aboutie et doit être encore développée pour mieux rendre compte des transports possibles de valeurs d'attributs. En particulier, le transport d'attribut doit être étudié au regard du cotexte de leurs apparitions. L'interprétation d'une métaphore n'est pas singulière mais s'inscrit dans le cadre plus générique de l'exploration d'un texte, d'un corpus, d'un parcours interprétatif et la systématisation des transports n'est pas possible automatiquement. On peut cependant envisager de proposer des interactions spécifiques en forme de suggestions produites automatiquement dans ces cas de figures.

La notion déjà connue de métaphore conventionnelle a été observée lors de nos analyses. Dans le corpus d'articles de bourse, nous avons relevé de nombreux emplois de métaphores relevant de la météorologie boursière. On peut effectivement considérer alors, que pour ce corpus, cette métaphore est effectivement conventionnelle (au sens de [Lakoff et Johnson, 1980*]). Il s'agit d'une convention relative à un domaine mais également au genre textuel étudié (le genre journalistique). Faute d'autre corpus, nous ne pouvons confirmer le statut de cette métaphore pour d'autres genres. Dans [Ferrari, 1997*], l'auteur propose une analyse des moyens que peut fournir l'informatique pour automatiser, ou semi-automatiser, la phase d'expertise préalable à la constitution des données relatives aux métaphores conventionnelles. La notion de domaine relativement à la métaphore conceptuelle ne s'avère pas facile à manipuler dans les faits, en particulier en ce qui concerne le domaine source de la métaphore lorsqu'on étudie un corpus du domaine cible. Des exemples analogues au (14) nous amènent à pointer les difficultés que l'on peut rencontrer lorsque l'on cherche à rassembler des entités lexicales relevant d'un domaine source dans le cadre d'une telle étude de fait de langue. Ce fait a déjà été abordé à la suite des travaux de Lakoff et Johnson sur des corpus où l'on constate l'utilisation conjointe de sources multiples que l'on peut considérer parfois comme proches. Dans quelle mesure peut-on par exemple faire correspondre le verbe *s'envoler* au domaine de la météorologie ? Si ce type de question peut se résoudre facilement dans un cadre de veille documentaire où le point de vue de

l'utilisateur et sa façon d'appréhender sa tâche suffisent pour rendre valides les ressources utilisées, il n'en est pas de même pour une étude de fait de langue où l'expertise des usagers détermine la validité des résultats. Ainsi, si les résultats obtenus en terme d'indications pour la détection et de réécritures possibles d'emplois métaphoriques à l'aide de LUCIA s'avèrent satisfaisants et encourageants, il semble que le concept-même de métaphore conceptuelle pourrait être revu à l'aune des domaines et des dimensions au sens de la SI, plutôt que relativement aux entités lexicales elles-mêmes préalablement décrites à l'aide d'attributs. Il est prévisible que l'utilisation de LUCIA dans une telle perspective puisse apporter de nouvelles modifications au modèle lui-même. Nous pensons en particulier que l'utilisation de la représentation des dispositifs en forme de groupes d'attributs/valeurs permettraient de faire *a posteriori* la distinction entre domaines à partir des propositions de structuration des processus automatiques.

Une régularité intéressante a été observée dans certains articles du corpus. Elle concerne les extensions d'une métaphore « au fil » d'un texte qui donnent lieu à des faisceaux de récurrences d'attributs associés à des entités lexicales du domaine source. Jusqu'ici, nous n'avons pas encore exploité cette particularité dans ISOMETA. La distinction automatique entre des emplois métaphoriques filés et la simple présence redondante de termes en rapport avec le domaine source est difficile à caractériser. Il est d'autant plus difficile d'opérer cette distinction que l'on trouve, de manière épisodique, dans le corpus étudié, des textes où cohabitent des emplois métaphoriques et d'autres non métaphoriques. Dans l'exemple (27), des entités associées au domaine de la météorologie s'entrecroisent dans la chaîne paradigmatique avec des entités du domaine du corpus en présentant parfois des emplois métaphoriques (pour *tourbillon*). C'est l'expert qui est ici encore sollicité pour analyser ces cas. Dans de telles circonstances, les outils de visualisation et d'interactions qui ont été présentés apparaissent comme une aide précieuse à l'analyse (c.f. 5.2.2.2).

Le décor changeait de nouveau mercredi. La ***neige*** *faisait une apparition plus que remarquée sur les marches du palais. L'****action*** *Damart, fabricant bien connu de sous-vêtements adaptés contre le* ***froid****, s'envolait, gagnant en* ***pourcentage*** *ce que la* ***température*** *perdait en* ***degrés****. Ce sont encore les* ***intempéries*** *qui poussaient vigoureusement les* ***valeurs pétrolières****, comme Raffinage (+6, 55%) et* ***ELF****-Aquitaine (le* ***titre*** *est monté jusqu'à 351 F jeudi, avant d'être « victime » le lendemain de* ***prises de bénéfices****).*

(…) Certes, les avancées successives de ***Wall Street****, l'envolée de Tokyo et le redressement de Francfort en fin de période donnaient, à Paris, le sentiment agréable d'être pris dans un* ***tourbillon*** *général.*

(27) Extrait de l'article 4 du corpus « Le monde sur CD-ROM »

Les perspectives du projet sont encore multiples. En plus d'évaluer nos propositions et donc entreprendre l'automatisation de certains de leurs aspects (voir 5.5), nous pensons que les aides à l'interprétation que l'on peut fournir peuvent devenir la base d'un système de génération automatique de métaphores. Ce domaine de recherche est déjà étudié pour, par exemple, rendre plus spontanées les

interactions homme/machine en langue naturelle. Il apparaît notamment qu'un manque lexical pour une combinaison d'attributs donnée dans le domaine cible comblé dans le domaine source est un lieu propice à la métaphore. Cependant, il n'est pas possible de systématiser les métaphores possibles. Ainsi dans l'exemple (15), il est envisageable de proposer les entités *bourrasque* et *tempête* pour la construction de métaphores dans le domaine de la bourse ayant trait à des phénomènes violents, évalués comme mauvais. Ceci est rendu possible par la spécification de ces entités dans une catégorie à part entière utilisant l'attribut [Force]. En l'absence de cet attribut, ces entités auraient été présentes dans la ligne de la table de niveau supérieur regroupant des entités telles que *vent*. Or il paraît peu envisageable d'utiliser *vent* dans ce cotexte pour mettre en jeu une métaphore de signification très proche. Certains problèmes restent donc posés : comment distinguer strictement les entités lexicales des domaines sources et cibles, comment juger de la pertinence d'un emploi métaphorique produit pour un certain contexte…

L'absence d'analyse grammaticale du corpus nous permet de limiter les temps de traitements pour analyser rapidement différents corpus sans être dépendants des éventuelles erreurs des analyseurs. Cependant, nous avons déjà abordé le fait que ce principe, qui relève d'une *sémantique légère* pour le TAL, peut amener quelques erreurs d'appariement. Dans le domaine de la météorologie, les entités lexicales ayant trait aux saisons n'ont ainsi pas été intégrées au dispositif. La distinction entre *été* en tant que nom ou forme de l'auxiliaire *être* est en effet impossible sans analyse des textes. Sur cette simple entité, nous avons pu repérer 1176 occurrences dans l'ensemble du corpus. Parmi ces occurrences, seules 74 correspondent au nom, c'est-à-dire moins de 0,6% des occurrences et aucune d'entre elles n'a été envisagée comme métaphorique. Cependant, une analyse plus précise de ces phénomènes d'homographies pourrait apporter quelques informations plus précises sur l'intérêt que pourrait apporter une analyse grammaticale dans ce cadre.

Du point de vue du modèle, la distinction entre attributs propres et attributs partagés a permis également de revenir sur la filiation avec les sèmes de la SI. En corrélation avec la SI, les attributs propres peuvent être les pendants de sèmes mésogénériques (i.e. relatifs à un domaine dans le sens de la SI). Par exemple, dans l'analyse de Hébert [Hébert, 2001] de la tirade «… *Cachez ce sein que je ne saurais voir. …* » dans le Tartuffe de Molière (1622-1673), le chercheur met au jour le sème mésogénérique /sexualité/ pour *sein*[129]. Plus en rapport avec notre corpus d'étude, le sème /bourse/ pour *bourse* et *mi-séance* dans «… *La Bourse de Paris est à l'équilibre à mi-séance...* » est lui-aussi mésogénérique. Les libertés prises avec la SI, nous amènent également à considérer les attributs partageables de LUCIA comme les pendants de sèmes afférents de dimensions inférieures au domaine.

[129] Notons au passage, que comme il est précisé dans [Hébert, 2001*], l'identité des signifiants phoniques *sein* et *saint* permet d'actualiser simultanément dans cette fameuse tirade non seulement le sème mésogénérique /sexualité / mais également /religion/ qui peut être considéré également comme mésogénérique.

Cependant, l'afférence, c'est-à-dire l'inférence qui permet l'actualisation d'un sème afférent en contexte, est systématiquement proposée par le logiciel quelle que soit la validité effective de cette opération : c'est à l'utilisateur d'en juger la conformité avec son interprétation. En SI, les sèmes afférents n'apparaissent que par instruction contextuelle, le contexte de la tâche et le corpus étant fixe pour une étude du type ISOMETA, on peut envisager une actualisation automatique mais pas une validité systématique de cette actualisation. D'une manière générale, il est difficile de trouver une corrélation stricte entre les types d'attributs et les types de sèmes puisque les attributs sont utilisés pour décrire et structurer des significations potentielles d'entités lexicales alors que les sèmes sont le résultat de l'interprétation d'un texte. Cependant, la notion d'attribut propre est particulièrement intéressante dans notre approche de la métaphore : le parallèle avec les sèmes mésogénériques peut être poursuivi jusqu'à l'isotopie et la notion d'*impression référentielle*. Rappelons que pour la SI, les domaines dépendent de normes sociales. Comme il a été souligné dans [Tanguy, 1997b* : 71], la stabilité socioculturelle[130] du découpage de l'univers référentiel par des domaines induit des *effets de référence*. Ces effets sont une ou des isotopies présentes le long de la chaîne syntagmatique de sémèmes relevant d'un même domaine – dans le cadre de LUCIA, nous parlons d'entités lexicales appartenant à un même dispositif ou partageant un même attribut propre.

La partie suivante présente les résultats obtenus avec LUCIA dans un autre champ d'application : celui de la veille documentaire.

5.4 Veille documentaire

Dans cette partie, nous présentons des exemples de résultats de l'utilisation de LUCIA dans un cadre de veille documentaire. L'usager potentiel n'est pas nécessairement un expert en terminologie, en linguistique ou en lexicologie – tout au plus pourra-t-il être un spécialiste de son domaine d'intérêt relativement à la situation de sa tâche, un spécialiste d'un domaine de discours. Le détail des descriptions et leur nature même pourront ainsi être différents que ceux construits dans un cadre d'étude d'un fait de langue. Du point de vue de la finesse des descriptions, un utilisateur peut par exemple n'être intéressé que par des documents qui abordent de façon générale son centre d'intérêt, dans ce cas, le dispositif correspondant pourra être composé d'un nombre restreint de tables rassemblant des entités lexicales très générales par rapport au domaine. D'autres usagers peuvent s'impliquer plus personnellement dans les descriptions et donc mettre l'accent dans les dispositifs sur des attributs en relation avec un jugement ou une évaluation particulière d'un thème ou la façon d'aborder ce thème. D'autres encore peuvent proposer des descriptions fines avec un nombre de tables important. Ce nombre de table peut augmenter à la suite de plusieurs utilisations du système si les premiers résul-

[130] L'identification des domaines se situe au palier sociolectal puisqu'ils sont dépendants de normes sociales.

tats ne sont pas pleinement satisfaisants. Ces possibilités peuvent avoir des répercussions non seulement sur la quantité de données manipulée automatiquement, mais aussi sur la nature des attributs mis en jeu dans les dispositifs. Dans le cas d'une description générale, on peut avoir une majorité d'attributs partageables réinvestis d'autres dispositifs. Dans le cas d'une description précise, les dispositifs peuvent présenter une majorité d'attributs propres spécifiant les entités lexicales retenues. Nous allons voir dans cette partie que ces particularités peuvent influer sur les processus automatiques nécessaires à la veille et qu'elles peuvent être cooccurrentes dans un même dispositif. Ce sont alors les choix de critères de filtrage et d'ordonnancement par rapport à ces données qui permettent de se focaliser sur un aspect particulier du domaine décrit.

Les analyses automatiques qui permettent un repérage des informations des dispositifs à l'intérieur de textes sont utilisées pour une caractérisation chiffrée des textes. Ces caractéristiques sont exploitables pour des processus de filtrage et d'ordonnancement personnalisés. Nous avons vu dans la partie 5.2 l'importance de la présentation des résultats pour la tâche de veille documentaire : les documents peuvent être représentés sous la forme de schémas accompagnés de rapports d'exploration sous la forme de textes qui rassemblent selon divers niveaux de détail les informations repérées et calculées par les logiciels. Dans cette partie, nous présentons un logiciel d'étude, nommé LUCIASearch, qui permet ces opérations automatiques sur des textes inconnus. Ce logiciel est dédié à la veille documentaire sur des documents non formatés de l'Internet (5.4.1). Un exemple d'utilisation nous permet de présenter des exemples de résultats que l'on peut en obtenir (5.4.2). Enfin, les conclusions et les perspectives de ces travaux sont présentées dans une dernière section (5.4.3).

5.4.1 LUCIASearch

LUCIASearch est un logiciel d'étude qui a pour but de montrer la faisabilité et la valeur ajoutée que peuvent apporter des analyses automatiques de documents inconnus à partir de dispositifs. Ce logiciel est un métamoteur : il permet l'interrogation simultanée de plusieurs moteurs de recherche de l'internet. Il filtre et réordonnance les résultats proposés. Il exploite ainsi l'index et le système d'interrogation par requêtes des moteurs tout en palliant leurs manques en terme d'analyse personnalisée de contenu. Il propose une technique de *pull* dans le cadre d'une veille documentaire.

LUCIASearch a été implanté en Java. Il utilise en particulier les classes du *package* `java.net` pour accéder à travers les réseaux à des documents distants de la machine de l'utilisateur. En outre, il exploite les techniques de projections des données des dispositifs exposées en 5.1 et permet la création automatique des interfaces de lecture d'ensemble de documents et de documents spécifiques à la veille documentaire présentées dans 5.2.

La figure suivante propose une vision globale du fonctionnement du système (figure 104), le point de départ correspond à la situation où un utilisateur a déjà constitué un ou plusieurs dispositifs pour la tâche.

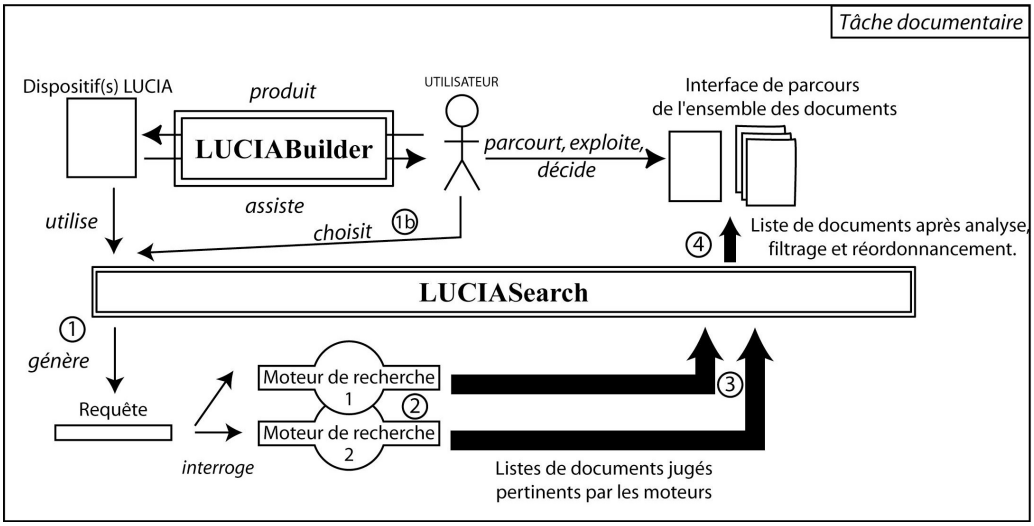


Figure 104 – Fonctionnement général de LUCIASearch.

Un ou plusieurs dispositifs peuvent être exploités par LUCIASearch pour produire une requête par sélection d'éléments particuliers (1 sur la figure 104). On peut sélectionner une ou plusieurs lignes de tables, une table ou plusieurs tables ou un dispositif dans son ensemble. La requête soumise au moteur de recherche est générée en fonction des entités lexicales des éléments sélectionnés. Si le nombre d'élément est trop important, le fonctionnement des moteurs de recherche implique souvent des listes de résultats très courtes voire vides. Pour pallier cet inconvénient, il est également proposé de créer sa propre requête à partir d'une simple liste de mots (1b sur la figure 104). Cette liste peut être proposée à la main par l'utilisateur ou créée à partir de la sélection d'entités lexicales du dispositif utilisé.

Dans sa version actuelle (version 1.0 de juillet 2004), LUCIASearch propose d'interroger soit le moteur de recherche Yahoo, soit le moteur Lycos, soit les deux[131]. La syntaxe de la requête peut intégrer les options de positionnement d'expressions, c'est-à-dire la caractérisation d'une entité complexe pour qu'elle ne soit pas considérée par le moteur comme une suite d'entités simples. Cette fonctionnalité est possible à partir de la caractérisation de la syntaxe des URL de requêtes générées par les moteurs interrogeables par le logiciel. Les pages de résultat des moteurs, obtenues automatiquement à partir de la génération de l'URL correspondante, sont ensuite analysées : les parties intéressan-

[131] Des problèmes inhérents au fonctionnement du moteur Google nous empêchent pour l'instant de proposer ce moteur pourtant très populaire parmi ceux proposés gratuitement sur l'Internet. Ces problèmes concernent la nécessité d'obtenir un *cookie* pour analyser à distance ses résultats.

tes y sont repérées. Les parties intéressantes des pages de résultats sont celles qui correspondent aux informations relatives aux documents résultats. Une étude préalable de l'aspect des pages de résultat des moteurs a été nécessaire pour, par exemple, distinguer les liens publicitaires ou de navigation parmi les autres services du moteur (2 sur la figure 104). Pour les moteurs Yahoo et Lycos, les résultats apparaissent comme les items d'une liste numérotée (balises `<li>`) après une liste de liens sponsorisés. Les documents d'arrivée des liens sponsorisés ne sont pas analysés. Ces liens sont caractérisés par la mise en page HTML (la CSS de Yahoo les placent dans des balises correspondant à une classe nommée `sponsored` tandis que Lycos utilise une liste d'items titrée `Liens sponsoris&eacute;s`).

Les informations proposées par le moteur pour chaque document résultat sont stockées et pourront être affichées plus tard à l'issue de la chaîne de traitements en complément du rapport d'exploration textuel et des représentations schématiques. Ces informations sont par exemple la taille du document (généralement donnée en kilooctets), le titre (récupéré par le moteur à partir de la balise `<title>` de la page HTML), un extrait du texte du document contenant les mots de la requête et bien sûr l'URL qui permet d'y accéder. Le prototype présenté dans la figure 92 (p.202) nous a déjà permis de montrer ces fonctionnalités – rappelons que l'intégration de HTMLtoSVG n'a pas été finalisée et que cette figure présente un prototype qui n'a pas été généré entièrement de façon automatique

L'ensemble des documents proposés par les moteurs est analysé à distance (*parsing* à distance de chaque document correspondant à une URL de résultat – 3 sur la figure 104) grâce à l'URL récupérée. À l'heure actuelle, cette analyse est proposée pour les documents aux formats HTML et TXT – les documents HTML générés dynamiquement par PHP sont bien entendus traités de façon identique au document directement en HTML.

Les premières analyses permettent d'éliminer les documents redondants pour les moteurs interrogés et les URL invalides (type `erreur 404`). Il s'agit de deux fonctionnalités classiques des métamoteurs. Les URL invalides sont repérées lorsque l'accès à travers le réseau produit une erreur. En effet, la mise à jour des index des moteurs demande un certain temps et des documents initialement indexés peuvent avoir été supprimé : l'Internet est une base documentaire très volatile[132]. Si plusieurs moteurs sont interrogés, les documents redondants sont éliminés par comparaison des URL.

Les URL valides sont ensuite explorées. Les documents correspondants sont analysés par projection des ressources contenues dans les dispositifs. Des processus spécifiques permettent de repérer dans les pages HTML les zones textuelles des documents et de les distinguer des zones de navigation, des publicités ou des images. Ces processus ont été décrits en 5.2.2.1 car ils sont également utilisés pour la mise en œuvre de l'interface de parcours rapide de l'ensemble documentaire résultant. Les zo-

[132] L'association américaine *Internet Archive* tente de constituer depuis de nombreuses années une archive générale des documents publiés sur l'Internet pour pallier ce genre de désagréments. http://www.archive.org. La tâche semble cependant bien vaste et ne saurait de toutes façons être exhaustive.

nes de textes sont transformées au format TXT (à l'aide du module HTMLtoTXT dans le cas du HTML) et analysées en terme de repérage des entités lexicales des dispositifs. Les zones textuelles sont repérées à l'aide de la structure HTML des documents. En présence de tableaux, si une cellule de d'un tableau n'a pas été considérée comme une zone de navigation ou ne présente pas uniquement une image ou animation, celle-ci est considérée comme une zone textuelle et fera donc l'objet d'une analyse. Si la structure du document utilise les balises de titres et paragraphes HTML (type `<h1>`, `<h2>`, `<p>`, etc.), une zone textuelle correspond à l'ensemble du texte contenu entre deux balises de plus haut niveau (par exemple entre `<h2>titre1</h2>` et `<h2>titre2</h2>`). Comme nous l'avons déjà signalé, l'utilisation d'HTML est très libre sur l'Internet et il est très courant de ne pas trouver ce type de balises dans les documents. Pour ces cas de figure, les zones de textes correspondent à des paragraphes (`<p>` ou `<br>`) consécutifs. Si les paragraphes présentent moins de 15 mots, ils sont alors agglomérés à la zone textuelle précédente (ou suivante s'il s'agit de la première zone de texte). Ce seuil arbitraire est critiquable car il ne permet pas de distinguer les titres longs des petits paragraphes. Il a été obtenu à la suite d'une expérimentation menée lors d'un projet de DESS[133] [Hubert, 2003*]. Lorsque les documents traités sont au format TXT, les zones textuelles correspondent aux parties de textes distinguées par des sauts de lignes. La même technique d'agglomération de texte que pour les documents HTLM est appliquée.

La mise en corrélation des entités lexicales avec les descriptions des dispositifs permet la mise en place de calculs sur les récurrences d'attributs dans les zones remarquables : les parties de textes sont annotées sur le modèle de ce qui a été présenté en 5.1. En parallèle, l'interface de parcours rapide de l'ensemble documentaire est générée selon les principes exposés en 5.2.2 (4 sur la figure 104).

Dans la partie suivante, nous proposons un exemple d'utilisation de LUCIASearch. Cet exemple permet de montrer la faisabilité et l'intérêt des analyses automatiques de textes inconnus à partir de dispositifs.

[133] DESS RADI : Réseaux, Applications Documentaires et Image du département d'informatique de l'université de Caen.

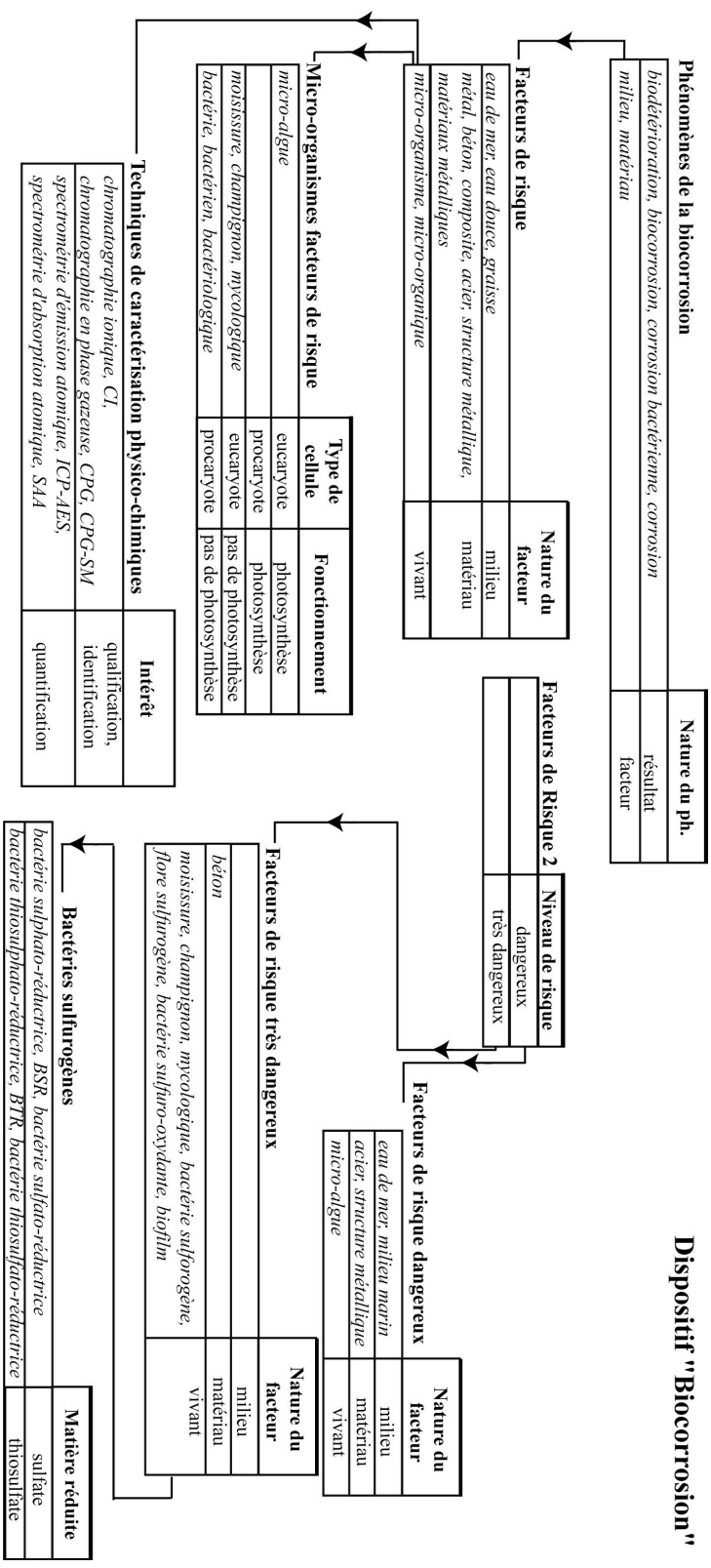


Figure 105 – Dispositif « Biocorrosion »

5.4.2 Exemple d'utilisation

L'exemple d'utilisation de LUCIASearch que nous proposons a été élaboré pour une démonstration de LUCIASearch aux décideurs du CRITT BNC qui nous ont approché en juillet 2004 pour une exploitation du système à des fins professionnelles. Un des partenaires du CRITT BNC est la société CORRODYS - (entité mixte créée par l'Université de Caen et le CRITT BNC, avec le soutien du CEA). CORRODYS propose des services de détection et de solutions préventives et curatives pour les phénomènes de la biocorrosion. C'est donc ce sujet qui a été choisi pour l'élaboration d'un dispositif et la présentation d'un exemple d'utilisation du logiciel d'étude.

5.4.2.1 Dispositif utilisé

L'exemple d'utilisation de LUCIASearch que nous proposons est fondé sur l'exploitation du dispositif présenté en figure 105, p.231 relatif au domaine de la biocorrosion. Ce dispositif a été élaboré à la suite d'une étude (manuelle et statistique en fonction du protocole de construction des dispositifs proposé en p.174 partie 4.5) d'un corpus de textes présentant les causes et les solutions industrielles du problème de la corrosion d'origine bactérienne. Ce corpus constitué de 9 pages du site de la société CORRODYS[134] contient 3 445 mots. Le dispositif a été créé de façon à présenter différents points de vue cooccurrents sur le domaine afin de mieux exposer les possibilités du système. Ce dispositif a été utilisé dans LUCIASearch pour obtenir des documents en rapport avec la *biocorrosion du béton* en fonction de la manière dont a été décrit le domaine dans le dispositif. La biocorrosion du béton est l'un des domaines d'intérêt particulier de la société CORRODYS.

L'examen du dispositif proposé en figure 105 permet de voir que le phénomène de la biocorrosion est appelé également *biodétérioration*, *corrosion bactérienne* ou simplement *corrosion*. Des entités lexicales identiques apparaissent dans des tables distinctes. Nous avons déjà présenté cette possibilité dans le chapitre 3. Dans le cas présent, cette configuration permet l'expression de deux points de vue complémentaires sur ces entités. Par exemple, l'entité *eau de mer* est placée dans la catégorie des « Facteurs de risque » et associée à [Nature du facteur : vivant]. Les tables qui sont des sous-catégories de « Facteurs de risque » permettent de distinguer des entités en fonction des attributs [Type de cellule] et [Fonctionnement]. Il s'agit là d'une description à caractère plutôt référentiel par rapport à des observations biologiques. Ceci traduit une connaissance encyclopédique sur les concepts associables aux entités en question dans le domaine décrit. Ce sont les différences entre ces connaissances qui sont exprimées dans la table en question. La même entité *eau de mer* apparaît également dans la table « Facteurs de risque dangereux » qui est une sous-catégorisation de la table « Facteurs de risque 2 » distinguant des facteurs de risque de la biocorrosion en fonction de leur dangerosité. Il s'agit

[134] http://www.corrodys.com.

là d'un critère de catégorisation plus subjectif, difficilement envisageable dans le cadre d'une description d'un autre domaine – nous avons déjà présenté ce fait dans le chapitre 3 p.100. Les tables qui sont des sous-catégories de « Facteurs de risque 2 » sont renseignées avec les entités lexicales qui apparaissent également parfois dans d'autres tables mais on y a également ajouté des entités plus spécifiques comme *bactérie sulfuro-oxdydante* ou *bactérie sulfurogène*. Ce fait traduit un intérêt particulier à décrire ces entités en terme de risque plutôt qu'en fonction de considérations biologiques. Les critères de validité des structures sont respectés car, entre autres, aucune entité lexicale n'est associée dans le dispositif avec des valeurs différentes d'un même attribut. Une faute d'orthographe courante a également été intégrée au dispositif, il s'agit de *bactérie thiosulphato-réductrice* et *bactérie sulphato-réductrice* qui sont deux formes inspirées de l'orthographe anglo-saxonne redondante dans le corpus d'observation, utilisées pour désigner des bactéries thiosul*f*ato-réductrices ou sul*f*ato-réductrices. La présence d'entités comme *bactérie* (table « Micro-organismes facteurs de risque ») et *bactérie sulfurogène* (table « Facteurs de risque très dangereux ») ne pose pas de problèmes particuliers lors des analyses. Comme nous l'avons indiqué en 5.1, l'heuristique de repérage des entités lexicales dans les textes donne le primat aux entités complexes : les analyses automatiques tiendront compte de la distinction posée dans le dispositif.

5.4.2.2 Expérience

Pour l'exemple, nous avons utilisé une requête constituée des deux mots *béton* et *biocorrosion*. Différentes modalités d'ordonnancement et de filtrage ont été utilisées sur les résultats.

De nombreux termes sont utilisés pour désignés la biocorrosion. Le nombre de requêtes potentielles est donc très important si l'on examine les combinaisons possibles de ces termes avec *béton*[135] pour la recherche prévue. Dans le cadre d'une tâche véritable, il est possible d'analyser les documents obtenus de la simple requête « *béton* » et d'obtenir du logiciel les documents présentant le plus de récurrence des attributs présents dans le dispositif en question classés selon divers critères en fonction de la nécessité de la tâche : on peut favoriser la présence de récurrences d'un ou plusieurs attributs (pour une approche au niveau des isotopies) ou favoriser la présence de certaines entités lexicales associées à une ou plusieurs tables particulières (pour une approche thématique au niveau des tables). Une telle démarche est longue en temps de calcul et produit une quantité très importante d'information : elle n'a donc pas été retenue pour être présentée dans ce tapuscrit.

L'exemple correspond à la configuration du logiciel présentée dans la figure 106. Dans la partie en haut à gauche, une représentation arborescente permet d'avoir accès à toutes les tables du dispo-

[135] Pour donner un ordre d'idée du nombre de documents que proposent un moteur comme Lycos censés être en rapport avec ce sujet, la requête « béton biocorrosion » aboutit à 6 documents, « béton » à plus de 10 000 documents, et « biocorrosion » aboutit à 39 documents.

sitif et de pouvoir y sélectionner des mots pour la requête. Une fois sélectionnée, une entité lexicale est présentée dans la partie en haut à droite avec le nom de la table ainsi que les attributs/valeurs de catégorie correspondants. Le panel du bas permet de sélectionner une table, une ligne de table ou tout le dispositif comme critère de filtrage et d'ordonnancement. Dans la figure 106, il s'agit de la configuration par défaut : l'ensemble du dispositif va être utilisé pour classer les documents lors de l'analyse : la partie en bas à droite est vierge. Si l'on désire favoriser la présence d'une ou plusieurs tables et/ou la présence d'un ou plusieurs attributs ou valeurs d'attributs, la sélection se fait sur le même principe que pour les entités lexicales.

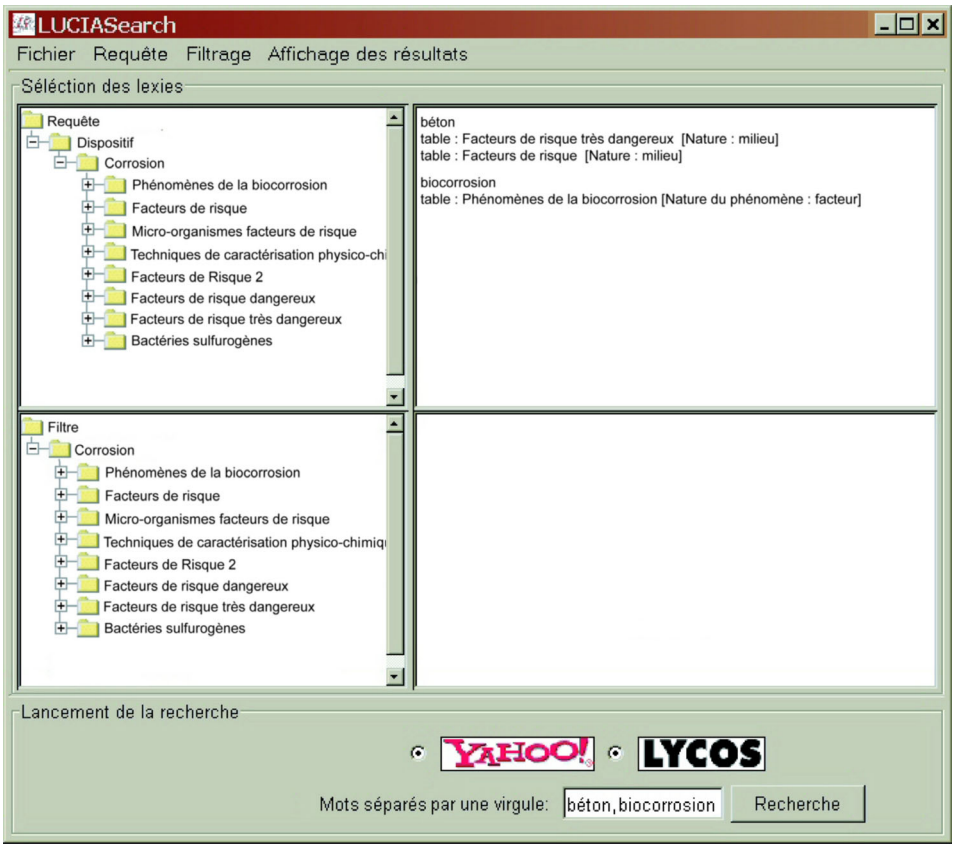


Figure 106 – Copie d'écran de LUCIASearch pour la requête « *béton biocorrosion* » avec tout le dispositif comme critère de filtrage.

L' URL de la requête générée automatiquement pour le moteur Yahoo est :

```
http://fr.search.yahoo.com/search/fr?ei=UTF-8&fr=fp-tab-web-
t&x=op&va=biocorrosion+b%C3%A9ton&va_vt=any&vp=&vp_vt=any&vo=&vo_vt=any&ve=&ve_vt=any&vd=all&v
st=0&vs=&vf=all&vc=countryFR&fl=1&vl=lang_fr&n=20&brfw=Lancer+la+recherche
```

Pour le moteur Lycos, l'URL de la requête est la suivante :

```
http://vachercher.lycos.fr/cgi- bin/pursuit?query=b%E9ton+biocorrosion&x=0&y=0& cat=fr&tld
=com&family=off
```

Documents proposés (titre, URL, présentation)	Classement	
	Lycos	Yahoo
BTP – Génie Civil http://www.adit.fr/adit_edition/domaines/trans_btp_gc.html *ensemble d'hyperliens vers des brèves scientifiques.*	-	1
Le laboratoire de biocorrosion / biodétérioration… http://www.crittbnc.com/pageLibre000100c4.html *présentation à caractère publicitaire de problèmes de biocorrosion soumis au laboratoire CORRODYS.*	1	2
CEFRACOR, Centre français de l'anticorrosion http://www.euroanticorrosion.fr/cefracor/welcome.html *présentation et coordonnées du CEFRACOR*	2	3
LABORATOIRE CORROSION–FRAGILISATION HYDROGÈNE – CFH www.ecp.fr/fr/F-recherche/F1-laboratoires/cfh.pdf *rapport d'activité 1998-1999 du laboratoire CFH de Châtenay-Malabry*	-	4
Université Paul Sabatier - Toulouse III - France : Les laboratoires de recherche www.ups-tlse.fr/RECHERCHE/acces_labo_par_motcle.php3?oper=1&lettre=B *liste d'hyperliens vers les thématiques des laboratoires de recherche de l'université de Toulouse III.*	-	5
Matériaux – Chimie www.adit.fr/adit_edition/domaines/mat_chimie.html *ensemble d'hyperliens vers des brèves scientifiques.*	3	6
www.u-picardie.fr/~beaucham/duee/vrignaud.htm *Article scientifique intitulé « La détérioration des canalisations »*	4	7
Opération 8 www.ccr.jussieu.fr/lple/OP8.html *Présentation d'une « opération de recherche » intitulé « Electrochimie dans les eaux naturelles »*	5	8

Figure 107 – Ensemble des documents obtenus.

L'ensemble des documents proposés par les moteurs et analysés automatiquement par le logiciel sont présentés dans la figure 107. Les titres proposés sont ceux repérés par les moteurs, la courte présentation du document a été rédigée par nos soins. Cette figure montre également l'ordre de classement de ces documents pour chacun des moteurs. Tous les documents proposés par le moteur Yahoo sont présentés également par le moteur Lycos qui ne propose pas de document nouveau par rapport à l'autre moteur – son intérêt est donc nul dans l'exemple. Notons que le document intitulé « *Laboratoire Corrosion – Fragilisation Hydrogène – CFH* » n'est pas analysable par le logiciel pour l'instant car il est au format PDF (c.f. 5.4.3).

Pour représenter l'ensemble des analyses et des calculs automatiques, nous proposons de nous attarder particulièrement sur le document intitulé « *Le Laboratoire de biocorrosion / biodétérioration* ». La figure 108 propose une représentation schématique des zones analysées dans ce document. Elle précise les entités lexicales du dispositif qui y ont été repérées avec les attributs de catégorie et les attributs hérités correspondants (les attributs/valeurs hérités sont précédés d'un +).

Figure 108 – Schématisation du repérage automatique des récurrences.

La structure HTML du document est constituée d'un ensemble de tables. Les cellules de ces tables et ces tables elles-mêmes ont été considérées comme des zones textuelles si elles n'avaient pas été repérées en tant que zones de navigation (comme c'est le cas pour la partie sous le titre du document « *Exemples d'études confiées au laboratoire de biodétérioration* » et celle à gauche sous le texte « *Laboratoire de biocorrosion / biodétérioration CORRODYS* »**)**. Sur la figure, les zones textuelles ont été encadrées de noir, le document a été tronqué car aucune zone remarquable (présentant au moins une occurrence d'une entité du dispositif) n'est présente au-delà de la partie présentée. L'analyse du document a mis au jour la présence de 168 graphies n'appartenant pas à la *stop-list* (par abus de langage nous utiliserons le terme *entité non vide*) : les calculs sur les entités lexicales supports de récurrences d'attributs et valeurs d'attributs pourront ainsi être relativisés par rapport à l'ensemble des entités non vides du document.

Les figures suivantes rassemblent les résultats du calcul fondé sur ces informations selon deux paliers d'analyse : le document dans son ensemble et les zones textuelles repérées.

Attribut	Nombre et entités lexicales supports de la récurrence	Valeur d'attribut	Nombre et entités lexicales supports de la récurrence
[Nature du phénomène]	9 (5) 13,3% (3%) biocorrosion, biodétérioration (3), structure métallique, milieu marin (2), corrosion (2), béton	[résultat]	6 (3) 3,6% (1,8%) biocorrosion, biodétérioration (3), corrosion (2),
		[facteur]	4 (3) 2,4% (1,8%) milieu marin (2), structures métalliques, béton
[Nature du facteur]	6 (5) 3,6% (3%) flore sulfurogène, milieu marin (2), structure métallique, bactérie sulfuro-oxydante, béton	[vivant]	2 (2) 1,2% (1,2%) flore sulfurogène, bactérie sulfuro-oxydante
		[milieu]	2 (1) 1,2% (0,6%) milieu marin (2)
		[matériau]	2 (2) 1,2% (1,2%) structure métallique, béton
[Niveau de risque]	6 (5) 3,6% (3%) flore sulfurogène, milieu marin (2), structure métallique, bactérie sulfuro-oxydante, béton	[dangereux]	2 (1) 1,2% (0,6%) milieu marin (2)
		[très dangereux]	4 (4) 2,4% (2,4%) flore sulfurogène, structure métallique, bactérie sulfuro-oxydante, béton

Figure 109 – Résultat du calcul des récurrences d'attributs et valeurs d'attributs pour le document « *Le Laboratoire de biocorrosion / biodétérioration* ».

La figure 109 présente le résultat du calcul des récurrences d'attributs et valeurs d'attributs pour l'ensemble du document. Les chiffres présentés entre parenthèses correspondent aux calculs relatifs au nombre d'entités différentes supports d'une récurrence. Les pourcentages correspondent au nombre relatif d'entités support d'une récurrence par rapport au nombre d'entités non vides. L'attribut majoritairement redondant dans le document est [Nature du phénomène]. La valeur d'attribut majoritairement redondante dans le document est [Nature du phénomène : résultat]. Les attributs [Nature du facteur] et [Niveau de risque] présentent des nombres de récurrences absolues et relatives identiques. Le contenu du document semble correspondre au domaine du dispositif dans son ensemble.

Zone de texte	Attribut	Nombre et entités lexicales supports de la récurrence	Valeur d'attribut	Nombre et entités lexicales supports de la récurrence
1	[Nature du phénomène]	2 (2) 50% (50%) biocorrosion, biodétérioration	[résultat]	2 (2) 50% (50%) biocorrosion, biodétérioration
3	[Nature du facteur]	2 (2) 22% (22%) flore sulfurogène, milieu marin	[vivant]	1 (1) 11,1% (11,1%) flore sulfurogène
			[milieu]	1 (1) 11,1% (11,1%) milieu marin
	[Niveau de risque]	2 (2) 22% (22%) flore sulfurogène, milieu marin	[dangereux]	1 (1) 11,1% (11,1%) milieu marin
			[très dangereux]	1 (1) 11,1% (11,1%) flore sulfurogène
4	[Nature du phénomène]	3 (3) 12% (12%) corrosion, structure métallique, milieu marin	[résultat]	1 (1) 4% (4%) corrosion
			[facteur]	2 (2) 8% (8%) structure métallique, milieu marin
6	[Nature du phénomène]	2 (2) 13,3% (13,3%) corrosion, béton	[résultat]	1(1) 6,7% (6,7%) corrosion
			[facteur]	1 (1) 6,7% (6,7%) béton

Zone 1 : 4 entités non vides, Zone 3 : 9 entités non vides, Zone 4 : 25 entités non vide, Zone 6 :15 entités non vides.

Figure 110 - Résultat du calcul des récurrences d'attributs et valeurs d'attributs par zones textuelles du document « *Le Laboratoire de biocorrosion / biodétérioration* ».

La figure 110 présente le nombre de récurrences d'attributs et valeurs d'attributs par zone de document. Elle montre que c'est la zone textuelle numéro 3 qui présente le plus de récurrences d'attributs et valeurs d'attributs différentes de manière absolue, tandis que c'est la zone numéro 1 qui en présente le plus de façon relative. Cependant, l'analyse n'a repéré dans cette zone que 4 entités non vides, c'est donc une petite zone et il est donc probable que son contenu n'est pas très intéressant (il s'agit en l'occurrence simplement du nom du laboratoire). Les attributs [Nature du facteur] et [Niveau de risque] sont récurrents dans la zone 3 du fait de la présence des mêmes entités lexicales. Dans la zone 6, cooccurrent deux entités lexicales en rapport avec les mots de la requête : *béton* qui est présent tel quel dans la requête et *corrosion* qui partage la même valeur d'attribut que *biocorrosion* dans la table des « Phénomènes de la biocorrosion ». Cette zone comporte 15 entités non vides et l'attribut [Nature du phénomène] est récurrent pour 13,3% des entités de cette zone. Elle apparaît donc comme la zone la plus intéressante relativement à la requête.

Table	Entités lexicales de la table présentes dans le document
Phénomènes de la biocorrosion	7 (3) 4,2% (1,8%) biocorrosion (1), biodétérioration(3), corrosion (3)
Facteurs de risque	3 (2) 1,8% (1,2%) milieu marin (2), béton (1)
Facteurs de risque dangereux	3 (2) 1,8% (1,2%) milieu marin(2) , structure métallique (1)
Facteurs de risque très dangereux	3 (3) 1,8% (1,8%) flore sulfurogène(1), bactérie sulfuro-oxydante (1), béton (1)

Figure 111 - Résultat du calcul des occurrences d'entités lexicales en fonction des tables du dispositif sur l'ensemble du document « *Le Laboratoire de biocorrosion / biodétérioration* ».

La figure 111 présente le résultat du calcul des occurrences d'entités lexicales en fonction des tables dans lesquelles elles apparaissent. Elle montre que la table majoritairement représentée dans le document est l'une des plus générales au dispositif : « Phénomènes de la biocorrosion » (4,2% des entités du document appartiennent à cette table) et que les tables « Facteurs de risque », « Facteurs de risque dangereux » et « Facteurs de risque très dangereux » sont présentes en proportions équivalentes. Tous les thèmes décrits dans le dispositif ne sont donc pas présents. En particulier, aucune des tables les plus spécifiques au domaine (*Technique de caractérisation physico-chimiques* et *Bactéries sulfurogènes*) n'est représentée.

L'ensemble de ces calculs peut être réitéré sur tous les documents retenus. Ils sont proposés ici pour l'exemple mais ils peuvent être envisagés un par un, une fois les informations des dispositifs projetées sur les documents et les zones de ces documents repérées par le système. C'est là l'intérêt d'annoter les textes et non de produire la projection des données au coup par coup.

Les figures suivantes présentent les résultats obtenus pour l'ensemble des documents présenté dans la figure 107. Dans les documents intitulés *BTP-Génie Civil*, *Matériaux – Chimie* et *Université Paul Sabatier – Toulouse III – France : Les laboratoires de recherche* n'ont été repérées que des zones de navigation ou des zones textuelles qui ne présentent aucune récurrence d'attributs : ils ne sont pas présentés dans cette liste de résultats (c.f. 5.4.2.3).

Pour le document *CEFRACOR*, les trois entités *corrosion*, *biodétérioration* et *matériaux* sont présentes dans une même zone textuelle. Cette zone est la seule du document à présenter une récurrence d'attributs/valeurs. Celui-ci semble donc peu intéressant pour la tâche. L'analyse a permis de repérer 272 entités non vides dans l'ensemble du document et 33 dans la zone en question.

Attribut	**Nombre et entités lexicales supports de la récurrence**	**Valeur d'attribut**	**Nombre et entités lexicales supports de la récurrence**
[Nature du phénomène]	4 (3) 1,5% (1,1%) corrosion (2), biodétérioration, matériau	**[résultat]**	3 (2) 1,1% (0,7%) corrosion (2), biodétérioration
		[facteur]	1 (1) 0,4% (0,4%) matériau

Figure 112 - Résultat du calcul des récurrences d'attributs et valeurs d'attributs pour l'ensemble du document *CEFRACOR*.

Zone de texte	Attribut	Nombre et entités lexicales supports de la récurrence	Valeur d'attribut	Nombre et entités lexicales supports de la récurrence
5	[Nature du phénomène]	4 (3) 12,1% (9%) corrosion (2), biodétérioration, matériau	[résultat]	3 (2) 9% (6%) corrosion (2), biodétérioration
			[facteur]	1 (1) 3% (3%) matériau

Zone 5 : 33 entités non vides.

Figure 113 - Résultat du calcul des récurrences d'attributs et valeurs d'attributs pour la zone textuelle remarquable du document *CEFRACOR*.

Table	Entités lexicales de la table présentes dans le document
Phénomènes de la biocorrosion	4 (3) 1,5% (1,1%) corrosion (2), biodétérioration, matériau

Figure 114 - Résultat du calcul des occurrences d'entités lexicales en fonction des tables du dispositif sur l'ensemble du document *CEFRACOR*.

Pour le document *Opération 8*, l'analyse a permis de repérer 351 graphies non vides et de distinguer six zones textuelles dont quatre présentent des récurrences d'attributs (zones 2, 3, 4 et 6). Une zone supplémentaire fait apparaître une occurrence de *corrosion* (zone 2). Elle présente donc une récurrence supportée par la même entité lexicale. Cette récurrence est moins intéressante que si elle était supportée par deux entités distinctes.

Attribut	Nombre et entités lexicales supports de la récurrence	Valeur d'attribut	Nombre et entités lexicales supports de la récurrence
[Nature du phénomène]	14 (6) 4% (1,7%) corrosion (5), biocorrosion, matériau (3), métal (3), eau de mer , acier	[résultat]	6 (2) 1.7% (0,6%) corrosion (5), biocorrosion (1)
		[facteur]	8 (4) 2.3% (1,1%) matériau (3), métal (3), eau de mer (1), acier (1)
[Nature du facteur]	8 (5) 2,3% (1,4%) biofilm (2), béton , métal (3), eau de mer, acier	[milieu]	1 (1) 0,3% (0,3%) eau de mer
		[matériau]	5 (3) 1,4% (0,8%) béton, métal (3), acier
		[vivant]	2 (1) 0,6% (0,3%) biofilm (2)
[Niveau de risque]	5 (4) 1,4% (1,1%) biofilm (2), béton, eau de mer, acier	[dangereux]	2 (2) 0,6% (0,6%) eau de mer, acier
		[très dangereux]	3 (2) 0,8% (0,6%) biofilm (2), béton

Figure 115 - Résultat du calcul des récurrences d'attributs et valeurs d'attributs pour l'ensemble le document *Opération 8*.

Zone de texte	Attribut	Nombre et entités lexicales supports de la récurrence	Valeur d'attribut	Nombre et entités lexicales supports de la récurrence
2	[Nature du phénomène]	2 (1) 7,1% (0,4%) corrosion (2)	[résultat]	2 (1) 7,1% (0,4%) corrosion (2)
3	[Nature du phénomène]	6 (3) 8,9% (4,4%) matériau (3), corrosin (2), métal	[facteur]	4 (2) 5,9% (3%) matériau (3), métal
			[résultat]	2 (1) 3% (1,5%) corrosion (2)
	[Nature du facteur]	4 (3) 5,9% (4,4%) biofilm (2), béton, métal	[vivant]	2 (1) 3% (1,5%) biofilm (2)
			[résultat]	2 (2) 3% (3%) béton, métal
	[Niveau de risque]	3 (2) 4,4% (3%) biofilm (2), béton	[très dange-reux]	3 (2) 4,4% (3%) biofilm (2), béton
4	[Nature du phénomène]	3 (2) 8,8% (5,9%) corrosion, métal (2)	[facteur]	2 (1) 5,9% (2,9%) métal (2)
6	[Nature du phénomène]	4 (4) 15,3% (15,3%) eau de mer, corrosion, acier, biocorrosion	[facteur]	2 (2) 7,7% (7,7%) eau de mer, acier
			[résultat]	2 (2) 7,7% (7,7%) corrosion, biocorrosion
	[Nature du facteur]	2 (2) 7,7% (7,7%) eau de mer, acier		
	[Niveau de risque]	2 (2) 7,7% (7,7%) eau de mer, acier	[dangereux]	2 (2) 7,7% (7,7%) eau de mer, acier

Zone 2: 28 entités non vides Zone 3: 68 entités non vides Zone 4: 34 entités non vides Zone 6: 26 entités non vides

Figure 116 - Résultat du calcul des récurrences d'attributs et valeurs d'attributs par zones textuelles du document *Opération 8*.

Table	Entités lexicales de la table présentes dans le document
Phénomènes de la biocorrosion	9 (3) 2,6% (0,8%) corrosion (5), biocorrosion (1), matériau (3)
Facteurs de risque	6 (4) 1,7% (1,1%) eau de mer (1), métal (3), béton (1), acier (1)
Facteurs de risque dangereux	2 (2) 0,6% (0,6%) eau de mer (1), acier (1)
Facteurs de risque très dangereux	3 (2) 1,2% (0,6%) béton (1), biofilm (2)

Figure 117 - Résultat du calcul des occurrences d'entités lexicales en fonction des tables du dispositif sur l'ensemble du document *Opération 8*.

Pour le document sans titre classé numéro 7 par le moteur Yahoo (nous l'appellerons *Vrignaud* du nom de son auteur), l'analyse a permis de repérer 4 896 graphies non vides et de distinguer 75 zones textuelles en tout. Vu la taille du document, nous ne présentons pas le détail des résultats du calcul des récurrences d'attributs et valeurs d'attributs pour toutes les zones textuelles. Nous présenterons celles qui peuvent s'avérer intéressantes en fonction des critères d'ordonnancement et de filtrage choisis en fin de partie.

Attribut	Nombre et entités lexicales support de la récurrence	Valeur d'attribut	Nombre et entités lexicales support de la récurrence
[Nature du phénomène]	126 (12) 2,6% (0,2%) biocorrosion (5), corrosion (61), corrosion bactérienne, milieu (9), matériau (5), eau douce, métal (17), béton (5), acier (6), micro -organisme (5), bactérie (9), bactérien (2)	[résultat]	67 (3) 1,4% biocorrosion (5), corrosion (61), corrosion bactérienne
		[facteur]	59 (9) 1,2% (0,2%) milieu (9), matériau (5), eau douce, métal (17), béton (5), acier (6), micro -organisme (5), bactérie (9), bactérien (2)
[Nature du facteur]	52 (8) 1% (0,2%) eau douce, métal (17), béton (5), acier (6), micro -organisme (5), bactérie (9), bactérien (2), bactérie sulfatoréductrice (7)	[matériau]	28 (3) 0,6% métal (17), béton (5), acier (6)
		[vivant]	23 (4) 0,1% micro-organisme (5), bactérie (9), bactérien (2), bactérie sulfato-réductrice (7)
[Niveau de risque]	20 (4) 0,4% (0%) béton (5), biofilm (2), bactérie sulfato-réductrice (7), acier (6)	[dangereux]	6 (1) 0,1% acier (6)
		[très dangereux]	14 (3) 0,3% béton (5), biofilm (2), bactérie-sulfato-réductrice (7)
[Type de cellule]	11 (2) 0,2% bactérie (9), bactérien (2)	[procaryote]	11 (2) 0,2% bactérie (9), bactérien (2)
[Fonctionnement]	11 (2) 0,2% bactérie (9), bactérien (2)	[pas de photosynthèse]	11 (2) 0,2% bactérie (9), bactérien (2)
[Matière réduite]	7 (1) 0,1% bactérie sulfato - réductrice (7)	[sulfate]	7 (1) 0,1% bactérie sulfato - réductrice (7)

Figure 118 - Résultat du calcul des récurrences d'attributs et valeurs d'attributs pour l'ensemble le document *Vrignaud*.

Table	Entités lexicales de la table présentes dans le document
Phénomènes de la biocorrosion	81 (5) 1,6% biocorrosion (5), corrosion (61), corrosion bactérienne, milieu (9), matériau (5)
Facteurs de risque	34 (5) 0,7% eau douce, métal (17), béton (5), acier (6), micro-organisme (5)
Micro-organismes facteurs de risque	11 (2) 0,2% bactérie (9), bactérien (2)
Facteurs de risque dangereux	6 (1) 0,1% acier (6)
Facteurs de risque très dangereux	5 (2) 0,1% béton (5), biofilm (2)
Bactéries sulfurogènes	7 (1) 0,1% bactérie sulfato-réductrice (7)

Figure 119 - Résultat du calcul des occurrences d'entités lexicales en fonction des tables du dispositif sur l'ensemble du document *Vrignaud*.

Pour l'ensemble des résultats obtenus, différents critères de filtrage et d'ordonnancement sont possibles. Le critère « Nombre d'attributs représentés » permet d'ordonnancer les documents en fonction de la présence du plus grand nombre d'attributs différents. Ce critère permet d'obtenir des documents qui maximisent les aspects différents du domaine décrit dans le dispositif. Il est possible de donner un nombre minimum d'attributs qu'un document doit présenter pour être retenu dans la liste. Dans l'exemple étudié, si l'on désire avoir aux moins trois attributs représentés, le classement obtenu est le suivant : *Vrignaud*, *Le Laboratoire de biocorrosion / biodétérioration* et *Opération 8*. Ce classement permet d'aboutir à des documents qui abordent au moins plusieurs aspects du domaine.

Le critère « Présence d'attributs particuliers » permet de ne garder dans les résultats que des documents qui présentent au minimum des récurrences d'attributs ou attributs/valeurs sélectionnés.

Dans l'exemple, si l'on désire ne voir apparaître que des documents qui présente des récurrences de [Niveau de risque : très dangereux], l'ordre final en se basant sur les valeurs absolues d'entités supports est : *Vrignaud*, *Le Laboratoire de biocorrosion / biodétérioration* et *Opération 8*. Le même classement en valeurs relatives d'entités supports de l'attribut/valeur inverse la place du premier et du dernier en favorisant les documents courts où l'attribut est récurrent. Si l'on désire en revanche ne voir apparaître que des documents spécifiques sur le domaine, on pourra dans le cas présent forcer la présence de l'attribut [Matière réduite], ce qui en fonction de la configuration du dispositif utilisé revient à forcer la présence de la table *Bactéries sulfurogènes* (la présence requise d'une table est aussi un critère possible). Dans ce cas, seul le document intitulé *Vrignaud* est présenté. C'est le seul document proposé qui aborde ce thème. Dans les faits, c'est le seul document de la liste de résultats qui aborde le sujet de la corrosion de manière précise et scientifique. La présence d'une table spécifique au domaine est donc cohérente.

Les critères de présence d'un attribut particulier ou d'une table particulière peuvent être choisis en fonction des zones textuelles. Dans ce cas, les documents seront filtrés et ordonnés en fonction de la présence de zones textuelles où cooccurrent soit des entités supports de l'attribut sélectionné, soit des entités de la table sélectionnée. Dans l'exemple, si l'on opte pour cette option avec l'attribut [Niveau de risque], cela traduit un intérêt particulier pour un document présentant une zone spécifique au domaine où sont potentiellement évoqués des facteurs de risque de la biocorrosion. Dans l'exemple, la liste de documents obtenue est la suivante : *Opération 8*, *Vrignaud* et *Le Laboratoire de biocorrosion / biodétérioration*. En effet, les zones de ces documents qui présentent cet attribut sont respectivement :

- les zones 3 et 6 pour *Opération 8* où les récurrences sont respectivement supportées par 3(2) entités et 2(2) entités ;
- les zones 52 et 60 pour *Vrignaud* où les récurrences sont respectivement supportées par 2(2) entités (*bactérie-sulfuro-oxydante* et *béton*) et 2(1) entités (*bactérie-sulfuro-oxydante*) ;
- la zone 3 pour *Le Laboratoire de biocorrosion / biodétérioration* où les récurrences sont supportées par 2(2) entités.

D'autres critères de filtrage et d'ordonnancement sont possibles. L'évaluation de leur intérêt n'est pas facile car ils apparaissent dépendants du dispositif initial. Par exemple, la simple présence de récurrences de l'attribut [Niveau de risque] permet de pressentir la présence de facteurs de risque du domaine. Vu les valeurs de l'attribut, ces facteurs sont considérés au minimum comme dangereux. De manière générale, si l'on s'en tient au protocole de construction de dispositif proposé, les tables rassemblant les entités les plus spécifiques au domaine apparaissent sans sous-catégorisation, c'est-à-dire en bas des arbres de la forêt d'arbres que représente le dispositif. Ainsi, le critère qui tend à favoriser la présence d'une table ou d'un attribut très spécifique permet d'obtenir des documents classés et filtrés en fonction de la présence d'entités spécifiques au domaine et donc susceptible d'aborder le do-

maine de façon précise, voire à l'aune d'une thématique particulière du domaine. En revanche, la présence d'un attribut très générique (type [Nature du phénomène] dans l'exemple), tend à favoriser la présence de documents simplement en rapport avec le domaine tel qu'il a été décrit, ce qui peut être intéressant pour un premier filtrage si l'on n'utilise qu'une entité générale pour la requête et qu'on laisse le logiciel procéder aux analyses sur l'ensemble des résultats.

5.4.2.3 Discussion

L'expérience décrite dans la partie précédente a pour but de montrer la faisabilité et l'intérêt des analyses automatiques de documents à partir d'un ou plusieurs dispositifs. LUCIASearch est un logiciel d'étude qui permet une analyse en terme de contenu de documents. Il permet de pallier les manques d'autres systèmes documentaires (tout en profitant de leurs index). Par-là même, les analyses proposées permettent un accès personnalisé aux documents. L'exploration assistée de leur contenu est supportée par les interfaces de lecture créées.

La requête proposée en exemple a été utilisée pour simuler l'intérêt particulier d'un utilisateur pour un aspect précis du domaine décrit dans le dispositif proposé en exemple. Le critère de filtrage le plus adéquat pour la tâche simulée est celui qui force la présence de la table [Facteurs de risque très dangereux]. Dans ce cas, à partir des 7 documents de départ analysables proposés par les moteurs de recherche, seuls 3 sont proposés. Ils sont classés dans l'ordre suivant :

1 : *Vrignaud* où une récurrence de l'attribut est supportée par 14 (3) entités sur l'ensemble du document.

2 : *Le Laboratoire de biocorrosion / biodétérioration* où une récurrence de l'attribut est supportée par 4 (4) entités sur l'ensemble du document.

3 : *Opération 8* où une récurrence de l'attribut est supportée par 3 (2) entités sur l'ensemble du document.

À la lecture de ces documents, on peut voir que *Vrignaud* présente de façon très détaillée le mécanisme de détérioration des canalisations. Le document aborde effectivement le sujet de la biodétérioration du béton et traite du phénomène de biocorrosion. Le second document de la liste présente spécifiquement des problèmes de biocorrosion dont celui du béton. C'est un document court à caractère général. La présence de nombreuses photos et de courtes zones de texte peut être repérée à l'aide d'une présentation schématisée. Cette présentation peut éviter la lecture du document si la tâche nécessite d'obtenir des documents précis et non une présentation générale du domaine. Le dernier document de la liste présente une opération de recherche d'un laboratoire de physique des liquides et d'électrochimie. Cette opération aborde effectivement le problème de la biocorrosion mais la détérioration du béton n'apparaît pas comme un sujet de recherche à part entière. Ce document est particuliè-

rement intéressant pour montrer l'intérêt des analyses. Une seule occurrence de l'attribut valeur [Nature du facteur : vivant] apparaît : cette valeur ne prend part à aucune récurrence. À la lecture, on remarque que la corrosion de nature biologique (la biocorrosion) n'est qu'un thème de recherche du laboratoire parmi d'autres et qu'il n'est pas abordé en tant que tel dans le texte. De même, une seule récurrence de l'attribut/valeur [Nature du phénomène : facteur] apparaît et elle est étayée dans les deux zones textuelles (zone 2 et 4) par deux occurrences de la même entité lexicale. Cette récurrence est donc moins significative que si elle était appuyée par deux entités distinctes alors même que l'attribut [Nature du phénomène] est hérité par de nombreux termes généraux et spécifiques au domaine dans le dispositif.

Certains problèmes apparaissent à l'issue de cette expérience. Si les analyses automatiques à partir de données personnalisées montrent leur intérêt, certains problèmes inhérents au traitement de documents de sources hétérogènes perdurent. D'une part, l'impossibilité de traiter d'autres formats que le HTML est une limite importante à l'utilisation de LUCIASearch dans des conditions autres qu'expérimentales. D'autre part, l'exemple a été choisi à dessein pour montrer la difficulté de traiter de documents de tailles très différentes. L'utilisation des pourcentages par rapport aux nombres d'entités lexicales non vides des documents ne permet pas de résoudre totalement la relativisation de l'importance des thèmes représentés par les tables ou les récurrences d'attributs/valeurs. Les principes de sémantique légère montrent ici une limite car plus les documents sont longs, plus ils sont susceptibles de receler des entités non présentes dans les dispositifs et ce fait n'est pas pris en compte pour l'instant dans le calcul des résultats.

L'exemple étudié permet de pointer l'aspect itératif de l'utilisation du système. À partir de l'examen des résultats obtenus, on peut envisager un retour à la fois quantitatif et qualitatif sur les données de départ. D'un point de vue quantitatif, de nouvelles entités lexicales de documents jugés pertinents peuvent être ajoutées au dispositif de départ. D'un point de vue général au domaine, la lecture de *Vrignaud* fait apparaître par exemple de nombreuses occurrences de termes comme *corrosivité* ou *oxydation* qui semblent en rapport avec le domaine du dispositif et peuvent donc faire l'objet d'une catégorisation. De même, d'un point de vue plus spécifique au domaine, des termes qui ont trait à des mécanismes de la corrosion comme *oxydo-réduction* ou *enduit biologique* (utilisé indifféremment pour *biofilm*) pourraient être ajoutés. D'un point de vue qualitatif, l'ajout de ces entités pourrait amener à modifier la structure du dispositif de départ. De plus, des termes comme *matériau* et *corrosion* apparaissent finalement comme trop généraux et pourraient être supprimés de la table [Phénomènes de la biocorrosion]. Enfin, la biocorrosion en milieu maritime n'apparaît pas comme un thème général du domaine et pourrait faire l'objet d'une catégorisation particulière de termes comme *eau de mer* et *micro-algue*.

Le choix des critères de filtrage et d'ordonnancement n'est pas trivial. Cependant, la technique d'annotation des documents permet de générer plusieurs classements consécutifs pour choisir les plus adaptés à la tâche. Une fois déterminés, ces critères pourront être réutilisés pour une recherche analogue. Les techniques de filtrage qui ne dépendent pas de critères formulés par l'usager sont appréciables. L'examen manuel des pages *BTP-Génie Civil* et *Matériaux – Chimie* permet d'apprécier qu'elles sont effectivement très majoritairement composées d'hyperliens et d'apprécier la validité du choix automatique dans ce cas présent. Le document de l'université de Toulouse (intitulé *Université Paul Sabatier*) présente une liste de mots-clefs associés à des hyperliens censés aboutir aux sites de laboratoires de recherche concernés par le sujet ainsi représenté. Si *biocorrosion* fait partie de cette liste, l'hyperlien correspondant aboutit finalement à une page générée dynamiquement qui indique qu' « *aucun laboratoire ne correspond à ce critère de recherche* »[136]. Le cas des deux autres documents est plus discutable. Le document *BTP-Génie Civil* propose ainsi un lien intitulé *Production de béton à partir de résidus plastiques* qui aboutit à une brève scientifique qui n'aborde par le problème de la corrosion de ce matériau. Le même document ainsi que celui intitulé *Matériaux – Chimie* proposent un autre lien intitulé *Entreprise – L'avenir très prometteur d'une science encore assez méconnue : la biocorrosion* qui aboutit à un document fournissant une présentation générale de CORRODYS. Dans ce dernier document, il n'est pas fait mention de la corrosion du béton. Ces remarques ne remettent pas en cause l'intérêt du système LUCIA mais limitent l'utilisation de LUCIA-Search dans sa version actuelle. Il est envisageable d'analyser les textes des zones considérées comme des zones de navigation et de parcourir le graphe web à partir du point représenté par un hyperlien cohérent avec les ressources du dispositif utilisé.

Notre première présentation à la société CORRODYS a permis de voir quels étaient les points forts de nos propositions :

- l'analyse de documents non formatés ;
- la présentation interactive des résultats d'analyse tant du point de vue des ensembles de documents que pour les documents et leur contenu ;
- l'utilisation de données personnalisables et modifiables.

Les points faibles concernent essentiellement le fait que LUCIASearch est un logiciel d'étude et non pas un logiciel utilisable tel quel dans un cadre professionnel. Nous revenons sur ce point dans la partie suivante.

[136] http://www.ups-tlse.fr/RECHERCHE/acces_labo_par_motcle.php3?mot=96

5.4.3 Conclusions et perspectives sur le projet de veille documentaire

Une expérience dans le champ de la veille documentaire a permis de montrer comment, à partir de ressources structurées de façon différentielle, LUCIA peut assister un utilisateur pour l'accès à des documents inconnus.

Comme nous l'avons vu dans cette partie, la multiplicité des critères possibles de filtrage et d'ordonnancement et l'incidence de la structuration des ressources sur les résultats obtenus des logiciels ne permettent pas de considérer les résultats obtenus automatiquement comme suffisants en soit - au moins lors des premières utilisations dans un processus itératif. C'est pour cela que les interfaces de parcours rapide d'ensembles documentaires et d'assistance à l'exploration du contenu des documents apparaissent indispensables dans ce cadre. Plus que le seul travail des logiciels, c'est l'interaction avec l'utilisateur qui permet d'obtenir des résultats satisfaisants. Le repérage de zones coloriées dans les documents ou dans leur représentation schématique permet par exemple d'accéder rapidement aux parties de documents potentiellement les plus intéressantes en fonction des désirs de l'usager.

Les adaptations pour une utilisation professionnelle de nos propositions sont multiples. Nos contacts nous ont amené à envisager les modifications nécessaires pour l'élaboration d'un logiciel d'emploi (en opposition à un logiciel d'étude[137]) à partir des existants. Ces modifications concernent tout d'abord l'intégration de tous les logiciels d'étude proposés dans une même plate-forme, de THEMEEDITOR à LUCIASearch en passant par LUCIABUILDER. Ces logiciels devront en outre subir des adaptations pour qu'ils soient utilisables dans des conditions réelles (rédaction d'une aide avec des professionnels, modifications techniques pour uniformiser les standards de sauvegarde, etc.). Spécifiquement à LUCIASearch, ces adaptions concernent également les formats analysables : le format PDF est par exemple envisageable car il existe déjà des solutions pour des conversions de PDF vers TXT (qui supportent difficilement entre autres les mises en page en colonnes). D'autres concernent la présentation des données. LUCIASearch ne propose pas à l'heure actuelle d'utilisation des dispositifs à partir de leur présentation en tables, c'est-à-dire à partir de la présentation utilisée pour les construire en interactions avec les logiciels. Il faut remédier à cet inconvénient pour parvenir à préserver des vues cohérentes sur les données pour un plus grand confort des utilisateurs. LUCIASearch peut être adapté pour ne plus être seulement un métamoteur mais également être utilisable pour analyser des documents déjà en possession de l'utilisateur et pour procéder à des recherches précises. Une fois les documents annotés, les résultats d'analyse pourraient être conservés et constituer ainsi une sorte d'index personnel en rapport avec un domaine précis. D'autres techniques pourraient parfaire le fonctionnement itératif du système. Par exemple, l'ajout d'entités lexicales dans les dispositifs pourrait être facilité par une interface dédiée (basée par exemple sur la représentation en groupes d'attributs/valeurs

[137] Au sens de [Nicolle, 1996], c'est-à-dire un logiciel conçu dans le but de vérifier des hypothèses (sur les langues) en les expérimentant sur du matériau textuel attesté.

des dispositifs). Le repérage des entités potentiellement intéressantes pourrait également faire l'objet de calcul sur la base de techniques simples du type `tf.idf` [Salton et MacGill, 1983*] ou celle proposée dans [Vergne, 2003*]. Ces techniques seraient bien entendues utilisées toujours dans une optique d'assistance.

5.5 Évaluation

Il est temps d'aborder l'évaluation de nos travaux et pour ce faire, les techniques d'évaluation des modèles de TAL centrés sur l'utilisateur. Si les parties précédentes ont montré l'intérêt et la faisabilité de nos techniques pour l'accès personnalisé aux documents et l'assistance à leur contenu, l'évaluation de nos travaux est perfectible et doit se poursuivre.

Pour le projet ISOMETA, de la nature-même du fait de langue étudié et en l'absence de consensus autour de sa définition, une évaluation quantitative pourrait passer par le pré-traitement manuel de corpus judicieusement sélectionnés. Ce type d'activité nécessite un investissement en temps très important. Si le temps nous a justement manqué pour le réaliser, nous avons entamé un projet en ce sens dans le cadre du pôle MODESCOS (Modélisation en Sciences Cognitives et Sociales) de la Maison de la Recherche en Sciences Humaines de l'Université de Caen-Basse Normandie. Ce projet vise l'annotation par plusieurs utilisateurs des emplois métaphoriques correspondants aux domaines décrits dans le corpus utilisé. Il devra permettre la mise au point de règles de distinction entre des emplois métaphoriques et des emplois non métaphoriques d'entités lexicales du domaine source. L'aspect centré-utilisateur du modèle, fut il utilisé par des experts, nécessite une prise en considération de la dimension subjective de certaines facettes de l'interprétation que l'on peut faire d'un emploi métaphorique. Pour le projet ISOMETA, le consensus était fondé sur l'avis concordant des trois membres. La mise en place d'expériences avec d'autres utilisateurs permettra également d'évaluer ce consensus.

Dans le cadre de la veille documentaire (5.4), des mesures existent pour évaluer les systèmes. Mais comme il est souligné dans [Loupy, 2000*], il est très difficile d'effectuer des mesures ayant pour but une évaluation dès qu'un facteur mis en jeu dans un système est soumis au jugement humain. Or, notre système est justement centré sur l'individu. Comment dès lors prétendre à une évaluation ?

Les méthodes classiques d'évaluation issues de la problématique de la recherche d'information visent la quantification de la qualité des résultats produits par un système. Les taux de *rappel* et de *précision* sont les plus couramment employés. Le *rappel* est la proportion de bons résultats fournis par le programme par rapport aux bons résultats qu'il aurait dû idéalement fournir. La *précision* est la proportion de résultats corrects parmi l'ensemble des réponses fournies par le programme. Deux mesures duales découlent de ces deux taux : le *bruit,* qui est la proportion de fausses réponses parmi l'ensemble des résultats fournis, et le *silence,* qui est la proportion de bonnes réponses qui au-

raient dû être proposées par le système et qui ne l'étaient pas. Ces taux permettent des évaluations comparatives entre systèmes, en particulier au cours de campagnes d'évaluation. Ces taux ne peuvent être calculés que relativement à un travail effectué à la main par des experts. Or, nous avons déjà souligné la difficulté de mettre d'accord des experts, simples être humains, sur un problème particulier (chapitre 4, p.129). Le calcul de ces taux suppose que l'on sache combien de bonnes réponses sont possibles. Dans ces conditions, il ne s'agit toujours que d'une approximation de la qualité d'un système calculée par rapport aux résultats d'un ou plusieurs agents humains. La généricité de ces taux est ainsi contestable. En premier lieu, ils nécessitent la constitution d'un corpus de test. Dans une problématique de veille documentaire comme celle proposée avec LUCIASearch, il est impossible de calculer de façon incontestable le rappel des résultats proposés car personne ne peut établir en dehors de l'utilisation même du système sur un échantillon de documents combien il en existe précisément qui sont intéressants pour une tâche documentaire donnée. La constitution d'un corpus de test doit en outre correspondre aux conditions réelles d'utilisation des systèmes pour lesquels ils doivent servir. Ainsi dans [Lavenus et Lapalme, 2002], les auteurs montrent la différence entre les questions/réponses du corpus de référence des conférences TREC (dédiées à l'évaluation des systèmes de *Q/A*) par rapport à de véritables questions d'utilisateurs en recherche documentaire. Comme il est montré [Spark Jones, 2001], une évaluation en laboratoire est toujours fondée sur des présupposées concernant le contexte de la tâche. Selon notre point de vue, pour mieux considérer la tâche, il faut remettre l'humain, l'usager potentiel, au centre de l'interaction avec les systèmes. Cela implique que ces derniers ne soient plus évalués exclusivement de façon comparative, mais aussi individuellement pour éviter au minimum d'avoir des corpus de test en décalage avec les visées applicatives.

Dans la plupart des champs de recherche de l'informatique, l'évaluation des programmes est un souci premier car elle conditionne leur utilisation. Mais les systèmes informatiques, et d'autant plus ceux dédiés au TAL, ne se réduisent pas à un ensemble de programmes : d'autres aspects doivent être considérés pour leur évaluation globale. Une dimension supplémentaire est ainsi importante à prendre en compte : le temps. Évaluer un système à un instant donné ne garantit pas une utilisation future, en particulier lorsque l'on ne peut pas prédire les évolutions possibles des données à traiter. Nous avons montré que les langues évoluaient constamment : les mots changent en forme et en sens dans et par leur usage. Comme le propose Beust [Beust, 2004], des évaluations de systèmes tels que LUCIA doivent donc être menées de façon *syncho-diachronique* car c'est ainsi que vivent les langues. La propriété des langues de se décrire sur les plans synchroniques et diachronique a été mise en évidence par Saussure. C'est parce que la réflexivité est première dans les langues vivantes que leur usage amène à les modifier elles-mêmes, passant sans cesse d'un état synchronique à l'autre. Partant de ce constat, Beust [*ibid.*] affirme même qu' « *en ne tenant pas compte de cette réflexivité au centre de la sémiotique des interactions langagières, les évaluations classiques en TAL ne font finalement pas la différence entre langues vivantes et langues mortes ».* La dimension temporelle des données n'est pas la

seule à devoir être prise en considération, celle du système doit l'être également. Si les évaluations classiques s'effectuent sur des logiciels finalisés, fournis « clefs en main », il n'est pas prévu d'évaluer la capacité de l'utilisateur à utiliser de mieux en mieux le logiciel et celle du logiciel à contenir des données de mieux en mieux structurées ou décrites par l'utilisateur. Or, cette approche avec amorçage est justement un des points originaux de nos propositions. Nous avons montré, à la fois pour les travaux sur la métaphore et ceux sur la veille documentaire, combien le cycle itératif d'utilisation pouvait amener des modifications bénéfiques sur les résultats obtenus.

Notre démarche a été d'intégrer à mi-parcours une première évaluation relative à l'utilisabilité de notre modèle à travers l'expérience décrite dans le chapitre 3 (partie 3.4). Les résultats présentés dans ce chapitre concernent la démonstration de l'intérêt et de la faisabilité de nos propositions. L'évaluation du système en lui-même ne saurait passer que par des expériences en conditions réelles. Dans le cadre de la veille documentaire, nous espérons que nos collaborations avec des acteurs professionnels du domaine nous apporteront des solutions. Nous envisageons une démarche d'évaluation expérimentale comme celles pratiquées dans certaines sciences humaines avec des expérimentations avec plusieurs sujets, le recueil des résultats et des entretiens avec les sujets et enfin l'analyse du matériau ainsi constitué.

Pour clore ce document, nous proposons dans la partie suivante un bilan des travaux effectués. Nous replaçons nos travaux dans le champ plus général des STIC et nous retournons vers le principe de *sémantique légère* pour le TAL que nous avons défendu tout au long de ce tapuscrit.