

Ressources termino-ontologiques différentielles personnelles : construction et projection sur corpus

Thibault Roy et Pierre Beust

Laboratoire GREYC UMR 6072, Equipe ISLanD
& Pôle ModeSCoS (MRSH)
Université de Caen - Basse-Normandie
Boulevard Maréchal Juin, F14032 Caen
troy@info.unicaen.fr, beust@info.unicaen.fr

Résumé

Nous présentons dans cet article un modèle de représentation différentielle de domaines lexicaux et son application à travers un outil d'analyse visant à produire des cartes de corpus en fonction des ressources. Nous expliquons en quoi de telles analyses sur corpus permettent de se prononcer sur l'adéquation des ressources aux textes.

***Mots-clés** : sémantique interprétative, cartographie thématique, intertextualité, ressources centrées utilisateur.*

Abstract

We present in this paper a model of differential representation of terms and its application through a tool of corpus analysis. This tool provides maps from corpus and lexical resources. We present how analyses of corpora allow us to get information on the link between the resources and the analyzed texts.

***Keywords** : interpretative semantic, thematic cartography, intertextuality, user-centered lexical resources.*

1. INTRODUCTION

Une recherche d'information ponctuelle sur Internet peut être réalisée de manière assez efficace en soumettant quelques mots clés bien choisis à

quelques moteurs de recherche. Tout le monde en fait quotidiennement l'expérience. Il n'en est pas de même pour des objectifs plus complexes dans le domaine de la veille documentaire (résumé de textes, analyse thématique, navigation dans des espaces documentaires, etc.). Là, les spécificités socio-linguistiques des utilisateurs (par exemple leurs centres d'intérêt, leurs habitudes terminologiques) et leurs interprétations sont au centre de la problématique. Elles nécessitent d'une part des traitements fins (plus fins en tout cas qu'une indexation massive) et d'autre part d'avoir recours via ces traitements à des ressources terminologiques bien adaptées. Se pose alors les questions de trouver (et donc de ré-exploiter) ou de construire ces ressources ainsi que des les maintenir en complète adéquation avec le contexte visé. Par rapport à ces questions, l'originalité de nos travaux est de considérer qu'il faut avant tout donner une place prépondérante à l'utilisateur. Pour nous, c'est le plus souvent l'utilisateur qui est le seul à même de décrire les ressources dont il a besoin et c'est encore lui seul qui peut superviser l'évolution de ses ressources au fur et à mesure de leur utilisation. Cela ne veut pas dire pour autant qu'il n'y a pas de réutilisabilité possible d'une ressource entre différents utilisateurs, seulement cette réutilisabilité provient d'échanges entre des utilisateurs travaillant ensemble à un moment donné sur une même tâche. La question de la généricité d'une ressource n'est donc pas de notre point de vue un préalable, c'est éventuellement une conséquence. Nous allons montrer dans cet article quels outils nous proposons pour assister un utilisateur ou un groupe d'utilisateurs dans la construction et la maintenance de leurs propres ressources terminologiques.

Dans la suite nous adoptons le plan suivant. Après avoir introduit les principes de notre approche centrée utilisateur, nous présentons un modèle de structuration différentielle de terminologie, le modèle LUCIA. Nous expliquons alors quelles analyses nous mettons en œuvre à partir des ressources développées selon ce modèle et comment elles permettent de fournir en retour des indices sur la qualité de ces ressources.

2. L'APPROCHE CENTRÉE UTILISATEUR

Notre travail s'inscrit dans le cadre de recherches en Traitement Automatique des Langues (TAL) et en linguistique de corpus appliquées à la veille documentaire. Plus précisément, parmi les quatre familles de méthodes d'accès au contenu des documents (l'extraction d'information, les méthodes de question/réponse, le résumé automatique et l'aide à la navigation) décrit par A. Nazarenko dans (Condamines et *al.*, 2005, Chap. 6), nous nous intéressons plus particulièrement aux aides à la navigation.

D'un point de vue applicatif nos recherches visent l'aide à la lecture rapide (visualisation de corpus et de documents par exemple) et le regroupement en classes de documents thématiquement proches.

En matière d'ingénierie documentaire la tendance actuelle est de faire du Web une vaste base de connaissances pour un plus grand nombre d'utilisateurs. C'est la démarche considérée dans le projet du Web Sémantique où l'objectif annoncé par T. Berners-Lee (Berners-Lee, 1998), initiateur du projet et directeur du W3C, est d'enrichir (notamment au moyen des technologies développées autour du langage XML) les documents (à l'aide d'ontologies normalisées, soit automatiquement, soit en assistant leur auteurs) avec des informations sur leur propre sémantique qui soit directement interprétables par des agents logiciels sans la supervision d'une interprétation humaine. Ceci fait l'hypothèse que la valeur sémantique d'un passage de document est le fait de son auteur alors que c'est finalement bien plus celui de son lecteur.

Notre approche de la veille documentaire se situe à l'opposé de celles défendues dans le cadre du Web Sémantique. Elle s'en distingue essentiellement par le fait que nous mettons l'accent sur des traitements et des Ressources Termino-Ontologiques (RTO) avant tout centrés autour de leur utilisateur, de sa tâche, de ses besoins et de ses centres d'intérêt. Là où le Web Sémantique cherche à rendre le plus possible partagées de vastes ontologies qui normalisent et synthétisent une connaissance pensée comme objective et devant convenir à tous les utilisateurs, nous préférons manipuler des ressources propres à un utilisateur ou un petit groupe d'utilisateurs. Il en découle une certaine *légèreté* de ces ressources, au sens de (Perlerin, 2004), dans la mesure où elles ne représentent que ce qui est pertinent du point de vue de l'utilisateur et restent ainsi de taille raisonnable (par exemple une centaine de termes) ce qui les rend moins complexes à construire, à maintenir et à enrichir.

Cette approche centrée utilisateur conduit donc à opérer un certain renversement scientifique relativement aux ressources utilisées en TAL. Premièrement, d'un point de vue très pratique, force est de constater que des ressources très généralistes, valables pour tout type de traitement envisagé ainsi qu'à destination de tout utilisateur potentiel, ne sont pas facilement disponibles (sous forme électronique pour des traitements automatiques) et encore moins gratuites. Deuxièmement, nous soutenons que l'idée même d'une ressource généraliste est illusoire car elle dépend inévitablement du contexte qui lui préexiste (le but recherché par le ou les auteurs ainsi que leurs spécificités socioculturelles). Le rapport de l'Action Spécifique 32 du CNRS/STIC en 2003 (Charlet et al., 2003) va également dans ce sens en précisant un obstacle au projet du Web Sémantique.

tique : la détermination et l'ajout, même de simples méta-données, ne sont pas des activités naturelles pour la plupart des personnes. Cela traduit bien la difficulté à produire des ressources les plus objectivées possibles représentant des significations et/ou du sens.

Si les ressources doivent être personnalisées, la démarche centrée utilisateur amène également à défendre que les applications qui utilisent ces ressources doivent aussi être adaptées à leurs utilisateurs. Les traitements sémantiques appliqués à l'accès au contenu des documents ont tout à y gagner à être le plus possible subjectivés, tant du point de vue des ressources que du point de vue des résultats opératoires visés. Par exemple, Google permet de rechercher un mot clé ou un de ses synonymes avec l'opérateur *tilde* ~ (par exemple la requête *powerpoint ~help* effectue une recherche sur *powerpoint* ET *help* ou *tips, faq, tutorial*). Cependant, c'est le moteur lui-même qui établit ses listes de synonymes et il serait peut être plus judicieux que celles-ci soient validées par les utilisateurs quand ils les utilisent. Toujours à propos de Google, on trouve un exemple de résultat assez malheureux de l'opérateur *define* sur le *blog* de J. Véronis. L'opérateur *define* (disponible pour les pages en français depuis avril 2005) sert à rechercher à partir d'un mot des pages Web où ce mot ferait visiblement l'objet d'une définition. L'expérience relatée consiste à rechercher une définition du mot *femme* avec la requête *define:femme*. Les résultats donnés sont très contestables. On aurait donc tort de croire à la fiabilité de l'opérateur *define* (pourtant présenté par Google comme un outil de recherche de définitions sans plus de détails) comme on aurait également tort de considérer le Web dans son ensemble comme une encyclopédie dans lequel on puisse rechercher des définitions attestées pour tous, notamment d'un point de vue moral. Ces exemples illustrent le problème de la variabilité des terminologies pour différents utilisateurs et différentes tâches. La démarche centrée utilisateur nous paraît donc être une bonne réponse au constat que dressent D. Bourigault et N. Aussenac-Gilles (2003) : « ... le constat de la variabilité des terminologies s'impose : étant donné un domaine d'activité, il n'y a pas une terminologie, qui représenterait le savoir du domaine, mais autant de ressources termino-ontologiques que d'applications dans lesquelles ces ressources sont utilisées ».

Suivant ce constat, il nous paraît important de proposer des modèles et des outils pour permettre la création de RTO en fonction des utilisateurs et des applications visées. C'est dans ce but que nous proposons le modèle LUCIA ainsi que différents outils. Nous en rappelons les principes dans la partie suivante.

3. LE MODÈLE LUCIA

Le modèle LUCIA fait intervenir un certain nombre de notions linguistiques car il est fortement inspiré de la Sémantique Interprétative de F. Rastier (1987, 1994, 2001). Nous présentons tout d'abord ces notions puis les principes du modèle ainsi que les conditions de production et d'évolution de ressources LUCIA.

3.1. Notions

L'objectif des RTO LUCIA est de permettre des analyses interprétatives au sens de F. Rastier (1987). Dans ce but, il nous paraît important de clarifier les notions suivantes : lexies, textes, corpus, sèmes et isotopies.

3.1.1. Lexies

Les unités linguistiques manipulées dans les RTO LUCIA sont des lexies. Une lexie est une unité fonctionnelle (que l'on peut considérer également comme une entrée lexicale) regroupant plusieurs morphèmes et qui peut correspondre à plus d'un mot. Deux types de lexies peuvent être distingués : les lexies simples et les lexies complexes selon qu'elles consistent en un mot graphique (comme les lexies « eau » et « marcher ») ou en plusieurs (comme la lexie « pomme de terre »). Dans (Rastier et *al.*, 1994), la lexie est vue comme « *une unité de signification* » et les auteurs la définissent comme une élément de base à toutes analyses sémantiques de textes. Une lexie est représentée dans un texte par une ou plusieurs formes graphiques : ses flexions et éventuellement d'autres graphies. Ainsi, la lexie « *cassette* » peut apparaître en corpus sous plusieurs graphies : « *cassettes* » (flexion plurielle de la lexie) ou encore « *K7* », « *K-7* », etc. La lexie est une unité naturelle (plus que celle de mot dont la définition reste délicate) que l'utilisateur, au centre du processus de création et d'évolution des RTO selon une tâche donnée, peut décrire.

3.1.2. Textes et corpus

La notion de texte que nous adoptons est issue de (Rastier, 2001). L'auteur donne d'abord des définitions négatives schématisant les erreurs souvent commises lors de la définition du concept de texte. Tout d'abord, un texte ne doit pas être considéré comme une chaîne de caractères (seulement une infime partie des chaînes de caractères sont des textes). Un texte ne doit pas être non plus considéré comme une suite d'instructions (à la manière d'un programme informatique), cette considération rame-

nant la compréhension d'un texte à l'exécution d'un programme par un ordinateur. Enfin, un texte n'est pas non plus une suite de schémas cognitifs, la lecture d'un texte suscite généralement la création de schémas mentaux mais elle ne se limite pas à cela. F. Rastier donne ensuite une définition positive du texte : « *Un texte est une suite linguistique empirique attestée, produite dans une pratique sociale déterminée, et fixée sur un support quelconque.* ». Un texte n'est donc pas une création artificielle (comme pourrait l'être un exemple linguistique construit pour illustrer un fait de langue), il est créé dans le cadre d'une pratique sociale et il prend place sur un support (feuille de papier, fichier informatique, etc.)¹.

F. Rastier (2005) propose de considérer le texte comme une unité minimale d'une linguistique « évoluée ». Selon lui, cette unité prend son sens dans un contexte plus global, celui du corpus. L'auteur définit alors un corpus comme « *un regroupement structuré de textes intégraux, documentés, éventuellement enrichis par des étiquetages, et rassemblés : (i) de manière théorique réflexive en tenant compte des discours et des genres, et (ii) de manière pratique en vue d'une gamme d'applications* ». Un corpus doit être défini selon l'application pour laquelle il est destiné. Cette application détermine le choix même des textes structurant le corpus. Ainsi, toute analyse textuelle doit selon nous se baser sur des textes provenant de corpus construits par les utilisateurs en vue de tâches précises. Quand des ressources sont extraites de corpus, il en résulte qu'elles sont donc aussi fortement liées à l'utilisateur et à la tâche.

3.1.3. Sèmes et isotopies

Le sème est défini comme la plus petite unité de signification (Rastier, 1987), le plus souvent exprimé sous la forme d'un trait (entre deux « / »). Par exemple, la lexie « *chien* » pourra porter les sèmes /mammifère/, /possède des crocs/, /animal domestique/, etc.

Une isotopie est par définition (Rastier, 1987) un « *effet de la récurrence d'un même sème* ». Cet effet de récurrence permet alors de caractériser et d'identifier l'importance de certains éléments de signification (des sèmes) dans une phrase, un texte. Par « défaut », une isotopie est intratextuelle, c'est-à-dire que l'effet de répétition de sèmes se fait sur une entité textuelle dont la couverture la plus large est le texte. Dans l'énoncé « *le facteur m'a donné une lettre* » le sème /courrier/ est actualisé dans le contenu de « *lettre* » parce qu'il se répète dans le contenu de « *facteur* »

¹ Il en est de même des ressources utilisées en TAL car elles aussi prennent place sur un support (électronique) et sont créées dans le cadre d'une pratique sociale (dans le cadre d'une démarche de représentation de connaissances objectivées pour les ontologies).

formant ainsi une isotopie intra-textuelle. Cette actualisation permet de retenir la signification pertinente de « *lettre* » dans l'énoncé (on ne retient donc pas, par exemple, la signification de « *lettre* » en tant que caractère de l'alphabet). Quand la récurrence dépasse le palier du texte, on parle d'isotopie inter-textuelle, définie de la façon suivante par T. Thlavitis (1998) : « [une isotopie intertextuelle est constituée par] *la récurrence de traits sémantiques qui caractérisent des textes entiers mais au sein d'un intertexte* ». Dans nos travaux, l'intertexte est le corpus d'étude.

3.2. RTO LUCIA : principes et exemples

Le modèle LUCIA (*Located User-Centered Interpretative Analyser*) (Perlerin, 2004), d'inspiration saussurienne, est un modèle différentiel qui part du constat que pour désigner les choses dont on veut parler, pour établir leur valeur sémiotique, on les décrit juste assez pour les différencier des choses avec lesquelles elles pourraient être confondues dans la tâche. Le principe du modèle est d'exprimer des connaissances et un point de vue sur la terminologie d'un domaine en organisant des lexies selon deux principaux critères :

- des regroupements par similarité, témoignant de la proximité de certaines lexies ;
- des oppositions locales, précisant les différences entre lexies proches.

Le modèle LUCIA a pour objectif de permettre à un utilisateur de structurer et de décrire le lexique d'un domaine de son choix. Dans la pratique, l'utilisateur peut définir un *dispositif* pour chaque domaine qu'il souhaite représenter. Un dispositif contient un ensemble de *tables* en relation, chaque table contenant des lexies d'une même catégorie sémantique selon le point de vue de l'utilisateur. Au sein de chaque table, l'utilisateur doit expliciter des différences entre unités lexicales à l'aide de couples *attributs / valeurs*, qui pour nous sont l'expression de sèmes.

Afin d'avoir dès à présent une vision globale et concrète sur les principes fondamentaux du modèle LUCIA et sur la façon dont une RTO respectant ce modèle est construite, nous proposons un exemple de dispositif LUCIA (également appelé RTO LUCIA). L'exemple présenté ici a été produit lors d'une collaboration avec la société CORRODYS² proposant des services dans le traitement de la biocorrosion (corrosion de matériaux par des organismes vivants). Cette collaboration a été initiée entre des membres de notre équipe de recherche et cet organisme en juillet 2004

² Entité mixte créée par l'Université de Caen et le Centre Régional d'Innovation et de Transfert de Technologie de Basse-Normandie Cotentin : <http://www.corrodys.com/>

afin de réaliser un travail de veille documentaire sur Internet dans le domaine de la biocorrosion. Ce travail a amené à développer un dispositif LUCIA représentant un tel domaine du point de vue des experts. Des projections de ce dispositif sur des ensembles de pages Web ont ainsi permis d'obtenir une valeur ajoutée dans des tâches de recherche et de veille documentaire (projections non détaillées dans cet article, se reporter à (Perlerin, 2004, p.226) pour plus de détails).

Le premier travail pour construire une RTO LUCIA décrivant le domaine de la biocorrosion a été de rassembler des lexies s'y rapportant selon le point de vue des experts. Pour cela, une analyse des graphies répétées dans des documents spécialisés nous a permis de mettre en évidence avec l'aide d'experts des lexies typiques du domaine, comme l'illustre la liste suivante : eau de mer, eau douce, graisse, métal, béton, acier, structure métallique, micro-organisme, micro-organique, micro-algue, moisissure, champignon, mycologique, bactérie, bactérien, bactériologique, etc. À partir de ces lexies, des tables LUCIA ont été construites pour les regrouper et les différencier. Dans cette étude, nous avons été amené à construire, entre autres, les deux tables suivantes :

Facteurs de risque	Micro-organismes
eau de mer, eau douce, graisse, métal, béton, acier, structure métallique, micro-organisme, micro-organique	micro-algue, moisissure, champignon, mycologique, bactérie, bactérien, bactériologique

Le modèle LUCIA propose ensuite de caractériser chacune des tables à l'aide d'un ou plusieurs attributs et de différencier les lexies au sein d'une même table à l'aide de valeurs d'attributs opposées. Ainsi dans la table précédente Facteurs de risque, les experts ont choisi de faire intervenir un attribut Nature du facteur (attribut qui sera donc possédé par toutes les lexies de cette table) avec des valeurs opposées de cet attribut Milieu / Matériau / Vivant. Pour la table Micro-organismes, deux attributs ont été choisis. Les deux tables deviennent alors :

Facteurs de risque	Nature du facteur	
eau de mer, eau douce, graisse	<i>Milieu</i>	
métal, béton, composite, acier, structure métallique	<i>Matériau</i>	
micro-organisme, micro-organique	<i>Vivant</i>	

Micro-organismes	Fonctionnement	Type cellule
micro-algue	<i>Photosynthèse</i>	<i>Eucaryote</i>
	<i>Photosynthèse</i>	<i>Procaryote</i>
moisissure, champignon, mycologique	<i>Pas de photosynthèse</i>	<i>Eucaryote</i>
bactérie, bactérien, bactériologique	<i>Pas de photosynthèse</i>	<i>Procaryote</i>

Une fois l'ensemble des tables créées, le modèle propose de les regrouper et de les lier entre elles au sein d'un même dispositif³. Les liens entre tables sont appelés des liens d'héritage. Dans l'exemple présenté ici, nous avons considéré que la table Micro-organismes est reliée à la ligne Nature du facteur : vivant de la table Facteurs de risque. De cette manière, nous considérons que chaque lexie de la table Micro-organismes hérite du couple attribut/valeur Nature du facteur : vivant.

3.3. Production et évolution des RTO LUCIA

La construction de RTO LUCIA a fait l'objet d'une expérimentation et de développements logiciels que nous abordons dans cette partie. Il en ressort un lien particulier entre de telles RTO et les corpus utilisés.

3.3.1. Expérimentation

L'atelier formation du CNRS « *Variation, construction et instrumentation du sens* » qui s'est tenu en juillet 2002 (île de Tatihou, Manche) a été pour nous l'occasion de mettre en place une première expérimentation sur LUCIA (Perlerin et al. 2003). Nous souhaitions tester la capacité d'utilisateurs novices à s'approprier les principes généraux du modèle (attributs, tables, dispositifs) en leur demandant de construire dans un temps imparti, un dispositif sur un sujet précis (en l'occurrence la bourse) afin de pouvoir comparer les résultats. Cette expérience s'est déroulée au cours de deux séances de deux heures trente chacune et avec un total de 8 participants d'horizons différentes (linguistique, psychologie, ergonomie, informatique, microbiologie, etc.). Après un exposé introductif sur les principes du modèle, nous avons fourni aux participants une liste de 216 lexies issues du corpus *Le Monde sur CD-ROM*. Cette liste avait été obtenue à partir d'un calcul de type Zipf sur l'ensemble des articles traitant de la bourse et de l'économie de laquelle nous avons enlevé tous les éléments non verbaux et non substantivaux (cette liste contenait par exemple des lexies comme *action*, *back office*, *déévaluation*, *OPA* ou encore *palais Brongniart*). Les consignes données aux participants se bornaient à leur demander de construire sur papier un dispositif selon leur façon propre de parler du domaine (la consigne n'imposait pas nécessairement d'intégrer les 216 lexies dans le dispositif).

A l'issue des deux séances d'expérience, tous les participants ont au moins proposé des groupes de lexies, précisé les différences qu'ils consi-

³ Ce dispositif, ainsi que ceux présentés dans cet article, sont disponibles à l'adresse suivante : <http://www.info.unicaen.fr/~troy/dispositifs/>.

déraient effectives au sein de ces groupes et créé des tables avec un ou plusieurs attributs. Cependant aucun participant n'a estimé au bout de l'expérience être parvenu à un résultat finalisé. Après entretien avec eux, nous avons pu estimer tout d'abord que l'expérience présentait un certain nombre de biais. Le premier est certainement le temps imparti trop court pour la réalisation du travail demandé. L'absence du corpus d'origine et donc l'impossibilité de revenir sur un texte faisant intervenir les lexies proposées a également été ressentie comme un handicap par les participants. Cette expérience sans corpus, ni outil logiciel à disposition, permettait simplement de tester la faisabilité de la construction de RTO différentielles personnelles et donc d'apprécier la capacité des participants à amorcer un processus de construction cyclique. En l'occurrence, nous avons constaté que la méthode de construction des RTO LUCIA s'acquiert rapidement et que les principes qui la régissent sont facilement assimilables. Les différences et les points communs découverts au sein des travaux rendus par les participants nous encouragent à poursuivre dans notre voie : les utilisateurs intègrent leur propre sensibilité par rapport à un domaine (cette sensibilité pouvant relever d'une méconnaissance totale de ce domaine) tout en se conformant aux usages qu'ils ont pu rencontrer des lexies proposées.

Bien que la tâche de description des significations ne soit pas triviale, comme l'a souligné par exemple C. Kerbrat-Orecchioni⁴, nous avons pu constater que des utilisateurs arrivent à formuler dans un temps raisonnable des représentations terminologiques reflétant leur point de vue sur le domaine proposé. Ceci constituait l'une des hypothèses que nous cherchions à estimer. Durant cette expérimentation et, à travers les différences et les points communs entre les dispositifs et les groupes de mots construits par les participants, reflet de leurs capacités interprétatives, nous avons pu considérer à sa juste valeur la dimension sociale et partagée du langage latent en chaque locuteur qui ne peut pas être absente de ce type de construction (Nicolle et al., 2002, p.61). Ceci a renforcé notre point de vue résolument tourné vers l'utilisateur.

3.3.2. Aides logicielles

Pour réaliser la construction et la maintenance de dispositifs LUCIA, l'utilisateur est assisté de différents outils logiciels⁵. Parmi les outils pro-

⁴ « Un des problèmes majeurs que pose (...) la description des structurations lexicales réside dans le fait qu'elles tiennent à la fois des systèmes diacritiques (non hiérarchiques) et des systèmes taxinomiques (hiérarchiques) » (Kerbrat-Orecchioni, 1988)

⁵ Outils disponibles à l'adresse suivante : <http://www.info.unicaen.fr/island>

posés, citons le logiciel *MemLabor* qui permet une première assistance à l'extraction de lexies à partir d'un corpus. Le principe de cette extraction est que plus une graphie (hors graphies d'un anti-dictionnaire, contenant par exemple des mots grammaticaux) est répétée dans les textes d'un corpus, plus elle est susceptible de pouvoir être associée à l'un des domaines abordés dans le corpus.

Après avoir extrait les lexies caractéristiques du domaine, l'utilisateur peut les structurer et les expliciter selon le modèle LUCIA. Pour cela, nous proposons l'outil interactif *VisualLuciaBuilder* permettant un grand nombre d'opérations de création et de révision de dispositifs.

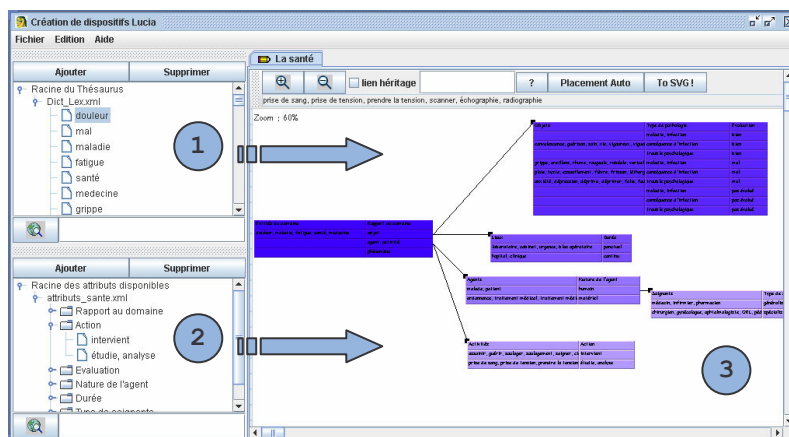


FIGURE 1 – Interface de *VisualLuciaBuilder*.

L'interface de l'outil comporte des zones distinctes marquées de 1 à 3 sur la figure précédente :

La zone 1 contient une ou plusieurs listes de lexies sélectionnées par l'utilisateur dans le cadre d'études sur corpus (par exemple, avec l'outil *MemLabor*). L'utilisateur peut ainsi ajouter de nouvelles lexies, modifier et supprimer des lexies existantes.

La zone 2 représente une ou plusieurs listes d'attributs et de valeurs d'attributs définis par l'utilisateur. Là encore, l'utilisateur peut mettre à jour les attributs d'une liste existante ou ajouter de nouveaux attributs et valeurs d'attributs.

Enfin, la zone 3 est une zone où l'utilisateur « dessine » son dispositif. Il crée de nouvelles tables, les nomme et glisse au sein de chacune d'elles les lexies (zone 1) et les attributs et valeurs d'attributs (zone 2) qu'il

estime décrire les tables concernées. De nombreuses mises à jour et corrections du dispositif sont possibles tout au long de cette phase de création. La création de plusieurs dispositifs en parallèle est également possible, un onglet de la zone de dessin est alors associé à chaque dispositif.

Dès que l'utilisateur considère ses dispositifs stabilisés, du moins momentanément, leur enregistrement sous forme de fichiers XML est réalisé ainsi que des sorties SVG permettant leur visualisation interactive et leur diffusion. Différentes projections sur corpus sont accessibles via cet outil en proposant des coloriages des textes d'un corpus à l'aide des lexies des dispositifs ou encore des représentations cartographiques du corpus prenant en considération ces dispositifs (projections détaillées en partie 4).

3.3.3. RTO LUCIA et corpus numériques

Par rapport à une ressource lexicale généraliste, du type dictionnaire électronique par exemple, une RTO LUCIA présente les deux différences principales suivantes :

1. La non exhaustivité et la subjectivité liée nécessairement à la démarche centrée utilisateur. Il serait illusoire de demander à tout utilisateur, qui n'est pas nécessairement lexicologue, de construire des ressources de large couverture adéquates pour un plus grand nombre. Dans une telle démarche, la généralité de la ressource lexicale et son éventuelle réutilisabilité par d'autres sont des objectifs seconds. Le rapport à la tâche menée par l'utilisateur est bien plus important.
2. L'aspect non nécessairement finalisé de la ressource dans le sens où la construction intégrale d'une ressource LUCIA n'est pas une tâche préalable à son utilisation. Une RTO LUCIA ne s'utilise pas dans une approche où l'on supposerait que ce qui n'est pas exprimé dans la ressource est faux à la manière de l'hypothèse du monde clos en logique. Si un terme est absent à un moment donné d'un dispositif cela ne veut pas dire qu'il n'y sera pas intégré dans une prochaine révision.

Les RTO LUCIA sont donc produites et révisées au fur et à mesure de leur utilisation de manière endogène dans une boucle d'interaction entre un outil logiciel, un utilisateur (ou un petit groupe d'utilisateurs) et des corpus où chaque pôle est déterminant. Il en découle une importance significative des corpus utilisés qui ne peuvent plus être considérés uniquement comme un réservoir de formes attestées sur lequel on tenterait de mettre en œuvre un calcul à base de ressources exogènes. Cette importance des corpus est clairement apparue lors de l'expérimentation avec les utilisateurs présentée précédemment.

Les corpus utilisés dans le cadre de nos recherches sont à la fois à l'origine des ressources lexicales construites et constituent en même temps le matériau d'expérimentation de ces ressources. Ainsi, notre démarche s'inscrit dans des processus de recherche et de développement en aller-retour entre des logiciels d'étude visant principalement des interfaces de lecture rapide d'ensembles documentaires, des corpus d'étude et des utilisateurs (ou des groupes d'utilisateurs), les uns étant conditionnés par les autres. Plutôt que de recourir à des ressources génériques et les plus exhaustives possibles, notre essai est d'exploiter au maximum des ressources légères et non exhaustives car spécifiques au besoin de l'utilisateur qui les crée. Une question qui en découle est celle de la stabilité dans le temps de ces ressources du point de vue de son (ou de ses) auteur(s). Le rapport au corpus est là encore prépondérant. Les outils de projection des ressources sur corpus présentés dans la partie suivante permettent un retour qualitatif sur les ressources. C'est donc en utilisant ses ressources qu'un utilisateur est amené à les mettre à l'épreuve et peut être incité à les rectifier et/ou les compléter. Par exemple, si on s'aperçoit qu'une lexie n'est jamais trouvée en corpus, on peut éventuellement la supprimer de la ressource. À l'inverse, si l'une est très présente avec des significations différentes, on peut vouloir en rendre compte dans la ressource (notamment, en créant une nouvelle table dans le dispositif). La taille raisonnable des ressources rend possible ce genre de révision incrémentale au cours de l'utilisation de la ressource.

Le Web (ou au moins quelques unes de ses parties) peut être assimilé à un « corpus », il est vrai un peu particulier⁶, qui peut aussi être mis à profit pour obtenir un retour qualitatif sur une ressource. C'est un projet que nous sommes en train de mener et dont le but est d'évaluer le rapport qu'entretient une ressource personnelle à une certaine forme d'actualité. Le principe est inspiré d'une expérience relatée sur le *blog* de J. Véronis⁷ à propos des candidatures à l'élection présidentielle de 2007. J. Véronis part de la liste des noms des candidats potentiels et soumet à un moteur de recherche autant de requêtes qu'il y a de couples formés de 2 noms de candidats.

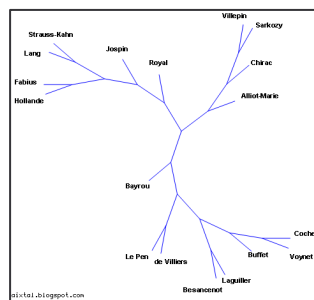


FIGURE 2 : Arbre des prétendants selon J. Veronis

⁶ F. Rastier ne le décrit pas comme un corpus mais comme « une aire de stockage, voire une décharge publique » (Rastier 2005, p. 32).

⁷ <http://aixtal.blogspot.com/2006/04/2007-larbre-des-pretendants.html>

En fonction du nombre de réponses il en déduit une distance entre les candidats 2 à 2. En somme plus il y a de réponses pour 2 noms de candidats, plus les 2 noms sont proches. En procédant à une visualisation reprenant le principe des arbres phylogénétiques des biologistes il visualise les proximités par graphes comme celui de la figure 2. Notre idée est de procéder de façon similaire en interrogeant régulièrement un moteur spécialisé dans l'actualité (par exemple <http://news.google.fr> ou encore un flux de dépêches d'agence de presse) avec tous les couples de termes que constitue un dispositif LUCIA. Le résultat visé serait un retour visuel à destination de l'utilisateur sur les proximités lexicales au sein d'un dispositif et leurs évolutions en fonction de l'actualité du moment.

4. PROJECTIONS DE RTO LUCIA SUR CORPUS

Nous avons insisté précédemment sur le lien très étroit existant entre des RTO LUCIA et des corpus. Pour appréhender ce lien, nous réalisons des projections de telles ressources sur corpus. Afin d'illustrer les différentes projections que nous proposons et leur retour sur les RTO, nous présentons dans cette partie des résultats d'une expérience faisant intervenir des dispositifs LUCIA dans une étude de trois métaphores conceptuelles sur un corpus journalistique.

4.1. Expérience réalisée et RTO utilisées

La métaphore conceptuelle (notée MC) est le phénomène cognitif et linguistique introduit dans (Lakoff et Johnson, 1980) pour qui elle est une projection d'un domaine source sur un domaine cible mettant en saillance certains éléments et en effaçant d'autres. L'étude abordée ici concerne trois MC que nous avons respectivement appelées : la *météorologie boursière*, la *guerre économique* et la *santé financière*⁸. Notre objectif principal à travers cette étude est de mettre en évidence comment se répartissent ces 3 MC dans un corpus en synchro-diachronie. Pour cela, nous avons réalisé différentes projections sur un corpus d'étude constitué d'articles journalistiques boursiers des ressources modélisant les domaines sources des MC étudiées (soit la météorologie, la guerre et la santé). Nous renvoyons à (Roy et al, 2006) pour plus de détails sur l'intérêt des projections en vue d'une caractérisation des trois MC. Nous insistons ici sur les retours de ces projections sur les RTO LUCIA utilisées, les pro-

⁸ Exemple de MC de la *météorologie boursière* extrait du corpus : « *Les monnaies et les marchés sont entrés en turbulence, un climat qui n'est guère favorable aux affaires.* »

jections permettant d'observer quelles ressources ou « parties de ressources⁹ » sont particulièrement utilisées ou au contraire très faiblement présentes dans le corpus étudié. De tels retours permettent alors de maintenir et de faire évoluer les ressources exploitées en tenant ainsi compte du lien très étroit que nous considérons entre corpus et RTO personnalisées.

Pour construire les dispositifs des trois domaines sources des MC étudiées, une exploration du corpus d'étude a été réalisée. La taille des dispositifs ainsi construits varie selon les domaines : 64 lexies réparties en 6 tables pour le dispositif de la guerre, 112 lexies réparties en 8 tables pour celui de la météorologie et 111 lexies réparties en 9 tables pour celui de la santé. Le corpus d'étude est constitué de 303 articles de taille variable (de 200 à 2000 graphies) provenant du journal *Le Monde sur CD-ROM*, tous relatifs au domaine de la bourse. Ce corpus couvre la période 1987 – 1989 et contient un grand nombre des trois MC étudiées en même temps que des emplois thématiques (*i.e.* non métaphoriques) des trois domaines.

Afin de suivre la répartition des RTO LUCIA sur le corpus d'étude, nous considérons les trois paliers textuels suivants :

Le **corpus** : nous proposons des représentations graphiques du corpus (que nous appelons des cartes) mettant en évidence des proximités entre textes et des regroupements de textes proches selon les domaines représentés dans les RTO LUCIA considérées. Ces cartes peuvent aussi bien représenter le corpus dans sa globalité que mettre dynamiquement en évidence l'évolution temporelle des domaines dans les textes d'une période choisie. Des « nuages » et « anti-nuages¹⁰ » du corpus représentant respectivement les lexies des RTO très fréquentes et très peu fréquentes sont également retournés à l'utilisateur.

Le **groupe de textes** : à partir de groupes de textes déterminés automatiquement à partir des RTO LUCIA considérées, nous construisons des rapports d'analyse pour chacun d'eux. Ces rapports font ressortir les particularités du groupe telles des lexies des RTO fréquemment partagées par les textes qu'il rassemble et ses principales isotopies inter-textuelles.

Et le **texte** : les lexies des RTO sont mises en évidence dans chaque texte à l'aide de couleurs. Pour chaque texte, un rapport d'analyse fait

⁹ Par « partie de ressources », nous faisons référence à un sous-ensemble de lexies de RTO LUCIA partageant une ou plusieurs propriétés : lexies présentes au sein d'une même table, partageant un même attribut, une même valeur d'attribut, etc.

¹⁰ Les nuages de mots ont été proposés initialement par le site *TagCloud* (<http://www.tagcloud.com/index.php>) et repris par J. Veronis dans son *blog* (<http://aixtal.blogspot.com/2005/11/blogs-un-nuage-sur-les-banlieues.html>).

ressortir ses particularités, telles les lexies des RTO fréquentes et les isotopies intra-textuelles majoritaires.

4.2. Résultats obtenus

Nous utilisons l’outil ProxiDocs (Roy et Beust, 2004) permettant différents traitements statistiques à partir d’un corpus de textes et de dispositifs LUCIA. Ces traitements conduisent à la construction de cartes thématiques de corpus et de rapports d’analyse.

Les premières projections présentées sont ce que nous appelons des nuages et anti-nuages de lexies, ils mettent respectivement en évidence les lexies des RTO LUCIA les plus fréquentes et les moins fréquentes dans le corpus. Ainsi, la figure 3 présente de tels nuages à partir du dispositif LUCIA décrivant le domaine de la guerre. En partie gauche de la figure, le nuage nous permet d’observer que les lexies *bataille*, *conflit*, *front*, *repli*, *résistance*, *stratégie* et *victoire* sont les plus fréquentes de la RTO dans le corpus. Au contraire, la partie droite, représentant un anti-nuage de lexies, permet d’observer que les lexies *bombarder*, *capitulation*, *char*, *défaite* et *hostilité* sont absentes du corpus.

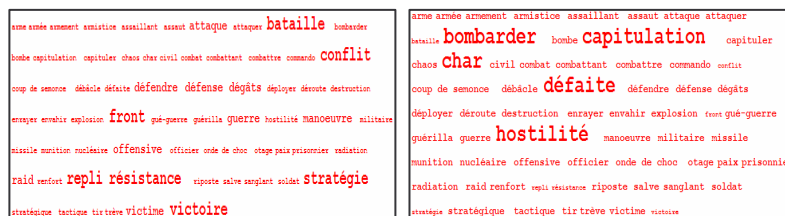


FIGURE 3 – Nuages et Anti-nuages des lexies du domaine de la guerre.

Ces nuages permettent à l'utilisateur de s'interroger sur la trop grande genericité des lexies très fréquentes ainsi que sur la cohérence de la présence dans les RTO des lexies absentes ou très peu présentes en corpus. Les projections suivantes (figure 5) sont des cartes du corpus obtenues après application d'une Analyse en Composantes Principales (ACP) et d'une Catégorisation Hiérarchique Ascendante (CHA) (Bouroche et Saporta, 1980). Le cercle des corrélations (figure 4) obtenu lors de l'ACP nous apporte une information sur la présence et le lien dans le corpus entre les RTO

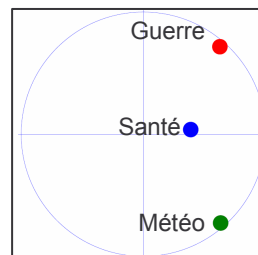


FIGURE 4 – Cercle des corrélations associé à l'ACP réalisée

représentant les domaines étudiés. Selon les particularités de ce cercle, un caractère plus exprimé est attribué aux éléments éloignés de son centre. Nous pouvons ainsi remarquer que les domaines de la météorologie et de la guerre sont fortement exprimés, c'est à dire très présents dans le corpus. Le domaine de la santé est par contre moins exprimé, car plus proche du centre. Sa position sur le cercle, entre les domaines de la guerre et de la météorologie, nous indique un lien entre ce domaine et les deux autres, ce lien se traduisant par des utilisations privilégiées du domaine de la santé avec les domaines de la météorologie et de la guerre. La carte de gauche (cartes des textes) représente chaque texte par un point de couleur correspondant au domaine majoritaire dans le texte (légende accessible en partie inférieure gauche).

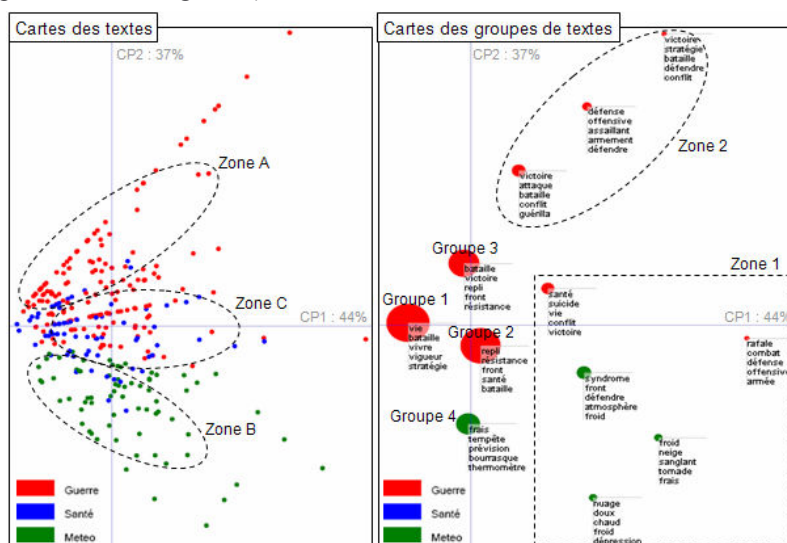


FIGURE 5 – Cartes du corpus construites par la plate-forme ProxiDocs à partir des domaines d'étude. Versions SVG de ces cartes : <http://www.info.unicaen.fr/~troy/isometa/cartes/>

L'utilisation conjointe en corpus du domaine de la santé avec les deux autres domaines se confirme sur la carte des textes : la zone C contenant des textes du domaine majoritaire de la santé est située entre les zones A et B respectivement des domaines majoritaires de la guerre et de la météorologie. Chaque point de cette carte est un lien hypertexte vers un rapport d'analyse contenant le texte où les occurrences des lexies des ressources sont coloriées ainsi qu'un classement par importance décroissante des lexies des RTO et des isotopies intra-textuelles détectées dans le

texte. De tels rapports donnent une idée sur la répartition des ressources au niveau du texte.

La carte de droite (carte des groupes de textes) représente 12 groupes de textes obtenus par CHA. Chaque groupe est représenté sur la carte des groupes par un disque de taille proportionnelle à la cardinalité du groupe. Un disque est colorié selon le domaine majoritaire dans le groupe et est étiqueté par les cinq lexies des RTO les plus occurrentes dans les textes du groupe. Chaque disque est un lien hypertexte vers un rapport d'analyse du groupe contenant un classement des lexies des RTO, ainsi qu'un classement des isotopies inter-textuelles partagées par les textes du groupe. Des groupes et des zones de groupes ont été marqués manuellement sur cette carte. Le tableau suivant illustre leurs particularités.

	Nombre d'articles	Nombre total de lexies de domaine		
		Guerre	Météorologie	Santé
Groupe 1	98	162	79	150
<i>Principales isotopies inter-textuelles</i>	1. Rapport au domaine : objet 2. Évaluation : mal			
Groupe 2	84	231	222	185
<i>Principales isotopies inter-textuelles</i>	1. Rapport au domaine : phénomène 2. Évaluation : mal			
Groupe 3	49	217	37	56
<i>Principales isotopies inter-textuelles</i>	1. Rapport au domaine : activité 2. Nature de l'agent : humain			
Groupe 4	29	18	110	31
<i>Principales isotopies inter-textuelles</i>	1. Rapport au domaine : phénomène 2. Axe : température			
Zone 1 (5 groupes)	25	90	105	54
<i>Principales isotopies inter-textuelles</i>	1. Rapport au domaine : objet 2. Évaluation : bien			
Zone 2 (3 groupes)	14	101	1	8
<i>Principales isotopies inter-textuelles</i>	1. Rapport au domaine : activité 2. Fonction : observation			

FIGURE 6 – Propriétés des groupes et des zones marqués sur la carte.

Ce tableau permet de localiser les emplois des ressources en corpus et met plus particulièrement en évidence les principaux éléments de signification partagés entre les textes d'un même groupe. Ces deux types d'informations présentes dans les rapports des groupes de textes permet-

tent à l'utilisateur d'avoir un regard plus précis sur la répartition de ses RTO en corpus : la navigation inter-textuelle proposée par la carte permet d'accéder directement aux groupes de textes où certaines parties de ressources sont présentes, et ainsi aide à déterminer en quoi une partie de la ressource est spécifique à tel(s) groupe(s) de textes. Les rapports de textes et de groupes de textes accessibles à travers les cartes mettent en évidence des informations les positionnant par rapport à leurs paliers textuels de niveau supérieur. Ainsi, chaque rapport de texte, en plus de contenir des informations relatives à la répartition des RTO dans le texte, contient des informations portant sur la répartition des ressources dans le groupe contenant ce texte. De manière similaire, chaque rapport de groupe de textes met en évidence les points communs et les différences de ce groupe par rapport au corpus dans sa globalité.

Ce positionnement des textes et des groupes de textes par rapport aux éléments de plus haut niveau les englobant (respectivement, groupe de textes et corpus) permet d'obtenir des informations pertinentes sur la répartition et la localisation des RTO LUCIA représentant les domaines étudiés en corpus. Pour aller plus loin dans de telles analyses, nous tenons compte des particularités du niveau global (que l'on pourrait appeler « signaux forts ») à un niveau plus local (où prennent place ce que l'on pourrait appeler des « signaux faibles »). Pour mettre en œuvre cette prise en considération du niveau global, nous avons pondéré les classements des isotopies inter-textuelles au niveau du groupe : si dans le groupe et dans le corpus, une même isotopie inter-textuelle est très présente, alors on diminue son importance dans le groupe ; au contraire, si dans le groupe, une isotopie inter-textuelle est présente et qu'elle l'est moins dans le corpus, alors on augmente son importance. Ces deux conditions permettent de faire ressortir des propriétés des groupes masquées par les propriétés globales du corpus dont, par définition du corpus, chaque texte hérite. Ce positionnement des groupes par rapport aux corpus aide ainsi à identifier comment les ressources sont projetées sur les différents paliers textuels (texte et groupe) du corpus analysé et en quoi elles permettent de différencier et d'isoler des groupes et des textes.

Le corpus d'étude considéré ici est constitué d'articles traitant de l'actualité boursière. La dimension temporelle a donc une forte influence sur le contenu de ce type de textes. Pour mettre visuellement en évidence cette influence, nous proposons à l'utilisateur une carte temporelle animée (également au format SVG) présentant l'enchaînement automatique de cartes réalisées à partir de ses RTO LUCIA et de textes compris sur une certaine fenêtre temporelle glissante. Dans les extraits présentés en figure 7, la fenêtre temporelle est d'un mois et l'unité de déplacement de cette fenêtre est d'un jour. Une telle carte temporelle permet de repérer

les moments auxquels certaines parties de ressources sont présentes en corpus. Ainsi, il est possible de mettre en évidence la « pertinence » d'une ressource au fil du temps dans le flux documentaire : une partie de la ressource fréquemment présente dans le flux laisse penser à une forte stabilité de cette dernière ; au contraire, il est possible de s'interroger sur une ressource très rarement présente et de penser à sa mise à jour.

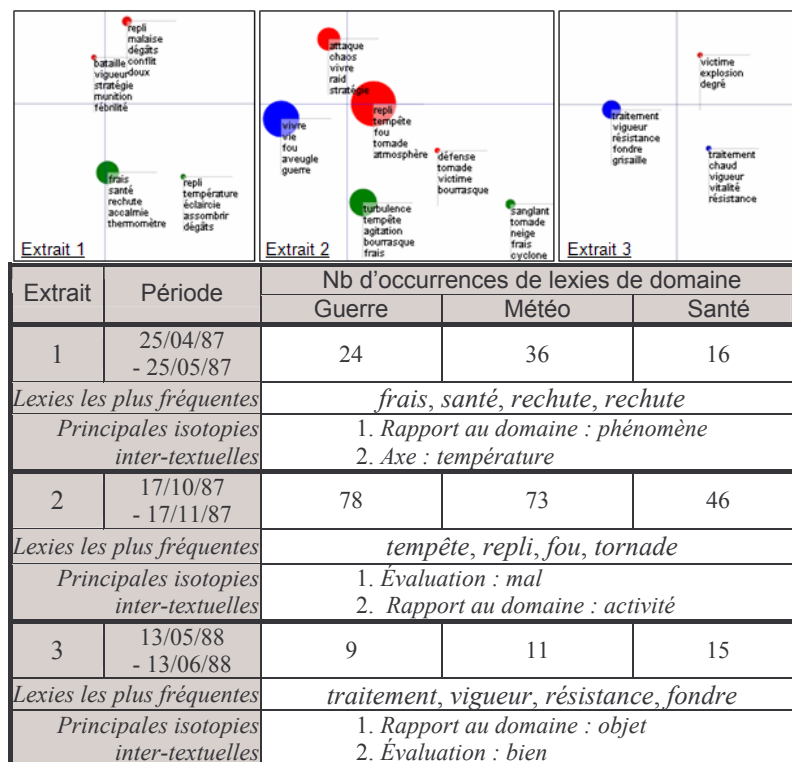


FIGURE 7 – Trois extraits de la carte temporelle du corpus et tableau présentant les extraits. Version SVG dynamique de cette carte : <http://www.info.unicaen.fr/~troy/isometa/cartes/>

Les cartes ont permis de déduire plusieurs informations synchro-diachroniques sur les ressources exploitées par rapport au corpus d'étude. Il a ainsi été possible de remarquer quelles parties de ressources étaient présentes dans le corpus ou dans un sous ensemble du corpus. En allant plus loin dans la « localisation » d'une partie de ressource en corpus, les cartes temporelles ont permis de mettre en évidence des usages termino-

logiques associés à une période donnée. Par exemple, dans le cadre de notre étude sur la métaphore conceptuelle (Roy et *al.*, 2006), nous avons pu mettre en évidence qu'un lexique guerrier était particulièrement présent lors des périodes de grandes tensions boursières. C'est un retour intéressant sur la ressource de savoir à quel moment elle se retrouve mise en avant par l'actualité. Cela incite éventuellement l'utilisateur à mettre à jour ses RTO. Si ce dernier observe une forte décroissance de l'utilisation d'une partie de sa ressource au fil du temps, alors il peut s'interroger sur la suppression de la partie concernée. Au contraire, si une partie de RTO est de plus en plus présente, alors il peut se demander si les parties « voisines » de la ressource ne devraient pas être étoffées, c'est-à-dire décrites plus finement (par exemple, si les lexies d'une ligne d'une table sont de plus en plus présentes, alors il est tout à fait envisageable de les détailler avec de nouveaux attributs et valeurs d'attributs à l'aide d'une table héritant de cette ligne). Enfin, si des lexies sont fréquemment présentes tout au long de la période étudiée, alors l'utilisateur peut se demander si elles ne sont pas trop générales et ainsi les décrire un peu plus.

5. CONCLUSIONS ET PERSPECTIVES

À travers nos travaux, nous cherchons à apporter une contribution conjointe à la linguistique de corpus et aux modèles de représentation de RTO. Dans notre démarche, ces deux aspects ne peuvent être dissociés l'un de l'autre car les ressources que nous considérons sont produites à partir d'extractions et d'analyses sur corpus, elles mêmes réalisées principalement par projections des ressources sur les corpus. Il en découle une boucle d'interdépendance entre les textes et les ressources, boucle au cœur de laquelle le sujet utilisateur / interprétant tient une place primordiale. Ainsi les RTO utilisées par un sujet dans une tâche interprétative de traitement de corpus ne sont ni un préalable à l'instrumentation, ni une finalité de cette instrumentation. Elles sont simplement un reflet de l'activité de l'utilisateur, activité qui, a priori, n'est pas bornée dans le temps.

Une ressource terminologique n'est jamais stabilisée et finalisée. C'est vrai des ressources terminologiques en général du fait de la diachronie inhérente aux langues naturelles. C'est encore plus vrai dans le cas de ressources produites de manière centrée utilisateurs car en tant que reflet d'une activité susceptible d'évoluer, elles ont, elles aussi, un caractère évolutif et par nature non finalisé. Ainsi, comme la question de l'extraction de ressources à partir de corpus a souvent fait l'objet

d'instrumentations logicielles, les questions de la maintenance dans le temps d'une ressource et de son partage dans une communauté d'individus doivent aussi faire l'objet de solutions techniques. C'est ce à quoi nous contribuons avec différents outils *open-source* tels que Proxi-Docs, par exemple. Il nous semble que ces solutions pour mieux gérer et mettre à profit des RTO personnelles doivent permettre plusieurs types d'analyses : des traitements chronologiques de corpus de documents datés (bien souvent des flux documentaires tels que des dépêches d'agences de presse) ainsi que des traitements multi-échelles (paragraphe, document, groupe de documents, corpus dans son ensemble par exemple) pour mieux appréhender les rapports entre le local et le global dans la dimension intertextuelle des corpus.

Enfin, en tant qu'une ressource terminologique centrée utilisateur est étroitement liée à un ou plusieurs corpus d'étude, il nous semble que mieux caractériser ce type de ressources peut aussi avoir comme intérêt de mieux caractériser certains genres textuels (documents d'actualité, messages issus de forums de discussion, bases documentaires thématiquement homogènes, etc.). Dans ce but, différentes expérimentations multipliant les contextes d'utilisations et les utilisateurs (exploration de corpus issus du Web et repérage des thèmes abordés dans un corpus d'articles journalistiques) sont toujours en cours de réalisation.

6. RÉFÉRENCES

- T. Berners-Lee. What the Semantic Web can represent? W3C, <http://www.w3.org/designissues/rdfnot.html>, 1998.
- D. Bourigault et N. Aussenac-Gilles. Construction d'ontologies à partir de textes, Actes de de *Traitement Automatique des Langues Naturelles (TALN)*, Tome 2, 27-47, Batz sur Mer, France, 2003.
- J.-M. Bourroche et G. Saporta. *L'analyse des données*, Paris : Presses Universitaires de France, 1980.
- J. Charlet, P. Laublet, C. Reynaud. *Web Sémantique*. Rapport de l'Action Spécifique 32 CNRS / STIC, 2003.
- A. Condamines (dir.) *Sémantique et corpus*, Paris : Hermès, 2005.
- C. Kerbrat-Orecchioni. Sémantique, *Encyclopedia Universalis*, 693-699, Understanding and creating sentences, *American Psychologist*, 18, 735-751, 1988.
- G. Lakoff et M. Johnson. *Metaphors we live by*, Chicago Press, 1980.
- A. Nicolle, P. Beust, V. Perlerin. Un analogue de la mémoire pour un agent logiciel interactif, revue *In Cogito*, n°21, 37-61, 2002.

- V. Perlerin et P. Beust. Pour une instrumentation informatique du sens, dans M. Siksou, *Variation, construction et instrumentation du sens*, 197-229, 2003.
- V. Perlerin. *Sémantique légère pour le document*, Thèse de doctorat en Informatique, Université de Caen Basse-Normandie, 2004.
- F. Rastier. *Sémantique interprétative*, Presses Universitaires de France, 1987.
- F. Rastier, M. Cavazza et A. Abeillé. *Sémantique pour l'analyse*, Masson, 1994.
- F. Rastier. *Arts et sciences du texte*, Presses Universitaires de France, 2001.
- F. Rastier. Enjeux épistémologiques de la linguistique de corpus, dans G. Williams, *La linguistique de corpus*, Presses Universitaires de Rennes, 2005.
- T. Roy et P. Beust. ProxiDocs, un outil de cartographie et de catégorisation thématique de corpus, Actes des *Journées internationales d'Analyse statistiques des Données Textuelles (JADT)*, Tome 2, 978-987, Louvain-la-Neuve, Belgique, 2004.
- T. Roy, S. Ferrari et P. Beust, Étude de métaphores conceptuelles à l'aide de vues globales et temporelles sur un corpus, Actes de *Traitement Automatique des Langues Naturelles (TALN)*, Leuven, Belgique, Tome 1, 580-589, 2006.
- T. Thlivitis. *Sémantique interprétative intertextuelle*. Thèse de doctorat en Informatique, Université de Rennes I, 1998.