Chapitre 4

Acquisition et structuration des ressources

À travers les chapitres précédents, nous avons vu émerger deux étapes majeures pour la mise en œuvre du système LUCIA : l'acquisition et la structuration de ressources d'une part (1 sur la figure 38) et l'utilisation de ces ressources à des fins d'analyse d'autre part (2 sur la figure 38). Ces étapes s'avèrent communes à de nombreuses applications de modèles de TAL et en particulier aux études sur corpus. Bien qu'elles entrent pour nous dans le cadre d'un processus itératif dans lequel l'utilisateur joue un rôle central (figure 38), c'est dans un souci de clarification que ces deux étapes sont présentées dans deux chapitres distincts. Le présent chapitre traite de l'étape d'acquisition et de structuration des ressources. Cette étape constitue l'amorce du processus itératif. Elle n'est pas définitive : les propositions initiales peuvent être révisées à la suite de l'évaluation d'analyses, d'un changement de tâche ou d'un changement de point de vue sur les données (étapes 3 et 4 sur la figure 38).

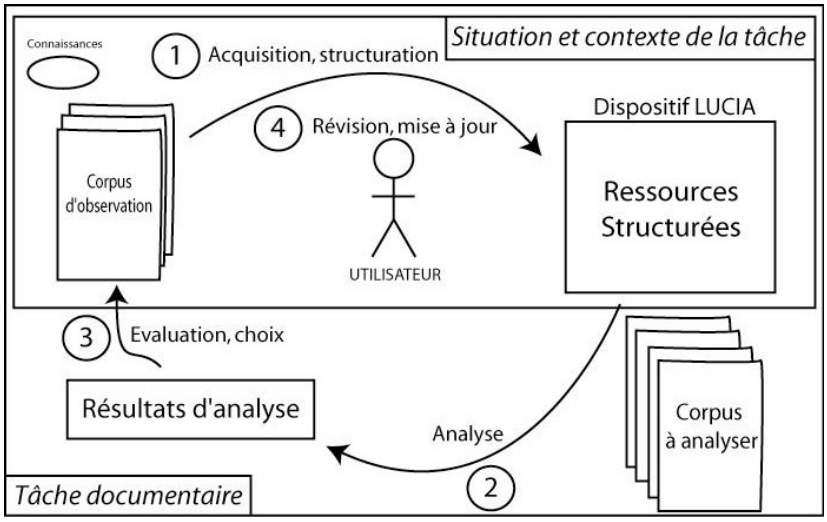

Figure 38 – Processus itératif d'utilisation du système

L'acquisition et la structuration des ressources du système sont deux opérations directement dépendantes de la tâche pour laquelle le système est employé. Du point de vue du modèle de catégorisation et de description lexicale, il n'y a pas de différence en fonction des tâches ; ce sont les données proposées par l'usager qui différeront, pas les concepts du modèle ni les solutions logicielles qui les mettent en œuvre. En revanche, comme nous le verrons dans le prochain chapitre, la spécificité des tâ-

ches possibles impliquent des modalités d'exploitation et de présentation des résultats d'analyse différentes.

Dans ce chapitre, nous débutons par la présentation des tâches pour lesquelles LUCIA peut être utilisé, nous verrons en particulier les influences spécifiques de ces tâches sur la phase d'acquisition et de structuration des ressources (4.1). Pour assister l'étape d'acquisition d'entités lexicales, nous préconisons l'utilisation d'un corpus d'observation. Nous présentons dans une seconde partie (4.2) quelles sont les conditions de constitution d'un tel corpus et pourquoi, malgré l'existence de nombreuses solutions d'extraction supervisée, nous proposons une solution logicielle minimale suffisante pour nos applications. Nous présentons également les solutions mises en place pour éventuellement prendre en considération les variantes morphosyntaxiques des entités lexicales lors des analyses ultérieures. Pour être utilisables par le système, les entités retenues doivent être placées dans les structures catégorielles et descriptives que nous avons présentées dans le précédent chapitre. Le logiciel LUCIABuilder (4.3) rend possible les interactions nécessaires à une telle construction. La construction de dispositifs n'est pas une tâche simple à réaliser. L'observation des attributs au sein de dispositifs construits permet la mise au jour de propriétés et de relations intéressantes entre attributs. Ces propriétés peuvent être exploitées pour l'élaboration de processus automatique d'assistance à la construction et à la révision de dispositifs (4.4). Enfin, une dernière forme d'assistance sera proposée sous la forme d'un protocole de construction d'un dispositif (4.5).

4.1 Les tâches

Nous avons déjà abordé l'importance de la tâche dans les étapes d'utilisation de LUCIA. Celles pour lesquelles notre système peut être utilisé doivent envisager une exploration assistée du contenu de textes : ceci constitue le premier aspect générique des tâches possibles (4.1.1). L'exploration du contenu de textes est par exemple utile à une tâche de veille documentaire (4.1.2) ou pour l'étude d'un fait de langue tel qu'une métaphore conventionnelle (4.1.3). Mais c'est également le cas pour d'autres types de tâches qui dépassent les champs de la linguistique informatisée ou de l'informatique linguistique (4.1.4). Quelle que soit cette tâche, il faut que l'utilisateur soit capable de faire les choix adéquats lors de l'acquisition et la structuration. Les contraintes du modèle et des implantations logicielles l'assisteront au cours de cette étape.

4.1.1 Aspects génériques des tâches

Quelles que soient les circonstances d'utilisation de LUCIA, l'emploi du système nécessite en premier lieu que l'usager puisse définir sa tâche, qu'il soit capable de savoir quels types de service il attend de la part des logiciels. Ce sont les buts de cette tâche qui dirigeront à la fois ses choix pour acquérir les ressources les structurer, et les biais d'exploitation et de visualisation des résultats qui lui seront proposés.

Comme nous le montrons par l'exemple dans le prochain chapitre, il y a deux aspects génériques aux tâches pour lesquelles LUCIA peut être utile :

- Premièrement, *le système permet l'accès à des documents particuliers parmi des documents inconnus ou dans des ensembles qu'il serait trop long de parcourir sans recours à un système informatique dédié*. Les particularités intéressantes des documents à proposer à l'utilisateur constituent une spécificité de la tâche. Les critères pour savoir si un document est susceptible de répondre aux attentes de l'usager relèvent de deux dimensions distinctes ; l'une concerne le travail de l'utilisateur, l'autre celui des logiciels. Nous avons vu dans le chapitre précédent que les dispositifs permettent l'expression d'un point de vue de l'utilisateur sur des données en rapport avec sa tâche. Dans le cadre d'une étude linguistique, ce point de vue peut avoir statut d'expertise. Dans le cadre d'une tâche documentaire plus répandue, il s'agit de l'expression de besoins particuliers. Les traitements effectués à partir des données proposées par l'utilisateur seront différents selon le type de tâche, mais les données seront proposées sous la même forme : des associations d'entités lexicales avec des attributs selon les modalités prévues par le modèle. Nous présenterons dans ce chapitre le logiciel qui permet de créer ses données (4.3) : nous verrons alors que, quelle que soit la tâche initiale, seul le répertoire où seront stockés les fichiers correspondant différera alors.

- Deuxièmement, *le système permet l'exploration assistée du contenu de documents*. Nous sommes alors dans le même cas de figure que pour l'analyse d'ensembles de documents : les données proposées en fonction des contraintes du modèle de description et de catégorisation seront manipulées de façon différente en fonction de la spécificité de la tâche mais c'est bien dans des dispositifs LUCIA tels que nous les avons présentés dans le chapitre précédent que seront fournies les données par l'usager.

Ainsi, l'acquisition et la structuration des données relèvent avant tout de l'utilisateur. C'est un fait inhérent aux systèmes centrés sur l'individu. La particularité du modèle LUCIA (par rapport à PASTEL [Tanguy et Thlivitis, 1996*] par exemple) est que les choix de structuration ne sont pas tant contraints par des considérations provenant d'une théorie sémantique - en l'occurrence la SI – que par la tâche précise pour laquelle le système est utilisé. C'est d'ailleurs pour cela que notre approche de la SI est plus opportuniste que dans des travaux informatiques qui voudraient l'implanter entièrement. Dans les parties suivantes, nous présentons les deux tâches pour lesquelles a été utilisé LUCIA en précisant quelles sont leurs spécificités et les incidences de ces spécificités sur l'acquisition et la structuration des ressources.

4.1.2 Veille documentaire

4.1.2.1 Présentation de la tâche

La veille documentaire informatisée est une activité médiatisée par l'ordinateur qui a pour but de répondre aux questions suivantes :

- Qu'est-ce qui se dit sur mon/mes sujets d'intérêt ?
- Comment les textes abordent-ils ce sujet ? Et donc quelles sont les parties des textes relatives à ce sujet ? Quelles sont les aspects de ce sujet abordés dans les textes ?

Dans l'industrie, la veille documentaire s'appuie principalement sur des bases de données spécialisées (bases de brevets comme celle de l'INPI ou l'USPTO[68], bases de données bibliographiques, factuelles, universitaires ou commerciales du type PASCAL, MEDLINE, CAB et INSPEC[69]) pour :

[68] Institut national de la propriété industrielle – http://www.inpi.fr et US Patent and Trademark Office – http://www.uspto.gov

[69] PASCAL de l'INIST-CNRS spécialisée en Sciences, Technologie et Médecine : http://www.inist.fr/PRODUITS/ pascal.php, MEDLINE de la National Library of Medicine : http://www.nlm.nih.gov, les CAB de la société Silver Platter - OVID sont des ensembles de bases de données spécialisées dans de nombreux domaines : http://www.ovid.com, INSPEC de l' Institution of Electrical Engineers est spécialisée dans les sciences physiques : http://www.iee.org

- constituer des dossiers de synthèse rassemblant par exemple un état de l'art sur un sujet, agrémentés de références bibliographiques, de brevets, de programmes de recherche, etc. ;
- produire des analyses statistiques et des graphiques pour évaluer la productivité par pays/laboratoire/entreprise sur un sujet, lister les organismes, les auteurs et les coopérations dans le domaine, etc. ;
- effectuer des traitements infométriques spécialisés tels que l'extraction terminologique liée à un thème pour la mise en évidence d'associations entre concepts (ou *clusters*), etc.

Lorsqu'il s'agit de traiter de l'information non structurée (en opposition à ce que l'on trouve dans les bases de données) ces services consistent à surveiller des sites Internet ou des publications sélectionnées par des spécialistes pour informer périodiquement les utilisateurs des nouveautés qui concernent leurs domaines d'intérêt. Le traitement automatique du contenu des textes tout-venant pose alors les mêmes problèmes que ceux que nous avons évoqués dans le chapitre 2 (assistance à l'accès au contenu, prise en considération des particularités de la tâche, etc.). Les pratiques de veille ont beaucoup évolué ces quinze dernières années. Les techniques étant la plupart du temps le fond de commerce des nombreuses entreprises spécialisées dans ce domaine, il est difficile d'en dresser l'état de l'art. Cependant, des publications comme Archimag ou CaptainDoc[70] permettent d'être tenu au courant des nouvelles offres logiciels ou de services dans le domaine. Les entretiens que nous avons pu mener avec des professionnels (le service de veille d'EADS[71] et le CRITT BNC) ainsi que la lecture des publications spécialisées précitées, montrent que les besoins se situent ici aussi au niveau de l'accès au contenu des documents, de la personnalisation des processus de recherche et des interactions proposées aux usagers[72].

La veille documentaire se distingue de la recherche documentaire par son aspect répétitif. La tâche consiste soit à assembler un fond documentaire sur un sujet donné, soit à suivre l'évolution de ce sujet dans le temps. Lorsque cette tâche, effectuée pour le compte d'une organisation, a pour but l'observation, la recherche, le traitement, l'analyse et la diffusion d'une information stratégique à but décisionnel, on parle alors de veille technologique (ou de veille stratégique, concurrentielle, commerciale…) [Rostaing, 1993 : 6-28]. Il ne s'agit donc pas de permettre l'accès à une information précise ou locale comme c'est par exemple le cas dans les systèmes de Question/Réponse (ou *Q/A* pour *question answering* [73]) ou dans certaines pratiques de la recherche d'information, mais de récupérer le plus

[70] http://www.archimag.com et http://www.captaindoc.com

[71] European Aeronautic Defence and Space Company – l'entretien a eu lieu en octobre 2002 en présence de Pierre Beust et du responsable de la veille technologique de l'agence de St Cloud M. Roussel.

[72] Voir en particulier les dossiers de Captain-Doc: L'hypertexte à l'épreuve du sens (oct. 2003), Le document électronique et le temps (juin, 2002) et la rubrique Ged (Gestion électronique de documents) et Workflow d'Archimag.

[73] http ://trec.nist.gov/

de documents pertinents possibles en rapport avec un sujet donné. Faciliter l'exploration des documents rapatriés de façon personnalisée apparaît comme une nécessité. La veille documentaire, tout comme la recherche documentaire, souffre de manques concernant une véritable analyse du contenu des textes lorsque ceux-ci ne sont pas formatés. L'utilisation de ressources indépendantes des utilisateurs et du rapport qu'ils entretiennent avec leur(s) sujet(s) d'intérêt est ici aussi selon nous un frein à la personnalisation des services. Cette étape est la plupart du temps déléguée à un expert qui à l'aide d'entretiens et de documentations constitue des données utilisables par les processus automatiques. Certains services, comme ceux de la société *Pertinence Mining*[74] spécialisée dans la veille économique et stratégique, proposent des moyens de personnalisation à leurs utilisateurs. Il s'agit non seulement de préciser par exemple la fréquence à laquelle on veut être averti d'une nouvelle information mais également d'une sélection des sources à analyser (pour Pertinence Mining des sites de l'Internet), et la définition de « centres d'intérêt ». Ces centres d'intérêt sont essentiellement définis par un ensemble de mots-clefs et des résumés automatiques sont proposés pour synthétiser les informations. Nous montrerons dans le chapitre 5 que la structuration de ces termes, acquis non pas seulement en fonction des connaissances de l'utilisateur, mais également à partir de corpus en rapport avec la pratique du domaine d'intérêt par l'usager permet d'apporter une valeur ajoutée en terme de personnalisation et d'aide à l'analyse de contenu. Il s'agit d'aborder les thèmes non pas seulement à partir du lexique mais également à partir des descriptions sémantiques que l'on peut produire de ce lexique. En outre, nous renonçons à avoir un recours systématique à des experts linguistes ou terminologues car la veille documentaire doit être réactive aux changements. Un rapport sur la veille dans l'entreprise de 2002 [Seive, 2002] met l'accent sur la possibilité nouvelle donnée aux non-spécialistes de la veille dans les entreprises (des ingénieurs aux responsables du marketing) d'avoir accès à une masse importante d'informations en particulier à partir de l'Internet. Les moteurs de recherche proposent effectivement des systèmes de recherche simples, à contrario des bases de données qui utilisent des langages d'interrogation complexes. Suivant un schéma classique, le premier temps fut celui de l'engouement, c'est-à-dire une recherche frénétique d'informations. Les personnes s'inscrivaient à de nombreuses listes de diffusion et se trouvaient confrontées à « *un raz de marée informationnel* (sic.) ». L'avantage était l'ouverture et la diversité de sources d'information tandis que le principal inconvénient provenait justement de la masse d'informations souvent ingérable et dont la pertinence n'était pas systématiquement avérée (beaucoup de sites ne présentent aucun suivi éditorial). Ce constat est encore valable pour nombre de pratiques professionnelles parmi lesquelles on compte celles des chercheurs informaticiens. Pour certaines entreprises, la solution passe à l'heure actuelle par la mise en place d'une veille collaborative où chaque collaborateur apporte de l'information sur son domaine de compétence, après

[74] http://www.pertinence.net

classement de celle-ci dans une catégorie. Dans ces circonstances les manques en terme d'assistance personnalisée à l'accès au contenu restent d'actualité.

Pour fournir des textes pertinents et répondre aux attentes d'une tâche de veille documentaire donnée, un système doit pouvoir utiliser des ressources permettant la mise en place d'analyses pour effectuer :

- un filtrage des documents relatifs au sujet de la recherche ;
- un ordonnancement des documents du plus intéressant au moins intéressant en fonction de critères définis ;
- une mise en évidence des textes pertinents et des parties de textes pertinentes, celles en rapport avec le sujet pour en accélérer la lecture.

Ce sont ces types de services que nous nous proposons de fournir à l'aide des ressources LUCIA, nous en verrons les modalités dans le chapitre suivant.

Notre étude, en tant que thèse de doctorat n'avait pas de buts commerciaux immédiats. Il s'agissait principalement de présenter et d'évaluer un système original sans pour autant prétendre à sa possible utilisation telle quelle dans des pratiques dépassant le milieu de la recherche. Ainsi, nos propositions quant à la veille documentaire sont principalement présentées ici en tant que moyen d'évaluation de nos propositions quant à l'analyse personnalisée du contenu de textes bien que nous ayons déjà envisagé certaines modifications techniques à entreprendre dans le cadre d'une utilisation industrielle au vu des contacts que nous avons pu avoir avec certains professionnels (voir partie 5.4.3 - chapitre 5).

4.1.2.2 Acquisition et structuration des données

Pour une veille documentaire, la définition de la tâche nécessite tout d'abord que l'utilisateur sache quel sujet d'intérêt il attend de voir abordé dans des documents. Les sujets d'intérêts sont généralement à la croisée de plusieurs *domaines*. C'est par exemple le cas lorsqu'on s'intéresse à un sujet tel que « la bourse ». Ce sujet est à la croisée du domaine de la banque, de l'entreprise voire des affaires sociales (l'expérience décrite dans le précédent chapitre nous l'a montré). Le terme domaine est employé ici dans une acception intégrant un entour social propre à une pratique, même si cette pratique n'est pas lexicalisée dans la dénomination choisie pour qualifier le domaine. Cette dimension praxéologique de la tâche participe de la prise en considération de la *situation* de l'utilisateur dans son interaction avec la machine symbolisée par le rectangle intitulé *Situation et contexte de la tâche* dans la figure 38 p.111. Par exemple, la médecine sera considérée comme un domaine différent selon que l'on sera en présence de textes de vulgarisation ou d'articles scientifiques relatifs à cette science, ce

qui peut correspondre à deux pratiques différentes, celle du médecin en tant que praticien et celle du médecin en tant que chercheur. Dans l'un ou l'autre de ces cas, les textes peuvent receler des entités lexicales différentes pour des significations analogues (exemple : *opération de l'appendicite* dans un texte de vulgarisation et *appendicectomie* ou *ablation de l'appendice iléo-cæcal* dans un article scientifique[75]) ou à l'inverse présenter une même entité lexicale interprétable de façon différente (exemple : *appendicite* avoir trait à l'opération d'ablation dans un texte de vulgarisation ou à l'inflammation de l'appendice en question dans un article scientifique). Même dans les limites d'une pratique donnée, la polysémie peut apparaître à l'intérieur d'un même domaine, indépendamment même parfois de tout entour social. Par exemple, le *gel* dans le domaine de la climatologie peut avoir trait aux résidus d'eau glacée sur une surface ou au changement d'état de l'eau, du liquide vers le solide[76]. En SI, la notion de domaine est définie comme suit : c'est un *groupe de taxèmes lié à une pratique sociale. Il est commun aux divers genres propres au discours qui correspond à cette pratique* [Rastier, 2001a* : 298]. Le domaine est l'une des trois classes sémantiques envisagées. Rappelons que le taxème est la classe de sémèmes minimale en langue, à l'intérieur de laquelle sont définis leurs sèmes spécifiques et leur sème micro-générique (ex. : //secours// pour 'Pompiers', 'SAMU', 'Police') et que la dimension est la classe de sémèmes de grande généralité indépendante des domaines et qu'elle induit un sème macro-générique (/animal/ vs /végétal/). Même si la démarche n'est pas la même attendu que le domaine est ici défini dans un premier temps sans support sémique, nous rejoignons la définition du domaine de la SI et cela d'autant plus que celle-ci considère désormais possible la polysémie dans un domaine déterminé. La proposition « *dans un domaine déterminé, il n'existe pas de polysémie* » [Rastier, 1987* : 124] s'est vue agrémentée d'un adverbe en limitant la détermination : *dans un domaine déterminé, il n'existe généralement pas de polysémie* [Rastier, 2001a* : 298]. Le fait qu'un sujet puisse se trouver à l'articulation entre plusieurs domaines (dans le sens utilisé en recherche documentaire) est couramment admis par les spécialistes de la recherche documentaire informatisée, en particulier à la vue des expériences de catégorisations manuelles de textes lors de la constitution d'ensembles de répertoires hiérarchiques pour l'Internet [Chaffee, 2000]. Ainsi, si nous parlons de domaine en rapport avec un dispositif, l'accent est mis essentiellement sur *la manière dont on parle de ce domaine* et dans le contexte d'une tâche précise, *la manière dont l'utilisateur parle de ce domaine ou la façon dont il souhaite le voir abordé*. Le sujet d'une veille documentaire peut être paraphrasé comme par exemple « les catastrophes naturelles météorologiques » ou encore « les rapports entre les Etats-Unis et l'Europe ». L'intitulé du domaine fourni par le lecteur/utilisateur intègre l'entour social et praxéolo-

[75] Voir Par exemple, à l'adresse http://www.famili.fr/bonasavoir/1008237439/ : « *SANTE : Enceinte, quels sont les risques d'une appendicite?* » (Famili.fr) et dans « *L'appendicectomie : travail de libération de la cicatrice et de ses adhérences ; objectivation des modifications myofaciales par la posturologie »* de M. Bousquet (1992), mémoire de fin d'étude de la Collégiale Académique de France.

[76] Un constat similaire est proposé dans [Kaiser, 1995] au sujet de la polysémie de *champ* dans les manuels de bases de données.

gique de la tâche mais ne limite donc pas le sujet d'étude à des contraintes lexicologiques ou encyclopédiques.

Pour un domaine donné, on peut être intéressé par plusieurs thèmes. Nous avons déjà abordé la notion de thème dans le chapitre précédent (c.f. 3.3.1). Elle s'avère centrale pour l'exploration du contenu des documents : nous aurons l'occasion d'y revenir encore en la confrontant à l'empirisme à travers la description de résultats obtenus de certains de nos logiciels (c.f. 4.2.3 et 4.2.4). Pour l'heure, nous l'évoquerons relativement à un domaine : en veille documentaire, un thème est considéré comme un aspect spécifique d'un domaine. Un même domaine de tâche ou un même sujet d'intérêt d'une veille documentaire peut se concevoir en thèmes et sous-thèmes. Dans le domaine de la politique par exemple, on peut ainsi parler du thème des sanctions pénales infligées aux dirigeants, voire du sous-thème de l'inéligibilité[77]. Dans le domaine de la bourse, on peut être intéressé par le thème des places boursières en tant que lieux géographiquement situés ou par les influents sur les cours boursiers. Ainsi, après avoir défini un domaine, l'utilisateur du système dans le cadre d'une veille documentaire devra cerner les thèmes qui l'intéressent particulièrement. Du point de vue des ressources, nos solutions informatiques pour l'acquisition de données lui permettront d'être assisté dans la tâche qui consiste à associer des entités lexicales à des thèmes. Les contraintes du modèle de description et de catégorisation lui permettront de décrire son domaine et ses thèmes pour les rendre utilisables de façon automatique. L'association entités/thèmes proposée relève d'une approche lexicale qui n'a pour dessein que d'aider la description sémantique, en termes de traits sémantiques (les attributs), qui sera effectuée plus tard. La notion de thème est souvent intuitive. La qualification proposée n'explicite ni l'implication du lecteur, ni l'indépendance entre l'expression et le contenu du thème. Par implication du lecteur, nous entendons par exemple l'acte interprétatif d'actualisation ou de virtualisation de sèmes. Cela est d'autant plus vrai pour des sèmes évaluatifs voire thymiques dont l'apparition ne dépend pas du seul cotexte et genre textuel mais également de dimensions plus personnelles à l'usager (historiques, sociales, etc.). L'absence d'isomorphie entre les plans de l'expression et du contenu, soutenue entre autres par Hjelmslev [Hjelmslev, 1943], n'est pas contradictoire avec nos propositions. Les entités associées à un thème le sont par un usager moyennant une tâche et un contexte précis : nous le rappelons une fois encore, ces mises en relations n'ont pas d'autres valeurs que celles de rendre des services partiellement automatisés à l'utilisateur qui en est l'auteur. Privilégiant dans un premier temps le signifié sur le signifiant, les descriptions finalement produites mettront en évidence des réseaux de récurrences sémantiques ; réseaux marquant l'implication de l'utilisateur et décrivant un thème indépendamment de ses

[77] On peut apprécier ici le caractère subjectif de la notion de thème et de hiérarchies thèmes/sous-thèmes. L'inéligibilité peut en effet faire l'objet d'un thème à part entière où l'implication des politiques constituerait un sous-thème. Cette sanction relève en effet d'une privation de droits civiques dont les intéressés n'ont pas de liens obligatoires avec les affaires publiques. Considérer un thème est souvent dépendant un point de vue particulier même indépendamment de textes.

représentants syntagmatiques ou plutôt, de façon corrélée à certains de ces représentants eux-mêmes associés à une façon de parler de ce thème.

4.1.3 Étude d'une métaphore conceptuelle

4.1.3.1 Présentation

Depuis une dizaine d'années, le renouveau de l'utilisation de corpus en linguistique et les progrès techniques connus en informatique ont permis la mise en place de nombreuses études sur corpus assistées par l'informatique. On parle souvent alors de « linguistique informatisée[78] ». Notons que ces travaux sont finalement apparus tardivement par rapport aux avancées des deux disciplines : l'influence de la critique chomskienne de l'empirisme y est pour beaucoup. Parmi les travaux de linguistique de corpus, nombreux sont ceux qui s'intéressent à la métaphore. La métaphore pose en effet certains problèmes pour les systèmes de résumé automatique de textes, la traduction automatique de documents ou encore la recherche documentaire (voir [Ferrari, 1997 : 7-12]).

On compte à l'heure actuelle des dizaines de définitions de la métaphore[79]. D'Aristote qui désignait ainsi *une substitution d'un mot à un autre* à Klinkenberg *in* [Charbonnel et Kleiber, 1999 : 157] qui y voit *un écart de catégorisation* qui *établie des connexions nouvelles dans nos structures encyclopédiques*, il semble difficile de trouver un consensus large sur le sujet. Tamba disait à ce propos que *dans le domaine linguistique notamment, le mot de métaphore sert à désigner des phénomènes mal circonscrits et si variés qu'il n'est pas toujours facile de savoir de quoi l'on parle au juste* [*ibid.* : 207]. Il s'avère que cette position se généralise finalement à bien des tropes, même celles moins populaires que la métaphore[80]. Les études sur la métaphore sont à l'heure actuelles largement dominées elles-aussi par l'approche conceptuelle, en particulier du fait de l'école californienne très prolixe en la matière. La théorie de Lakoff et Johnson [Lakoff et Johnson, 1980] connaît un très grand succès en accentuant le rôle de la métaphore comme une clef linguistique aux conceptualisations cognitives. Nous nous abstiendrons pour notre part d'aborder le phénomène dans cette dimension. En revanche, nous nous attarderons sur les métaphores conceptuelles telles qu'elles ont été introduites par ces deux chercheurs. Ce type de métaphores se caractérise par l'existence d'emplois métaphoriques récurrents qui font intervenir à la fois un même domaine source (aspect conventionnel) et un même domaine cible (aspect conceptuel). Selon l'hypothèse de Lakoff et Johnson, il existerait des corres-

[78] Ce terme se retrouve par exemple dans l'acronyme DELIC, l'équipe de recherche Description Linguistique Informatisée sur Corpus de l'Université de Provence et peut s'observer sous forme d'une traduction littérale en allemand avec *Computerlinguistik* et en anglais avec *Computational Linguistics*.

[79] Le site membres.lycos.fr/analogueman/resume.html en recense plus d'une vingtaine.

[80] Par exemple, Jean-Michel Messiaen remet en cause les définitions du Gradus [Dupriez, 1984] (qui fait souvent office de référence) et du Grand Robert (éd. de 1981) pour l'apologue et la parabole [Messiaen, 2000].

pondances ontologiques entre les entités du domaine source et les entités du domaine cible. C'est ce qui nous permettrait de raisonner sur le domaine cible en utilisant les connaissances utilisées pour raisonner sur le domaine source.

Les approches de la métaphore correspondent souvent à la définition de Klinkenberg ou celle de Dumarsais : *La Métaphore est une figure par laquelle on transporte, pour ainsi dire, la* signification propre *d'un mot à une* autre signification *qui ne lui convient qu'en vertu d'une comparaison*. [Dumarsais, 1977]. Une fois encore, l'approche rhétorique / herméneutique se distingue de ces conceptions. Elle envisage plutôt le phénomène relativement aux mécanismes de l'interprétation et conteste d'autant plus la définition de Dumarsais que l'opposition signification propre vs. signification littérale tend à s'appuyer sur une vérité factuelle ou historique (étymologique) éloignée des considérations contextuelles et synchroniques. Rastier s'intéresse ainsi plus à la sémantique des tropes en général qu'aux métaphores en particulier pour leur restituer leur dimension textuelle. Pour plus de détails sur les tropes et la SI, nous invitons à la lecture de [Rastier, 2001a* : 132-166]. La métaphore y est définie comme suit [*ibid.* : 160] et [Rastier *et al.*, 1994* : 98] : « *Pour la métaphore* in præsentia*, c'est la disparate des domaines ou des dimensions qui est l'interprétant. Si dans le contexte une isotopie générique est dominante, le sémème indexé sur cette isotopie sera comparé, l'autre comparant. (...) Quant à la métaphore* in absentia*, elle instaure une connexion symbolique qui doit être identifiée par des conjectures concordantes sur le discours, le type de l'œuvre, le genre de texte, la hiérarchisation idiolectale des isotopies*[81]*. »* Pour la SI, l'analyse du phénomène indépendamment d'un contexte est donc impossible. Les tropes ne sont que des moments singuliers de parcours interprétatifs qui mettent en jeu des opérations interprétatives remarquables.

En TAL, différentes approches se côtoient, qui apportent des solutions pour la détection et l'interprétation des métaphores. Sans en faire une présentation exhaustive, nous en retiendrons ici les principales orientations. Certaines recherches visent à expliciter la relation existant entre la source et la cible d'une métaphore, nécessitant de grandes bases de connaissances sémantiques, fortement structurées, pour leur mise en œuvre [Fass, 1997]. Selon les approches, la relation source–cible peut alors être vue comme essentiellement fondée sur l'analogie [Falkenhainer *et al.*, 1989] , [Gentner, 1983], ou plutôt porteuse de nouveauté, créatrice [Gineste *et al.*, 1997], [Indurkhya, 2002]. D'autres travaux n'ont plus pour objectif de rendre explicite la relation source–cible, mais plutôt de vérifier que l'interprétation des métaphores est similaire à celle des autres énoncés. Ainsi, Kintsch [Kintsch, 2000] montre que la signification d'une métaphore peut être représentée par un vecteur multidimensionnel,

[81] Rappelons que pour la SI, l'*interprétant* est une unité de contexte linguistique ou sémiotique permettant d'établir une relation sémantique pertinente entre des unités reliées par des parcours interprétatif. La hiérarchisation idiolectale, à laquelle il est fait allusion, consiste en l'évaluation relative des diverses classes définissant des isotopies génériques. Dans une métaphore, le comparé jouit par exemple d'une évaluation supérieure au comparant [Rastier, 2001a* : 299].

comme toute autre signification en LSA[82]. L'utilisation de LUCIA dans un tel cadre est en partie motivée par deux aspects spécifiques que les approches précédentes ne couvrent pas simultanément : l'utilisation de ressources simples à développer en terme de quantité et de temps de réalisation, d'une part, et la production pour un lecteur/utilisateur final d'une aide à l'interprétation qui lui soit rapidement accessible, d'autre part. Les travaux qui cherchent à expliciter la relation source–cible de la métaphore ont actuellement pour inconvénient de nécessiter des ressources complexes, selon nous difficilement réalisables par un utilisateur non-spécialiste. Ceux entrant dans la lignée de l'ASL présentent quant à eux l'inconvénient de ne pas fournir d'interprétation explicite de par la nature des représentations construites.

Nos propres travaux relatifs à la métaphore visent à montrer l'intérêt de LUCIA dans un cadre de linguistique informatisée. Notre but est de détecter et d'aider à l'interprétation du fait de langue par l'étude de contenu de textes. Ces travaux ont été entre autres motivés par l'intérêt porté à nos travaux par un spécialiste français du domaine : Stéphane Ferrari. L'étude de ce fait de langue peut paraître quelque peu antinomique au vu de certaines prises de positions de Rastier[83] et en particulier de sa critique justement de la prédominance de l'école californienne dans le domaine. Rastier regrette principalement le nombre des études sur les tropes telles que la métaphore et la métonymie aux dépens des autres figures telles que l'hypallage[84] ou encore l'antonomase[85] par exemple. Pour lui, la métaphore et la métonymie permettent de rester dans un cadre référentiel au monde en mettant en jeu respectivement un passage *d'une chose à une autre* et *d'une chose à sa voisine* dans les représentations hiérarchiques de concepts habituellement manipulées, au contraire de l'hypallage qui par exemple, tend justement à *troubler l'ordre du monde* et donc les représentations ontologiques [Rastier, 2001b* : 85]. L'étude de la métaphore n'est qu'un moyen d'évaluer le système LUCIA et les analyses du contenu de textes qu'il peut produire. Nous nous abstiendrons de ce fait de toutes conclusions linguistiques sur ce phénomène, le considérant quant à nos propositions d'aide à l'interprétation, comme tout autre fait de langue constituant un but analytique dans une tâche nécessitant un accès au contenu. Nous étudions une métaphore conventionnelle particulière, afin de dégager des principes généralisables par la suite. Il

[82] *Latent Semantic Analysis* ou ASL, Analyse Sémantique Latente, est une méthode de représentation où le sens d'un mot est défini par une position dans un espace à plusieurs dimensions [Landauer et Dumais, 1997].

[83] L'exemple le plus flagrant serait peut-être son entretien avec Pascal Michelucci intitulé pour sa retranscription sur le site de la revue Texto ! « *La métaphore est une figure outrageusement envahissante* » - http://www.revue-texto.net/Dialogues/Rastier-Michelucci.html

[84] L'hypallage, d'après le Littré cité dans [Dupriez, 1984* : 235], est le fait de paraître *« attribuer à certains mots d'une phrase ce qui appartient à d'autres mots de cette phrase, sans qu'il soit possible de se méprendre au sens »* comme dans « *Trahissant la vertu sur un papier coupable* » (Boileau) où grammaticalement parlant, c'est le papier qui est coupable.

[85] L'antonomase consiste d'après le Littré [Littré, 1964], à prendre un nom commun pour un nom propre, ou inversement un nom propre pour un nom commun comme dans « *balkanisation* » ou « *Washington s'embourbe dans la guerre* ».

s'agit des emplois récurrents de termes de la météorologie pour décrire des objets ou des phénomènes boursiers et économiques. L'étude d'une métaphore conceptuelle telle que nous l'avons menée, a pour but l'examen d'hypothèses linguistiques (analogie et nouveauté entre la source et la cible d'une métaphore lexicalisée) et l'évaluation sur du matériau empirique attesté de principes de détection et d'aide à l'interprétation de ce fait linguistique. Les résultats de cette étude seront présentés dans le chapitre 5 (partie 4.1.3).

4.1.3.2 Structuration des données

Dans le cadre d'une étude de fait de langue, la définition de la tâche nécessite tout d'abord que l'usager du système sache quel fait de langue il désire étudier. Cette lapalissade est utile car elle détermine le choix des ressources du système. Dans le cadre d'une étude sur une métaphore conventionnelle telle que nous l'avons menée, le fait de langue peut être exprimé par un syntagme ou un énoncé du type de ceux proposés dans [Lakoff et Johnson, 1980*] comme *argument is war* (« débattre c'est combattre[86] ») ou *love is a journey* (« l'amour est un voyage ») où les domaines cible et source de la métaphore sont désignés par un seul mot. La cible et la source doivent donc être préalablement identifiées si le cadre de l'étude implique justement d'en déterminer les relations en usage. La cible et la source d'une métaphore conventionnelle relèvent d'un domaine. Dans notre approche centrée sur l'utilisateur, nous utilisons la même définition de domaine que pour celle sur la veille documentaire.

4.1.4 Autres tâches

Les autres tâches envisagées concernent l'analyse thématique des textes ou leur étude linguistique. La veille documentaire comme la recherche d'informations sur le texte intégral ou la représentation de collections de documents amènent à rendre compte des thèmes présents dans les textes rapportés à un ou plusieurs domaines. Dans notre approche, il s'agit d'évaluer quantitativement et qualitativement cette présence au sein des textes. La quantité sera évaluée par la recherche et le comptage des récurrences d'attributs et valeurs d'attributs en fonction des ressources fournies[87]. La qualité sera évaluée en interaction avec l'utilisateur à travers des interfaces de lecture et de présentation des résultats. De nombreuses tâches sont possibles à partir de tels principes. Dans [Jenny, 1997], l'auteur présente des solutions informatiques pour l'analyse de contenu et de discours dans la recherche sociologique en français contemporain. Il regrette le peu d'intérêt suscité par les logiciels d'assistance à l'étude de textes dans ce champ d'étude mais présente les gains possibles d'une telle approche pour soulager le travail du chercheur. Notons que les travaux présentés dans [Souchard *et al.*, 1997] propo-

[86] Traduction proposée par Stéphane Ferrari [Ferrari, 1997*].

[87] Rappelons que si en SI, l'isotopie préexiste au repérage des sèmes dans notre système, les isotopies sont présupposées à l'aide des ressources de l'utilisateur et ne peuvent être validées en tant que telles que par lui.

sent une analyse d'un discours politique fondée en partie sur une analyse statistique lexicale supportée par le logiciel Termino [David et Souchard, 1995]. Il est possible d'utiliser LUCIA dans un tel champ d'application puisque le système permet une approche thématique de l'exploration du contenu de textes. La valeur ajoutée par rapport aux autres solutions passe dans ce cadre par la possibilité de structurer aussi finement qu'il est nécessaire les domaines d'intérêt pour mettre au jour des isotopies en rapport avec les thèmes intéressants.

Notre but est de proposer une instrumentation informatique pour l'assistance à l'accès à des documents et l'exploration de leur contenu. Dans le cadre d'une étude linguistique, l'une des difficultés lors de la constitution des ressources est de faire la part des choses entre ce qui concerne le fait observé et ce qui concerne l'interprétation de l'auteur de ces ressources. L'usager a alors statut d'expert. En revanche, dans le cadre d'autres études, la subjectivité des données est un principe majeur du système. Nous ne nous plaçons pas dans le cadre d'un modèle linguistique inductiviste, mais dans celui d'un modèle informatique qui présuppose une interprétation (celle proposée à travers la constitution des ressources) et qui instrumentalise cette interprétation (en permettant la validation ou l'invalidation des propositions formulées dans la constitution et l'agencement des ressources) dans une dimension calculatoire et dans une optique d'assistance à des tâches documentaires (en se basant sur le calcul des différences et des récurrences). La tâche, définie par le lecteur/utilisateur, peut souffrir d'un manque de précision du simple fait de n'être exprimée au départ que par une paraphrase, un syntagme ou un mot. L'expression de cette tâche relève du lecteur/utilisateur : c'est lui qui en mesure la portée (consciemment ou non) dans toutes les étapes d'utilisation du système, et cela en particulier, lors de l'éventuelle constitution d'un corpus d'observation. Pour permettre l'exploration du contenu de textes, le processus d'utilisation de LUCIA nécessite d'être amorcé par la constitution de ressources – cette étape est l'objet de la partie suivante. Les ressources vont être constituées à partir des connaissances de l'utilisateur ou d'un corpus d'observation.

4.2 Corpus d'observation et acquisition

Dans cette partie, nous présenterons la phase d'amorce indispensable à l'utilisation du système LUCIA : l'acquisition d'entités lexicales. Nous débuterons par la définition du corpus d'observation potentiellement utilisable pour acquérir et observer le comportement de ces entités dans un cotexte précis (4.2.1). Le corpus d'observation doit permettre l'extraction d'entités lexicales en rapport avec la tâche courante et l'observation du comportement de ces dernières dans un contexte précis. Nous apporterons ensuite quelques informations relatives aux solutions informatiques envisageables pour accélérer cette phase d'acquisition (4.2.2). Les deux parties suivantes, seront consacrées à deux propositions logicielles minimales entrant dans un cadre de *sémantique légère* qui peuvent être utilisées pour une telle tâche. Il s'agit d'une part, de MEMLABOR qui met en place des calculs statisti-

ques très simples pour proposer à l'utilisateur des listes de graphies représentatives d'un corpus (4.2.3) et de THEMEEDITOR d'autre part, qui permet une évaluation des résultats obtenus du premier logiciel (4.2.4).

4.2.1 Définition du corpus d'observation

Le corpus d'observation doit être constitué de textes véritables, de textes qui n'ont pas été créés artificiellement pour la tâche. Nous nous plaçons dans un premier temps dans le cadre d'observation empirique de phénomènes pour l'acquisition et la structuration des données. Les trois pôles intrinsèques des textes, à savoir l'impression référentielle que chacun d'entre eux suscite, et les foyers énonciatifs et interprétatifs représentés et situés en partie par des règles du genre textuel, restituent, comme en trace, les trois pôles extrinsèques du texte : l'auteur, l'entour et le destinataire [Rastier, 1996]. Aussi, pour pouvoir en retirer des indications sémantiques valables, le corpus d'observation doit être constitué en fonction des pôles intéressants pour la tâche : il ne doit pas être *du* texte mais bien *des* textes. Ces textes doivent être en rapport avec la pratique de l'auteur du corpus[88] en tant que lecteur et acteur central de la tâche. Du point de vue du corpus d'observation, nous rejoignons ainsi les propositions de [Thlivitis, 1998*] relatives à l'*anagnose*. Dans les travaux de Thlivitis relatifs à la Sémantique Interprétative Intertextuelle (SII), l'anagnose est un ensemble de textes que convoque chaque lecteur confronté à un nouveau texte afin d'amorcer le processus de construction de sens. L'anagnose est vue comme un contexte de lecture intertextuel propre à un lecteur. Un texte placé dans une anagnose particulière reçoit une charge sémantique du fait même des relations qu'il entretient avec les autres textes au sein de cette anagnose. Ainsi Thlivitis souligne que : « *… selon son point de vue et sa compétence interprétative, le lecteur situe le texte dans une société textuelle qui détermine la lecture qu'il fait de l'intra de ce texte. (…). Le sens des mots et des expressions d'un texte vient d'une réutilisation ou d'une modification des sens qui préexistent dans la société de textes où on le situe. Ces relations entre textes se situent précisément au niveau intertextuel. Là, les entités concernées sont soit les textes entiers soit leurs intratextes.* » En ce qui concerne les tâches que nous proposons d'assister, les textes du corpus d'observation constituent une anagnose à but, non pas normatif, mais descriptif et productif. Placer un texte dans un corpus d'observation intègre la dimension subjective du lecteur dont nous avons déjà souligné l'importance dans notre système. Dans [Pincemin, 1999a*], il est précisé qu' *un corpus n'est exploitable qu'en se référant à la manière dont on interprète sa constitution. Sa valeur n'est pas déterminée par sa forme, mais relève de son adéquation à une visée interprétative claire.* Bien que les buts ne soient pas ici tant descriptifs (il faudra associer aux entités lexicales repérées des éléments de signification), que productifs (le but final est de construire des ressources utilisa-

[88] L'auteur du corpus est celui qui rassemble les textes au sein de cet intertexte, il ne s'agit pas bien sûr de l'auteur des textes qui le composent.

bles automatiquement), le corpus d'observation doit être constitué dans la visée interprétative adaptée à la tâche, ce qui peut donner lieu à l'expression d'un certain cadre limitatif à travers le choix des genres des textes par exemple.

Pour une tâche de veille documentaire, le lecteur peut rassembler des textes avec comme objectif la description d'un domaine en rapport avec une pratique donnée. Les textes, qui doivent aborder le sujet d'intérêt de la tâche de veille sont assemblés selon le point de vue du lecteur par rapport à la fois à une pratique de ce domaine et au sujet d'intérêt qui détermine sa tâche. Le corpus d'observation a un but productif dans le sens où il constitue une partie de la base d'acquisition des entités lexicales. Il est également le support permettant plus tard de définir certaines modalités de catégorisation. C'est dans les textes du corpus d'observation que le lecteur peut puiser certains critères de catégorisation des entités lexicales en fonction de l'observation du comportement de ces dernières dans l'intertexte ainsi constitué (les exemples proposés dans ce tapuscrit s'appuient sur l'interprétation de textes).

Pour une étude de fait de langue telle que nous la proposons, le but descriptif du corpus d'observation est celui du fait de langue en question. Les textes du corpus d'observation doivent donc présenter des lexicalisations du fait de langue. Le caractère productif du corpus est le même que dans le cadre de la veille documentaire : il se situe au niveau de l'acquisition et de la structuration des entités lexicales. Si dans un tel champ d'application, les critères de sélection de textes du corpus doivent répondre aux exigences d'une telle étude (présence du fait de langue, objectivation des conditions de production des textes, conscience de la portée de conclusion en fonction du corpus), l'intérêt porté à la mise au jour de propositions d'aide à l'interprétation admet une certaine subjectivité dans le choix des textes à observer. Pour nos travaux sur la métaphore conceptuelle conventionnelle mettant en jeu les domaines de la bourse et de la météo, nous avons repris un corpus déjà utilisé pour une étude comparable [Ferrari, 1997*]. Ce corpus est constitué de l'ensemble des articles relatifs à la bourse entre 1987 et 1989 dans le journal *Le Monde sur CD-ROM*, en tout 594 articles pour un total de 450 000 tokens. Il contient notamment de nombreux emplois d'une métaphore conceptuelle conventionnelle décrivant les phénomènes boursiers en termes météorologiques que nous appellerons (comme l'avait fait avant Stéphane Ferrari) « la météorologie boursière ». Le domaine du corpus est celui de la cible de la métaphore, en l'occurrence pour notre étude sur « la météorologie boursière » : la bourse.

L'utilisation d'un corpus d'observation n'implique pas ici une approche terminologique textuelle (au sens de [Bourigault et Slodzian, 2000]) attendu qu'il n'y a ici aucun désir d'exhaustivité. Il but n'est pas de fixer la terminologie d'un domaine ou même d'une tâche. Cela a déjà été évoqué : l'acquisition d'entités lexicales a un statut d'amorce et ne constitue donc ni une fin en soi, ni une étape obligatoire – c'est un point qui nous distingue de nombreuses études analogues et qui permet de situer nos ressources par rapport à celles obtenues exclusivement de manière automatique ou par un expert, comme avant tout personnelles à l'utilisateur ou au groupe d'utilisateurs qui les ont élaborées dans le

cadre d'une tâche située. Pour assister l'éventuelle acquisition depuis un corpus d'observation des entités qui seront catégorisées, nous proposons des solutions peu nécessiteuses en ressources, fondées essentiellement sur des méthodes statistiques d'analyse de corpus.

4.2.2 Extraction supervisée

L'extraction supervisée a pour but d'assister l'utilisateur pour extraire du corpus d'observation des entités lexicales utiles à sa tâche. Dans le cadre d'une veille documentaire comme celui de l'étude d'une métaphore conceptuelle, il s'agit au final d'avoir des entités lexicales associées à des domaines pour les décrire dans des dispositifs. Cette étape ne constitue pas un axe d'étude central pour nous, attendu qu'elle constitue un champ de recherche à part entière déjà largement exploité par d'autres et dont les résultats satisfont nombre d'applications [Nazarenko et Hamon, 2002a]. Puisqu'elle entre dans le processus itératif d'utilisation du système, il apparaît cependant indispensable de l'aborder.

À partir d'un corpus d'observation homogène correspondant au(x) domaine(s) d'intérêt de la tâche en cours, on peut extraire les entités lexicales les plus redondantes. Si celles-ci ne permettent pas de discriminer les documents à l'intérieur du corpus, elles peuvent en revanche permettre à l'utilisateur de faire des associations entités / thèmes qui seront utiles à la description ultérieure. Du point de vue de l'extraction à partir des textes et dans une optique exclusivement computationnelle, l'entité lexicale est comparable au concept du *n-gramme* du TAL qui fait référence à un motif (parfois complexe : composé de *n* motifs) répété dans un texte ou un corpus de textes. Nous l'avons déjà évoqué dans le premier chapitre, nous tentons à travers nos réalisations informatiques d'être le moins dépendant possible des particularités des langues traitées pour conférer à nos propositions un caractère multilingue et parvenir à un système relevant d'une *sémantique légère*. Dans le chapitre 2, nous avons déjà présenté certaines méthodes d'extraction de termes utilisant des analyses syntaxiques du type de celle du système LEXTER [Bourigault, 1994*]. Ces systèmes sont performants même en terme d'*assistance* à l'acquisition d'entités lexicales. Les solutions « d'extraction terminologique » sont déjà nombreuses (voir pour un état de l'art [Jacquemin et Zweigenbaum, 2000] section G.2 et [Chalendar, 2002 : 55-56 et 63-70]) qui, pour la plupart extraient principalement des syntagmes nominaux récurrents (ou des syntagmes verbaux comme dans SYNTEX [Bourigault et Fabre, 2000*]) dans un corpus en utilisant des ressources au minimum propres à la langue manipulée et parfois même aux domaines des corpus traités. Certains algorithmes de recherche de *n-grammes* pour le TAL comme par exemple l'algorithme glouton[89] utilisé par [Ahonen-Myka *et al.*, 1999] se bornent à n'utiliser pour l'extraction qu'une liste de termes désignés comme peu informatifs. Ces listes appelées *stop-list* ou lis-

[89] En informatique, on désigne par glouton (ou *greedy*) les algorithmes qui construisent une solution de manière incrémentale, en faisant à chaque pas un choix minimisant une fonction de coût.

tes d'exclusion, sont faciles à élaborer attendu qu'elles recèlent principalement les termes grammaticaux de la langue utilisée, termes en nombre fini et exhaustivement connus. Vergne [Vergne, 2003] utilise le même type d'algorithme sans aucune ressource extérieure au corpus analysé et tente même de retrouver automatiquement ces listes d'exclusion. Ces solutions permettent d'extraire des motifs redondants. Pour les associer à des thèmes, on emploie des techniques de regroupement. Selon [Grefenstette, 1994], l'approche dominante pour le regroupement sémantique d'entités lexicales peut être divisée en trois étapes : extraction des cooccurrents d'un mot, association à chaque mot de l'ensemble de ses cooccurrents et mise en évidence de la proximité/distance des mots deux à deux en fonction des cooccurrents qu'ils partagent ou pas pour parvenir à un découpage en classes en fonction des plus ou moins grandes proximités entre mots. Les techniques varient à chaque étape : fenêtre de lecture, calculs statistiques [Lebart et Salem, 1994] ou exploration du contexte syntaxique pour le regroupement comme dans le système ZELLIG [Habert *et al.*, 1996]. Dans [Bouaud *et al.*, 2000 : 33], il est rappelé que les graphes issus du système ZELLIG peuvent être exploités comme une carte différentielle du corpus [Rastier *et al.*, 1994*] à partir duquel ils ont été créés. Les connaissances de la langue générale et une certaine familiarité avec le corpus et le domaine du corpus semblent permettre, en s'appuyant sur ces graphes, de caractériser différentiellement les propriétés attachées à un ensemble de formes et évaluer ainsi leur proximité sémantique[90]. Le contexte d'étude présenté dans [Bouaud *et al.*, 1997*] sur corpus homogène, domaine scientifique clos et langue de spécialité est propice à l'expression de propriétés référentielles attachées aux concepts évoqués. La comparaison des résultats obtenus à des ontologies du domaine amène à limiter la portée des regroupements à des « communautés de sens » locales sans valeur ontologique. Dans notre étude, retrouver un contexte analogue ou proche au corpus d'observation lors des analyses automatiques est justement un but recherché. Moyennant les limitations techniques (analyses syntaxiques nécessiteuses en ressources, dépendances de la langue…) que nous tentons de pallier au maximum, les systèmes type ZELLIG ou SYNTEX semblent adaptés à la tâche d'acquisition des entités lexicales depuis un corpus dans le cadre de LUCIA.

L'utilisateur occupe une place centrale dans toutes les phases d'utilisation du modèle. Cela implique d'envisager non pas des processus automatiques opaques mais des moyens d'assistance transparents et des processus semi-automatiques dont les résultats doivent être validés manuellement. C'est en particulier le cas pour l'acquisition des entités à partir d'un corpus d'observation. Nous nous plaçons donc d'emblée dans l'optique d'une tâche semi-automatique, où les logiciels ont à charge de faciliter le travail de l'usager tout en lui laissant le choix final pour les entités lexicales candidates à catégorisation et description. Dans l'optique d'une *sémantique légère*, nous proposons un logiciel d'étude d'aide à l'acquisition d'entités depuis un corpus non étiqueté et non analysé grammaticale-

[90] Notons que la représentation cartographique de corpus et la spécification des thèmes mis en jeu dans les documents à partir de ressources LUCIA est actuellement à l'étude au laboratoire GREYC [Roy et Beust, 2004].

ment ou syntaxiquement. Le principe est de minimiser les ressources extérieures au corpus qui seraient dépendantes d'une langue ou d'un point de vue grammatical ou syntaxique. Dans [Pincemin, 1999a*], il est en effet rappelé que, même dans ces domaines, le consensus général n'existe pas en TAL (choix des balises lors d'un marquage de corpus, choix du type de relation entre segments créés…). Pour expérimenter nos propositions, nous avons donc créé un prototype de base pour l'aide à l'acquisition de graphies depuis corpus. Ce logiciel minimal est le sujet de la partie 4.2.3, il se nomme MEMLABOR [Perlerin, 2002], il permet d'extraire des graphies redondantes dans un corpus. Après avoir sélectionné et composé ces graphies en entités lexicales, l'utilisateur pourra les affecter à des thèmes et évaluer des associations à l'aide d'un second logiciel : THEMEEDITOR, décrit dans la partie 4.2.4.

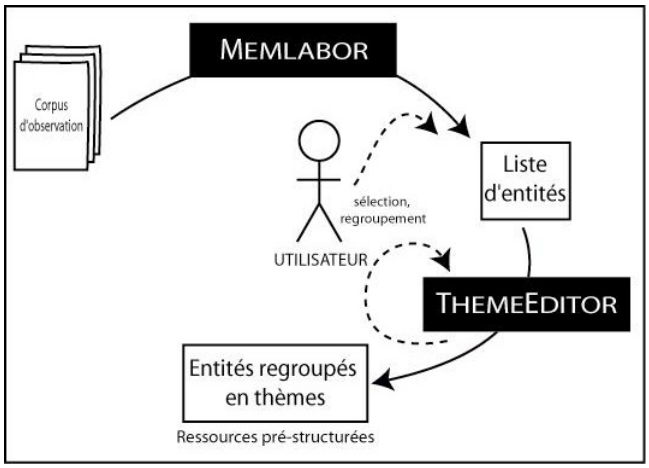

Figure 39 – MEMLABOR **et** THEMEEDITOR **pour l'acquisition et la pré-structuration des ressources.**

4.2.3 MEMLABOR – Logiciel d'aide à l'acquisition

MEMLABOR est une plate-forme informatique dédiée à la création, la gestion et la manipulation de corpus de textes. Ce logiciel d'étude permet de gérer des corpus sous la forme d'une description XML de fichiers de textes ou de fichiers de ressources linguistiques (listes d'exclusion, entités lexicales associées à un domaine). On peut y ajouter des composants (classes, méthodes ou exécutables) tout en conservant les techniques de manipulation fondées sur l'utilisation d'une DTD XML propre[91]. Les textes analysés peuvent être partagés entre plusieurs corpus puisqu'ils ne sont pas modifiés par les traitements et qu'ils sont référencés dans les documents XML par un chemin d'emplacement sur un support de stockage. Si les fichiers sont modifiés, ils sont alors dupliqués et leur version originelle est gardée intacte et toujours référencés dans le fichier XML du corpus. MEMLABOR a été implanté en Java (JDK 1.4.2).

[91] *MEMLABOR* est actuellement en cours d'utilisation par Pierre-Sylvain Luquet du GREYC pour une étude sur la construction d'un modèle de langue phonologique pour le français.

Débutons cette partie par une mise au point terminologique. Nous évoquerons ici les *segments* (ou *tokens*). Les segments sont les éléments obtenus d'un découpage informatisé d'une chaîne linguistique. La règle la plus simple pour une segmentation est celle qui préconise de retenir les éléments se trouvant entre deux espaces (comme « éléments » dans cette même phrase). Certains de ces segments correspondent, lorsque le découpage est efficace, à des *graphies*. Les graphies seront ici des exemplaires particuliers de mots, résultants de la mise en fonctionnement de ces mots dans une unité syntagmatique de langue. Par exemple, « appellons » et « appelons » sont des graphies (même si « appellons » correspond à une faute d'orthographe[92]). La segmentation pour le repérage des graphies d'un corpus constitue généralement la première étape à une exploitation informatique de corpus de textes.

MEMLABOR est un logiciel d'étude qui utilise principalement des méthodes statistiques pour assister l'utilisateur dans l'analyse de différentes distributions de tokens d'un corpus en vue d'une sélection pour une catégorisation et une description ultérieure. Il extrait les tokens les plus redondants d'un corpus. Dans notre étude, MEMLABOR pourrait être remplacé par certains logiciels cités dans la partie précédente et dans la partie 2.1 du chapitre 2, attendu qu'il faudrait alors se conformer aux contraintes prévues (étiquetage du corpus et/ou utilisation de ressources lexicales prédéfinies, etc.).

Un corpus créé à partir de MEMLABOR est représenté par un fichier XML soumis à une DTD permettant un stockage d'information du type proposé en figure 40. Cette DTD définit un document de type `CORPUS` déterminé par un nom, une suite de commentaires et une date de création (figure 40 - ligne 1). Chaque `CORPUS` est composé d'éléments de type `FICHIER_CORPUS` qui correspondent aux documents constituant le corpus. Ces derniers sont définis par un `nom` (emplacement sur un support de stockage) (figure 40- ligne 3 : un travail de type HTMLtoTXT correspondant à la transformation d'un document HTML au format TXT a été effectué sur le fichier du corpus `.\Corpus\LibMS\Reprise des hostilités contre Microsoft.htm`). Chaque document du corpus (de type `FICHIER_CORPUS`) peut être soumis à un travail particulier dont la trace sera sauvegardée dans le fichier XML du corpus selon le format défini pour les éléments de type `TRAVAIL_FICHIER` (figure 40- ligne 4 :) où les attributs `Nom`, `Result` et `Date` correspondent respectivement au type du travail correspondant, au fichier où sont stockés les résultats de ce travail et à la date de réalisation de ce travail. On peut ajouter à un élément `CORPUS` des éléments de type `TRAVAIL_CORPUS` qui correspondent à des travaux effectués sur l'ensemble des fichiers du corpus. Ils sont définis par un nom (le type de travail effectué sur le

[92] En vertu des critères de sélection des entités lexicales proposés dans le chapitre précédent, des entités lexicales présentant une faute d'orthographe peuvent être retenues en tant que ressource du système. En effet, certaines d'entre elles sont si courantes qu'il est difficile de ne pas en tenir compte si les éléments de signification auxquelles ont peu les associer sont intéressants pour l'étude. Nous pensons par exemple à l'expression « *au temps pour moi* » majoritairement écrite « *autant pour moi* » ou à des mots comme *connexion*, *langage* ou *danse* qui subissent l'influence anglo-saxonne et se retrouvent fréquemment sous la forme de *dance*, *connection* et *language*. Il s'agit ici d'un argument supplémentaire pour l'utilisation de ressources qui ne sont pas construites *a priori*.

corpus), un fichier où sont stockés les résultats de ce travail (emplacement sur un support de stockage) et une date (celle à laquelle le travail a été effectué sur le corpus) (figure 40 - ligne 2 : un travail de type Zipf – voir 4.2.3 - a été effectué sur l'ensemble des fichiers du corpus). Pour plus de détails sur la DTD du logiciel, on pourra se reporter à [Perlerin, 2002*].

```
1. <CORPUS   Nom="Microsoft_Libe"
          Commentaires="Procès Microsoft 1997-2001"
          Date="12/02/2002">
2. <TRAVAIL_CORPUS
          Nom="Zipf"
          Result=".\Corpus\LibMS\Microsoft_Libe.zipf.xml"
          Date="14/02/2002"/>
3.     <FICHIER_CORPUS Nom=".\Corpus\LibMS\Reprise des hostilités contre
       Microsoft.htm">
4.        <TRAVAIL_FICHIER
                    Nom="HTMLtoTXT"
                    Result=".\Corpus\LibMS\TXT\Reprise des hostilités contre
                    Microsoft.txt" Date="13/02/2002"/>
5.     </FICHIER_CORPUS>0
6. </CORPUS>
```

Figure 40 - Exemple de corpus codé pour MEMLABOR**.**

Dans MEMLABOR, les paliers d'observations sont la graphie, le paragraphe, le texte et le corpus. Dans sa dernière version (v.0.2 de novembre 2003), le logiciel propose les fonctionnalités suivantes :

- La normalisation de documents HTML[93] (pour pallier les éventuelles erreurs de syntaxe des fichiers). Les navigateurs utilisés pour visionner les fichiers HTML sont très permissifs : ils font abstraction des erreurs de syntaxe. Un fichier HTML peut être non-conforme aux recommandations du langage sans que les résultats observés à l'écran s'en ressentent. Nombre d'utilisateurs utilisent ainsi un pseudo-HTML qui satisfait l'œil mais pas les travaux de TAL en particulier lorsqu'il s'agit de traiter automatique de la dimension organisationnelle des documents (titres, sous-titres, listes, paragraphes, etc.). Si nous perdons ainsi une certaine dimension sémiotique des documents HTML, nous gagnons la possibilité de traiter rigoureusement la mise en forme en parties et paragraphes.
- La transformation de documents HTML d'un corpus au format TXT. Il s'agit principalement de créer un fichier TXT présentant le même texte que celui du document HTML mis en forme

[93] *MEMLABOR* exploite l'API Java du projet JTidy (http://lempinen.net/sami/jtidy/) qui permet la vérification et la correction de la syntaxe de documents HTML.

avec des passages à la ligne entre parties repérées. La normalisation préalable des documents HTML permet de procéder à une transformation rigoureuse au format TXT.

- La segmentation de fichiers d'un corpus (repérage des graphies à l'aide d'un *StreamTokenizer*[94] – segmenteur sur flux de données) où sont précisés les caractères délimiteurs et les caractères potentiels pour inscrire les graphies.
- La segmentation en paragraphes de fichiers d'un corpus (repérage des caractères délimitant les paragraphes dans les fichiers et constructions de fichiers correspondant à chaque paragraphe).

MEMLABOR permet également un calcul de type Zipf qui a présenté un grand intérêt dans le cadre de nos travaux. Le logiciel extrait des corpus analysés des tokens que l'utilisateur a à charge de les composer éventuellement en entités lexicales en fonction de leurs cotextes. Certaines techniques « d'extraction de termes » du TAL utilisent la méthode dite des cooccurrences [Smadja, 1992*]. Des statistiques sur la proximité syntagmatique de certains termes sont utilisés pour statuer sur leur composition automatique en termes composés. MEMLABOR propose un calcul de cooccurrences dans des entités syntagmatiques définies (document du corpus ou paragraphe de documents) pour à la fois assister l'utilisateur dans la tâche de composition des termes mais aussi éventuellement pour l'aider à former certaines catégories sémantiques. La relative petite taille des corpus nécessaires à l'élaboration d'un dispositif ne nous a pas amenés à utiliser cette fonctionnalité lors de notre expérimentation, nous nous abstiendrons donc de décrire ces fonctionnalités ici.

En 1935, le linguiste de Harvard Georges Kingsley Zipf vérifie manuellement que pour un corpus donné[95], la fréquence (F) d'un terme est inversement proportionnelle à son rang (R) et formalise ainsi la loi qui porte désormais son nom F x R = *c* [Zipf, 1949]. Cette loi se vérifie sur beaucoup de données linguistiques observables comme les règles syntaxiques appliquées sur un corpus par exemple [Gaizauskas, 1995] et dans beaucoup d'autres domaines : la répartition de la population des villes d'un état [Hill, 1970] ou les notes de partitions musicales [Zanetten, 2004] pour ne citer que ces références. Lorsqu'on classe par ordre décroissant de fréquence les termes d'un corpus, la loi de Zipf implique que la fréquence des termes décroît de manière exponentielle et ce quelle que soit la langue du corpus. Une représentation en histogrammes avec en abscisse le rang R et en ordonnée la fréquence F des termes rangés par ordre décroissant a une allure logarithmique (figure 42). Les mêmes couples R/F tracés dans un repère bilogarithmique forment un nuage de points approximativement linéaire dès lors que le corpus étudié est conséquent. Luhn [Luhn, 1958] s'est fondé sur ce type de calcul pour des travaux relatifs à l'indexation automatique de textes où l'enjeu est de repérer automatiquement les

[94] Classe Java du package java.io.

[95] Initialement, Zipf a utilisé le roman de Charles Dickens (1812-1870), *David Copperfield* mais la loi qui l'a formulée du comptage des mots de cette œuvre a été généralisée par la suite, en particulier à partir des années 1980 grâce l'utilisation des ordinateurs [Ha, Sicilia *et al.*, 2002 : 316].

termes d'un texte les plus discriminants par rapport à d'autres textes et les plus significatifs du texte analysé. L'auteur propose deux rangs critiques (*r1* et *r2* sur la figure 41) dont le positionnement dépend en particulier de la nature du texte considéré. Les termes au-delà du rang *r1* sont considérés comme trop redondants et ceux en deçà du rang *r2* sont considérés comme ne contribuant pas de manière suffisamment significative au contenu du texte.

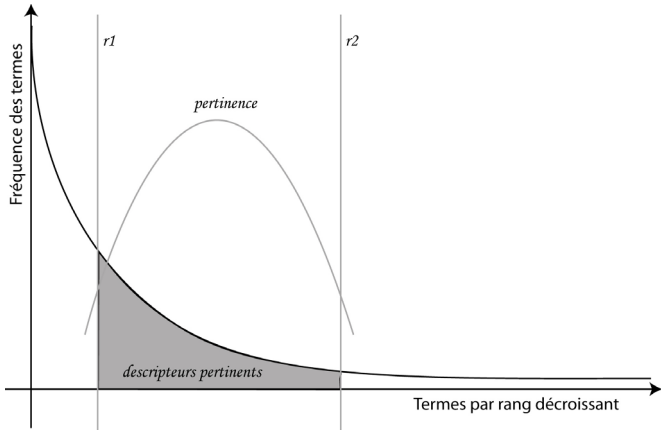

Figure 41 – Lois de Zipf et Luhn

Dans [Giguet, 1998], l'auteur affirme que l'étude de la liste des fréquences des termes issue d'un calcul de type Zipf amène à repérer approximativement deux sous listes distinctes et contiguës : la première regroupe essentiellement les termes grammaticaux (les termes à gauche du *r1* sur la figure 41), la seconde les termes « lexicaux thématisants » du corpus (les termes entre *r1* et *r2* sur la figure 41) et rejoint ainsi Luhn dans ses conclusions au niveau d'un corpus dans son ensemble. Confronté à l'étude, on peut constater que la détection de ces sous-listes s'avère très dépendante du type de corpus analysé, en particulier de son caractère « homogène », alors que la langue du corpus n'entre pas en compte. Ces deux listes sont difficilement repérables en tant que telles de façon rigoureuse à l'aide de critères formels. Un corpus de textes de même genre et traitant d'un même domaine fera ainsi généralement une distinction plus nette entre ces sous listes qu'un corpus de textes de genres différents. À l'usage, on remarque en particulier que si la première sous liste recèle bien une majorité de mots grammaticaux, il est difficile de statuer sur le caractère « thématisant » des éléments de la seconde qui regroupe, il est vrai, des termes non grammaticaux mais peut également présenter de nombreux mots grammaticaux, adverbes ou adjectifs. Par ailleurs, il est parfois difficile d'affirmer *a priori* que les termes trouvés font référence directement à un thème donné si l'on ne les examine pas dans le contexte du corpus – la SI nous invite à nous méfier d'une relation stricte terme / thème.

Rapportons les résultats obtenus d'un calcul de type Zipf sur un corpus constitué de 58 articles du journal Libération entre décembre 1997 et août 2001[96] (figure 42), tous relatifs au procès anti-monopole opposant l'entreprise Microsoft et l'état états-unien représenté par le juge T.P. Jackson. Ce corpus représente un total de 38 079 tokens parmi lesquelles on trouve 5 293 graphies différentes. Nous appellerons désormais ce corpus le « corpus Microsoft ».

- Le token le plus fréquent est *de* avec 2176 occurrences – rang 1, il est suivi par *le* (1040 occurrences, rang 2), *la* (997 occurrences, rang 3), *l* (821 occurrences, rang 4) puis *à* (712 occurrences, rang 5).
- Les premiers tokens (en terme de fréquence) pouvant faire référence à un procès (l'un des thèmes majeurs du corpus) se trouvent entre le rang 41 (*procès*) et le rang 224 (*jugement*). On peut noter que le token *libération* n'est pas retenu pour ce thème, attendu qu'il apparaît non seulement au moins une fois dans chaque texte du fait de la source des articles mais également à l'intérieur des articles pour des références internes au journal.
- Les premiers tokens (en terme de fréquence) pouvant faire référence à l'informatique (un autre thème majeur du corpus) se trouvent entre le rang 6 (*microsoft*) et le rang 2511 (*java*).
- Les hapax (tokens de fréquence 1) représentent 52% de l'ensemble, les tokens de fréquence 2 représentent 18% (figure 43).

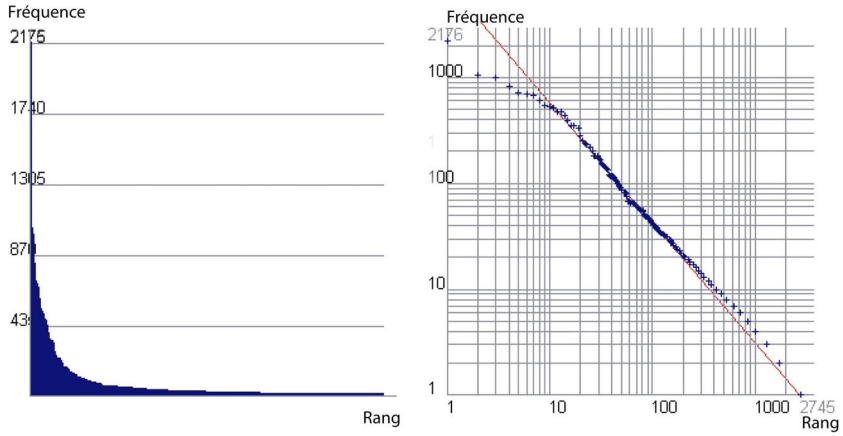

Figure 42 – Représentations graphiques des résultats d'un calcul Zipf sur le corpus Microsoft.

[96] La liste des graphies repérées n'a pas été reproduites car seule celle proposée en figure 45, p.137 effectuée sur le même corpus mais avec des ressources supplémentaires est véritablement intéressante.

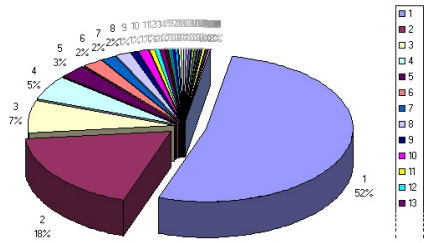

Figure 43 – Proportions des tokens en fonction de leur fréquence (sans liste d'exclusion).

L'examen des résultats obtenus fait apparaître que les tokens grammaticaux se retrouvent en majorité dans la première partie de la liste de tokens rangée par ordre décroissant de fréquence mais que l'on peut en trouver en fait, tout le long de cette liste. Les termes « thématisants » (au sens de Giguet [*ibid.*]) se trouvent disséminés tout le long de la liste et il est souvent difficile de statuer *a priori* quant à leur appartenance à un thème précis indépendamment de leur présence dans un cotexte (le cas le plus flagrant étant celui du token *libération* mais c'est aussi le cas pour des tokens comme *correction*, *avis, compétition, java* etc.). La loi de Zipf est une loi de tendance : les régularités qu'elle met en exergue sont d'autant plus vérifiées que le corpus est important et divers. Muller [Muller, 1971] (cité dans [Lainé-Cruzel, 2001 : 54]) a ainsi travaillé sur la fréquence des hapax et a montré que leur nombre – *lorsque le corpus atteint une taille suffisante* – s'accroît proportionnellement à la taille du corpus étudié. Dans les cadres applicatifs que nous avons choisis, les corpus d'observation éventuellement utilisés ne sont pas nécessairement très importants en terme de nombre de textes ou nombre de mots - c'est à l'utilisateur de les rassembler et il ne semble pas raisonnable de devoir avoir nécessairement en sa possession plusieurs centaines de documents électroniques pour faire fonctionner le système. Il est donc primordial de relativiser les opérations automatiques que l'on peut mettre en place à partir de ce type de calcul. De la première expérience sur le corpus Microsoft, nous pouvons conclure qu'utilisée pour une assistance à l'extraction d'entités lexicales répondants aux critères présentés en début de ce chapitre, ce type de calcul peut être fortement optimisé par l'utilisation d'une liste d'exclusion. À l'aide d'une liste d'exclusion, les mots difficilement associables à des thèmes comme les mots grammaticaux bien sûr mais également les formes des auxiliaires, certains adverbes et adjectifs courants peuvent être supprimés. Le caractère « thématisant » de certains tokens n'est pas d'accès trivial. Nous le récusons d'ailleurs préférant approcher les *thèmes* par la récurrence d'attributs ou attributs/valeurs plutôt que de les considérés inhérents *a priori* à certaines entités lexicales et elles-seules.

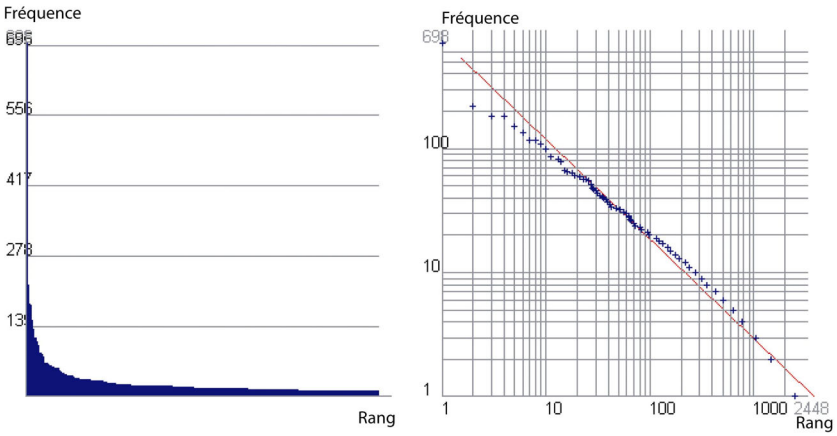

Figure 44 – Représentations en histogramme et dans un repère bilogarithmique des calculs de type Zipf sur le corpus Microsoft avec liste d'exclusion.

Pour MEMLABOR, nous avons donc choisi d'utiliser une liste d'exclusion. La liste d'exclusion pour le français actuellement proposée avec MEMLABOR est composée de 781 formes, représente 6,1Ko de mémoire et contient essentiellement des déterminants, pronoms, chiffres, auxiliaires et adverbes courants du français ainsi que quelques adjectifs. Elle a été élaborée à partir de celle proposée en ligne par Jean Véronis[97] de l'université d'Aix-en-Provence et complétée lors d'utilisations de notre logiciel. Il s'est en effet avéré que cette liste souffrait de certains manques car, comme nous verrons plus loin, de telles listes doivent parfois subir des modifications en fonction des corpus traités. La figure 45 p.137 présente un extrait des résultats obtenus d'un calcul de type Zipf avec une liste d'exclusion sur le « corpus Microsoft ». L'examen de cette liste permet d'apprécier le filtrage dû à la liste d'exclusion. On y retrouve les noms des principaux protagonistes de l'affaire (*Bill Gates*, le juge *Thomas Penfield Jackson*, l'entreprise *Microsoft* et son PDG *Steve Balmer*, le logiciel *Explorer*…), les lieux (*Washington*, *américain*, *New York* – composé par les deux termes correspondants, *Etats-Unis*…), des termes pouvant avoir trait au monde de l'entreprise (*compagnie*, *firme*, *entreprise*…) ainsi qu'un certain nombre pouvant avoir trait au monde de la justice (*procès*, *justice*, *appel*, *verdict*…). En étudiant ces termes dans le cotexte du corpus, on peut apprécier les significations mises en jeu dans leurs usages précis. Les représentations graphiques associées à ce calcul laissent apparaître que l'aspects des courbes aux deux échelles est identique avec ou sans l'utilisation de la liste d'exclusion (figure 44, p.136).

[97] http://www.up.univ-mrs.fr/~veronis/data/antidico.txt

1	microsoft	698	26	marché	44	37	actions	30	45	rousselot	22	48	première	19
2	gates	220	26	département	44	38	temps	29	45	accès	22	48	services	19
3	son	182	27	monde	42	39	york	28	45	jours	22	48	administration	19
4	bill	180	27	nouveau	42	39	entreprises	28	46	parties	21	49	sun	18
5	juge	150	28	mois	41	40	steve	27	46	concurrent	21	49	informations	18
6	windows	134	29	américaine	40	40	président	27	46	employés	21	49	effet	18
7	libération	117	29	ans	40	41	début	26	46	action	21	49	pouvoir	18
8	procès	116	30	verdict	39	41	ordinateurs	26	46	fédéral	21	49	1999	18
9	jackson	108	31	société	37	41	devrait	26	47	nouvelles	20	49	cas	18
10	gouvernement	100	31	américain	37	41	multimédia	26	47	juin	20	49	place	18
11	justice	87	32	pourrait	35	42	web	25	47	conclusions	20	49	exemple	18
11	logiciels	87	32	industrie	35	43	mardi	24	47	jugement	20	49	cours	18
12	internet	83	33	explorer	34	43	novembre	24	47	mettre	20	49	perdu	18
13	compagnie	79	33	redmond	34	43	prix	24	47	grand	20	50	téléphone	17
14	entreprise	66	33	oracle	34	43	igi	24	47	mai	20	50	dernière	17
15	démantèlement	65	33	ballmer	34	43	systèmes	24	47	etats	20	50	valeurs	17
15	firme	65	33	accord	34	43	affaire	24	47	linux	20	50	mercredi	17
16	informatique	63	34	dernier	33	43	consommateurs	24	47	ellison	20	50	dossier	17
17	washington	60	34	concurrents	33	43	concurrence	24	47	netscape	20	50	dit	17
17	exploitation	60	34	groupe	33	43	vendredi	24	47	seattle	20	50	années	17
18	nouvelle	59	34	décision	33	44	fin	23	47	secteur	20	50	avocat	17
18	appel	59	35	pratiques	32	44	semaine	23	47	stratégie	20	50	amiable	17
19	monopole	57	35	pc	32	44	an	23	47	street	20	50	penfield	17
20	système	56	35	produits	32	45	fois	22	47	wall	20	50	terme	17
20	sommaire	56	35	thomas	32	45	avril	22	47	1998	20	50	proposition	17
21	cour	55	36	géant	31	45	utilisateurs	22	47	solution	20	50	avocats	17
22	2000	51	36	new	31	45	suprême	22	48	états-unis	19	50	reste	17
23	logiciel	48	36	économie	31	45	année	22	48	procédure	19			
24	dollars	47	36	peut	31	45	sanctions	22	48	état	19			
25	antitrust	46	37	loi	30	45	jour	22	48	position	19			

Figure 45 – Premiers tokens de la liste de type Zipf obtenue sur le corpus Microsoft[98] (rang, token, fréquence).

Par rapport aux résultat présentés en figure 43 (p. 135) , on peut voir que le proportion de mots qui ne présentent qu'une seule occurrence est plus importante lorsqu'on utilise une liste d'exclusion : la proportion des hapax gagne 2 points (figure 46).

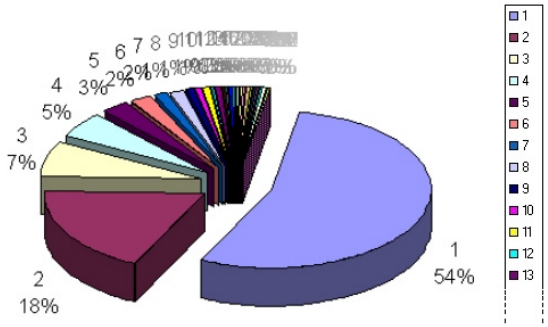

Figure 46 - Proportions des tokens en fonction de leur fréquence (avec liste d'exclusion).

[98] IGI (rang 43) correspond à l'acronyme *Investigative Group International* qui désigne une agence de détectives privés internationale engagée par l'entreprise Oracle (rang 33) au cours de cette affaire. Rousselot (rang 45) est le nom du principal journaliste de Libération relatant le procès en question.

Dans la figure 47 p.137, nous avons modifié manuellement les résultats du calcul sur le « corpus Microsoft » additionnant les fréquences des tokens relevant *a priori* d'un même paradigme flexionnel (l'*a priori* provient essentiellement du fait que nous n'avons pas eu recours à une analyse grammaticale du corpus) et celles des tokens pouvant être dérivés d'une même pseudo-racine sur le principe du *stemming*. Les résultats obtenus sur les 28 premiers tokens de la liste initiale amènent à constater des modifications sensibles quant à l'agencement des tokens présentés par ordre décroissant de fréquence d'apparition. Il n'y a rien d'étonnant à constater alors que le rang des noms propres a tendance à décroître avec la prise en considération de ces critères tandis que le rang des adjectifs et des verbes subit l'effet inverse. Des travaux ont montré que le paradigme flexionnel n'affecte pas profondément les résultats obtenus d'un calcul de type Zipf : il y a une corrélation forte entre les lemmes les plus représentés et les mots les plus représentés [Gelbukh et Sidorov, 2001] . Seule une étude a plus grande échelle permettrait de valider une hypothèse quelconque sur l'utilisation du *stemming* ou la prise en considération du paradigme flexionnel pour un calcul de type Zipf. Les résultats de cette première expérience (limitée) pour ce corpus ne font cependant pas apparaître de différences notoires notamment vus nos objectifs d'utilisation de telles listes.

Token initial	F	Tokens du même paradigme flexionnel	Nv F	Tokens à partir d'une pseudo-racine ou d'une orthographe avoisinante	Nv F
microsoft	698		=	MSN (5), MS (1)	704
gates	220		=		220
juge	150	juges (9)	159	jugement (20), jugeant (3), juger (2), jugements (1)	185
bill	180		=		180
informatique	63	Informatiques (69)	132	informations (18), information (10), non-informaticien (1), informait (1), informateurs (1), informaticiens (1)	150
procès	116		=	procédure (19), procédures (5), procureurs (1), procureur (1)	142
logiciels	87	Logiciel (48)	135		135
Windows	134		=		134
nouvelle	59	nouveau (42), nouvelles (20), nouveaux (6), nouvel (4)	131	nouveautés (2)	133
justice	100		=	injuste (3), justifier (2), justes (1), judiciaire (10), judiciaires (6), préjudiciable (1)	123
libération	117		=	délibération (1)	118
jackson	108		=	jackson (1)	109
gouvernement	100		=	gouvernementaux (3), gouvernementales (2), gouvernementale (1)	106
entreprise	66	entreprises (33)	99	entrepris (1)	100
compagnie	79	compagnies (12)	91		91
internet	83		=	internautes (2), internaute (1)	86
appel	59	appelé (13), appelle (3), appeler (2), appellent (1)	78	rappelé (3), rappeler (1), rappellent (1), rappelle (1)	84
cour	55	cours (18)	73	recours (5), encourt (2), recourir (1)	81
système	56	système (24)	80		80
monopole	57	monopoles (4)	61	monopolistiques (6), monopolistes (3), monopolistique (3), monopoliser (2), quasi-monopole (1)	76
démantèlement	65	démantèlements (1)	65	démanteler (7)	72
exploitation	60		=	exploiter (4)	64
marché	44	marchés (14)	58	démarche (2)	60
washington	60		=		60
antitrust	46	antitrust (2), anti-trust (7)	55	trusts (1)	56
monde	42		=	cybermonde (1), mondial (8), mondiale (1)	52
2000	51		=		51
dollars	47		=		47

Figure 47 – Nouveau calcul de type Zipf avec addition des tokens d'un même paradigme flexionnel (3e et 4e colonnes) et addition des tokens dérivant d'une même pseudo-racine ou intégrant une faute d'orthographe (5e et 6e colonnes). Nv F signifie nouvelle fréquence.

Il est important de noter ici que MEMLABOR est un logiciel multilingue en tant que les principes mis en jeu dans son fonctionnement sont indépendants des langues manipulées même si certaines ressources fournies en entrée sont dépendantes de la langue du corpus étudié. On pourrait ainsi considérer le logiciel comme alingue et ses ressources comme multilingues. En ce qui concerne ce logiciel, les seules ressources dépendantes des langues sont les listes d'exclusions et l'on en trouve en abondance sur l'Internet pour de nombreuses langues[99]. MEMLABOR permet de créer des listes d'exclusions singulières en fonction des corpus analysés car ces listes doivent parfois être adaptées aux corpus. MEMLABOR a par exemple été utilisé dans le cadre d'un projet mené par Vincent Rioux de l'IRCAM relatif à la description verbale de sources sonores et musicales[100] - *son* et *ton* ont donc été retirés de la *stop-list* par défaut. Il s'agissait d'utiliser notre logiciel pour une première approche d'un corpus de portraits verbaux d'échantillons sonores (projet ECRIN). Une liste d'exclusion particulière a été mise au point pour cette expérience. Les graphies *son* et *ton* étaient par exemple présentes dans la liste d'exclusion proposée par défaut pour le français avec le logiciel. *Son*, non plus considéré comme relatif à l'adjectif possessif masculin mais au nom homographe s'avère pourtant essentiel à prendre en considération lorsqu'on s'intéresse à la description d'échantillons sonores[101]. L'homographie est un phénomène très répandu dans les langues mais il n'est pas un frein à la mise en place de nos propositions même si certaines graphies peuvent être supprimées à tort lors de calculs de type Zipf du fait de leur présence dans la liste d'exclusion. C'est le principe de récurrence et le cycle itératif d'utilisation du système qui nous permet de ne pas prendre en considération strictement l'homographie.

Pour une illustration du multilinguisme de la loi de Zipf en même temps que l'interface principale de MEMLABOR, on peut apprécier une partie des résultats obtenus sur *Alice in Wonderland's* de L. Caroll[102] (1832-1898) dans la figure 48. La loi de Zipf se vérifie sur de nombreuses langues (naturelles), seule la constante *c* de la formule F x R = *c* est modifiée (voir par exemple, [Ha *et al.*, 2002*] pour une comparaison entre l'anglais et le chinois mandarin et [Gelbukh et Sidorov, 2001*] pour une comparaison entre l'anglais et le russe).

[99] A l'adresse http://www.unine.ch/info/clef/ (consultée le 1e octobre 2003), on trouve des listes d'exclusion directement utilisables par *MEMLABOR* (car au format TXT) pour le français, l'anglais, le russe, l'allemand, l'espagnol, l'italien, le finnois, le suédois et l'arabe.

[100] http://www.ircam.fr/equipes/analyse-synthese/rioux/ projets CUIDADO et ECRIN. Faute de financement, le projet conjoint avec l'équipe de l'IRCAM a avorté et n'a donc pas fait l'objet de publications.

[101] La figure 45 présente ainsi à dessein *son* au troisième rang alors que dans ce corpus, cette graphie aurait dû être exclue.

[102] Mis à disposition dans le cadre du projet Gutenberg - http://swww-2.cs.cmu.edu/People/rgs/alice-table.html

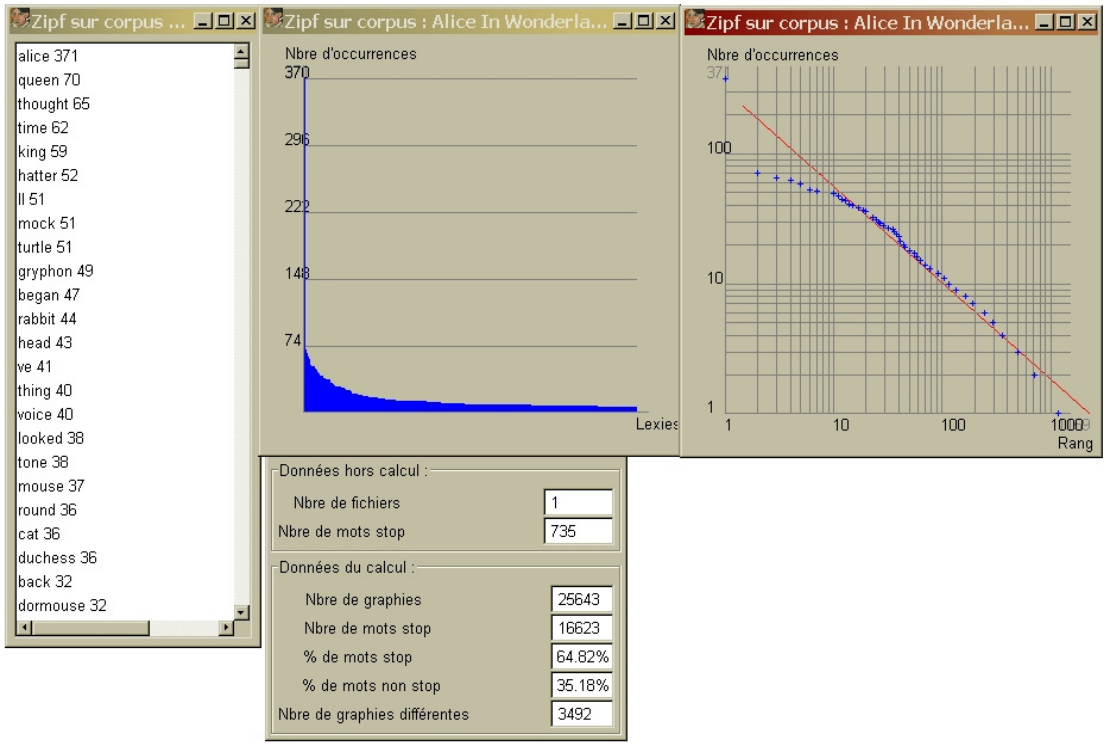

Figure 48 – Copie d'écran de MEMLABOR **: Calcul de type Zipf sur *Alice in Wonderland's*[103]**

Dans cette partie, nous avons vu comment était utilisés la loi de Zipf dans le logiciel MEMLABOR : les résultats obtenus permettent de présenter à l'utilisateur des tokens remarquables du point de vue de leur fréquence hors mots de la liste d'exclusion. Dans les faits, **ce n'est pas tant** la loi de Zipf en tant que telle que le calcul des occurrences et la présentation des résultats par ordre décroissant qui sont exploités. Pour faciliter l'exploration manuelle d'une telle liste, les tokens sont présentés dans l'ordre décroissant d'occurrences. Nous verrons plus loin que l'étape d'amorce que constitue l'acquisition de termes depuis un corpus d'observation peut s'effectuer à l'aide de quelques tokens (de l'ordre de la trentaine). La présentation d'une telle liste s'avère suffisante pour récupérer et composer des entités lexicales à partir d'un corpus d'observation et amorcer le processus d'utilisation de LUCIA si l'usager a déjà une bonne connaissance de son ou ses domaines d'intérêt. C'est en utilisant ce logiciel que nous avons débuté nos propres expériences d'évaluation. Cependant, nous pensons que MEMLABOR peut s'avérer très limité lorsque le corpus d'observation est important ou que le domaine d'intérêt de la tâche n'est pas maîtrisé par l'usager, en particulier pour la composition d'entités lexicales à partir de graphies. Dans ce cas, il serait préférable d'utiliser d'autres propositions plus efficientes pour cette étape. De même, pour l'étude d'un fait de langue particulier, la fréquence des termes n'est pas nécessairement le critère primordial à leur sélection pour faire l'objet d'une description. C'est

[103] Les vues proposées pour le calcul de type Zipf ont été inspirées par le logiciel en ligne d'Emmanuel Giguet – users.info.unicaen.fr/~giguet/java/zipf.html.

l'expertise de l'utilisateur qui doit alors prendre le pas sur les propositions issues de calculs statistiques.

MEMLABOR constitue une aide minimale pour l'acquisition d'entités lexicales à partir d'un corpus d'observation. Il permet de proposer à l'utilisateur des listes d'entités redondantes et potentiellement intéressantes pour la tâche. Parce que ces entités sont susceptibles de répondre aux exigences énoncées dans le chapitre 3 (partie 3.1), elles peuvent faire l'objet d'une sélection pour la constitution d'un dispositif. Or, et cela a déjà été évoqué, les thèmes ne sont pas donnés dans les textes mais à découvrir, la signification d'une entité dépend de son cotexte. Il peut s'avérer ainsi difficile de juger de la pertinence d'un ensemble d'entités lexicales acquises depuis un corpus d'observation par rapport à une tâche et/ou à un domaine sans une analyse de leur cotexte d'apparition. La question à se poser est alors la suivante : *Les entités sélectionnées sont-elles interprétables en cotexte comme ayant trait au(x) domaine(s) de ma tâche ?* Il est également difficile de juger du taux de recouvrement d'un ensemble d'entités sélectionnées pour une tâche par rapport aux entités présentes dans l'ensemble des textes d'un corpus d'observation. Si le corpus d'observation est thématiquement homogène et que le ou les thèmes abordé(s) dans les textes du corpus sont ceux de la structuration des ressources (i.e. de la tâche), il peut être intéressant d'évaluer ce taux. Ce dernier donnera une indication sur la répartition des entités lexicales sur l'ensemble du corpus : les analyses qui seront effectuées ultérieurement se baseront en particulier sur la redondance des entités au sein des textes analysés. En d'autres termes, l'autre question à se poser alors est : *Les entités sélectionnées sont-elles en nombre important par rapport à celles appartenant au(x) domaine(s) de ma tâche dans le corpus d'observation et par rapport aux autres entités du corpus ?* Dans cette partie, nous évoquerons une étape non obligatoire pour la constitution de ressources LUCIA : une première évaluation qualitative et quantitative des entités issues d'un corpus d'observation. Cette étape est facultative pour deux raisons principales. D'une part, l'utilisation d'un corpus d'observation l'est également. D'autre part, cette étape peut être effectuée lors de l'évaluation des résultats d'analyses dans le processus itératif d'utilisation des dispositifs (c.f. Chapitre 5). Elle peut s'avérer utile pour réduire le nombre d'itérations du processus jusqu'à des ressources satisfaisantes.

4.2.4 Première évaluation du lexique : THEMEEDITOR

Les différents outils informatiques proposés s'appuient principalement sur une spécificité des langues dites « naturelles » par opposition aux langages formels : la redondance dans le discours. Cette redondance apparaît au niveau morphosyntaxique avec, par exemple, les cas d'accords en nombre et en genre qui répètent plusieurs fois la même marque : l'*isosémie*. Elle peut être mise au jour également au niveau sémantique avec l'isotopie. Repérer les thématiques évoquées par un texte consiste à y retrouver une ou plusieurs isotopies [Rastier, 2001b* : 11] : celles qui parcourent le texte dans son en-

semble. Les signes qui composent les textes n'étant pas *a priori* dédiés à un thème, c'est une interprétation du texte qui permet d'identifier les signifiants et de les associer contextuellement à des signifiés pertinents. Cette association est également permise par certaines dimensions sociales de la langue. Une langue de spécialité peut par exemple favoriser voire systématiser certaines de ces associations. Pour les textes tout-venant, seul un retour sur corpus et les connaissances – et le point de vue – de l'utilisateur peuvent apporter une réponse définitive à la question : *Les entités sélectionnées sont-elles interprétables comme faisant référence au(x) domaine(s) de ma tâche ?*. On peut cependant assister cet examen en facilitant l'accès aux parties de textes présentant les entités sélectionnées et en proposant des outils pour entamer une structuration des entités.

Désirant limiter la quantité de données utilisées et octroyer une place centrale à l'interprétation et aux connaissances de l'utilisateur, nous avons choisi d'utiliser le principe de redondance évoqué ci-dessus accompagné de calculs statistiques pour mettre en place une assistance à cette tâche. Si l'on utilisait des entités considérées *a priori* comme appartenant à ces domaines (dans des dictionnaires par exemple), l'approche centrée utilisateur et la *sémantique légère* s'en trouveraient altérés. Les instrumentations proposées pour les deux tâches de validation s'appuient sur le coloriage et l'analyse distributionnelle des textes du corpus d'observation. Le principe de coloriage consiste à affecter une couleur à un ensemble d'entités lexicales susceptibles de partager des éléments de significations. Chaque entité peut être le support d'une ou plusieurs isotopies dans un texte. Le coloriage permet la visualisation de ces isotopies potentielles, l'examen de leur répartition, leurs alternances et enchaînement au long d'un texte et l'évaluation de leur pertinence. Il permet également d'apprécier grossièrement l'importance des isotopies. Ce repérage constitue une aide pour l'objectivation de ces isotopies et la découverte des sèmes mis en jeu.

Affecter une couleur à un ensemble d'entités peut s'accompagner d'une première classification en thèmes. Le regroupement en thème et sous-thèmes des entités lexicales acquises à ce stade constitue une pré-structuration de ces ressources qui amorce la catégorisation LUCIA faite ensuite. C'est ce que nous nous proposons de décrire ici avec l'exposition du logiciel THEMEEDITOR.

L'utilisation de la couleur est justifiée principalement par les lois structurales énoncées par certains psychologues de la forme [Cohen, 2000] et en particulier la *loi de la similitude* fondée sur les travaux allemands de la *Gestalt-theorie*. Cette loi stipule que l'œil regroupe dans le champ visuel des éléments qui présentent entre autres, des caractéristiques communes de luminance. La luminance est ici considérée comme la troisième variable perçue simultanément à l'écran ; les deux premières étant celles du plan de l'écran (abscisse et ordonnée). La luminance exploitée dans THEMEEDITOR correspond à la couleur attribuée à un thème (en plus bien sûr, de la couleur noire utilisée pour inscrire le

texte à l'écran). L'attribution d'une même couleur à un ensemble d'entités et la coloration[104] de ces entités à l'intérieur d'un corpus permettent de mettre en évidence leur caractéristique commune : appartenant à un même thème pour le lecteur/utilisateur, elles sont potentiellement support des mêmes isotopies (figure 49). Le coloriage est ainsi une méthode pour rendre inter-objectifs (partageables) certains aspects fondamentaux des interprétations que l'on peut produire.

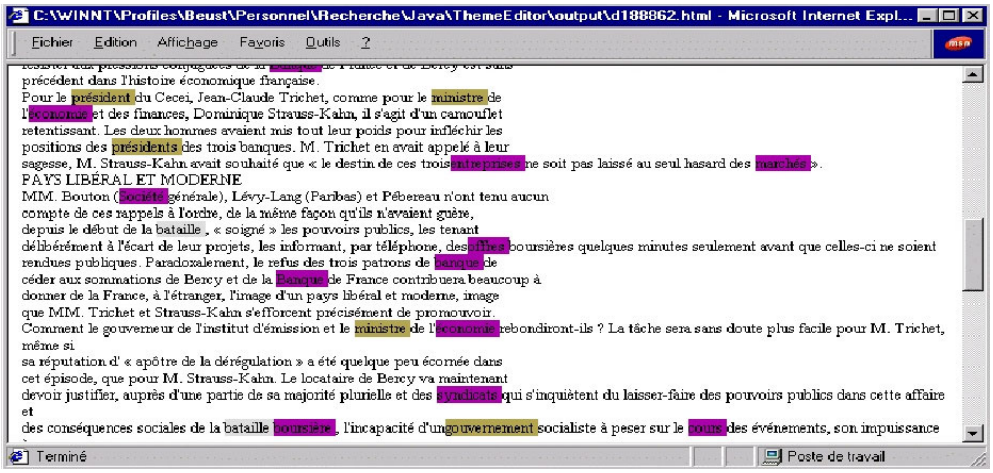

Figure 49 – Fichier HTML généré par *THEMEEDITOR* : texte coloré.

La coloration des entités appuyée par une analyse distributionnelle forme donc une aide à une première évaluation des entités. Ces deux aspects sont instrumentalisés par le logiciel THEMEEDITOR [Beust, 2002]. THEMEEDITOR est un logiciel d'étude de la dimension thématique des langues naturelles. Il a pour but de tester s'il est possible de mettre à profit la cohésion textuelle pour aider un utilisateur à construire de façon incrémentale des listes de mots associables à des thèmes dans le contexte d'un corpus donné. Il permet de produire des corpus automatiquement annotés en fonction des thèmes qui lui sont associés. Ces corpus sont produits dans un format XML de sorte à pouvoir être utilisés par d'autres applications – en particulier à l'intérieur de la chaîne de traitement que nous proposons. Il a été réalisé en Java par deux étudiants en maîtrise d'informatique à l'Université de Caen sous la direction de Pierre Beust en 2000 et modifié ultérieurement par nos soins. THEMEEDITOR procède à un coloriage d'entités lexicales regroupées en thèmes dans un corpus. Il permet de procéder au calcul de la proportion des entités d'un thème par rapport à un autre thème et la proportion des entités des thèmes créés par rapport à l'ensemble des entités d'un texte du corpus (dans les faits : les tokens du corpus). Ce logiciel interactif d'étude de corpus permet à de mettre à profit la cohésion textuelle pour préstructurer des entités lexicales en les rassemblant en thèmes et pour évaluer l'importance numérique des entités de ces thèmes dans les textes d'un corpus d'observation. Il permet d'assister l'évaluation

[104] Comme on peut l'apprécier dans la figure 49 et comme on pourra le voir dans le chapitre 5, nous parlons de coloration aussi bien lorsqu'il s'agit d'une mise en valeur par une coloration en arrière plan avec une couleur donnée sur le principe du surlignage.

qualitative des entités sélectionnées. D'une part, le coloriage permet d'apprécier grossièrement la proportion des thèmes à l'intérieur d'un texte du corpus d'autre part, des calculs de fréquences sont proposés par le logiciel permettant d'apprécier :

- la proportion des entités d'un thème par rapport aux autres entités d'un texte ;
- le pourcentage de présence de chaque thème par rapport aux autres ;
- le taux de recouvrement d'un thème, la proportion des entités d'un thème présentes dans un texte. Le taux de recouvrement d'un thème permet de relativiser les deux autres mesures. Dans l'exemple de la figure 50 issue de [Beust, 2002*], 99 entités ont été associées au thème de l'économie alors que seulement une seule d'entre elle a été repérée (3 fois) dans un texte.

Il est important de noter que l'absence de ressources thématiques construites *a priori* ne permet pas d'évaluer la proportion des entités d'un thème donné par rapport à celles que l'utilisateur associe à ce thème. En fait, la quantité d'entités acquises à ce stade de constitution des ressources n'est pas un enjeu majeur attendu que les expériences ont montré qu'au final, les dispositifs construits ont un nombre d'entités lexicales aux alentours de 100 et qu'une vingtaine d'entités de départ suffisait à amorcer le processus de construction des dispositifs qui va permettre de compléter la liste initiale [Perlerin *et al.*, 2003].

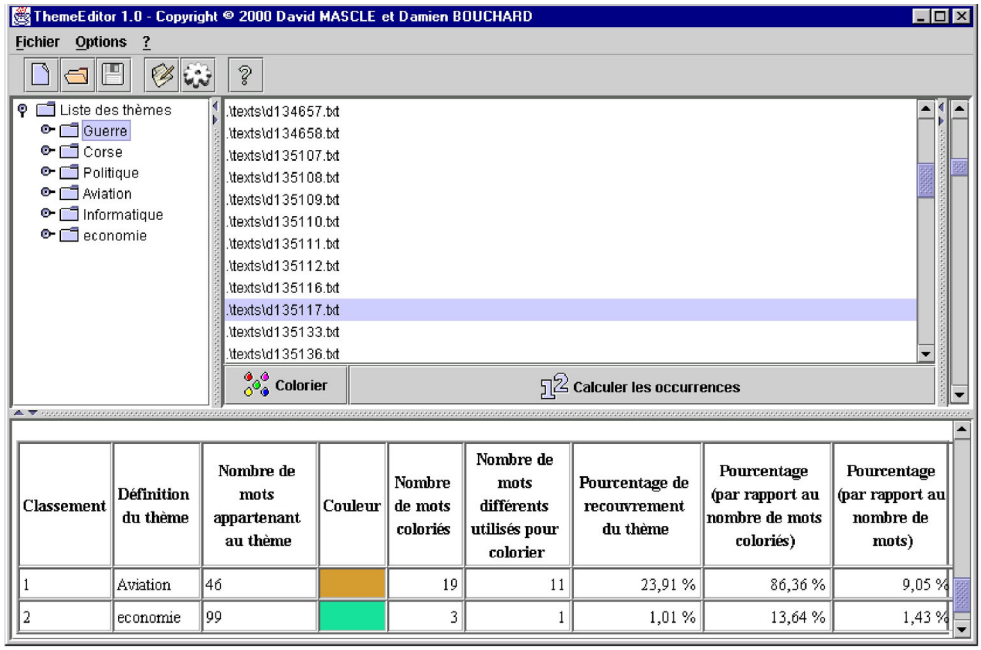

Figure 50 – Résultats d'un coloriage dans *THEMEEDITOR* d'après [Beust, 2002*].

Dans THEMEEDITOR, le regroupement thématique des entités lexicales est manuel. On peut assembler en thèmes des entités provenant de la composition de graphies obtenues depuis MEMLABOR et leur attribuer une couleur pour ensuite projeter cette coloration sur des textes du corpus. Une option

de l'application permet d'utiliser une base de données lexicales[105] pour ajouter au thème, en même temps que le mot sélectionné, certaines de ses flexions (c.f. partie 4.2.5). Si une entité est envisagée comme appartenant à plusieurs thèmes, l'heuristique de coloriage choisie dans THEMEEDITOR tend à prolonger le plus possible les isotopies, à favoriser la redondance, et attribue ainsi au mot la couleur la plus représentée dans le texte [Beust, 2002*]. En cas d'égalité, le logiciel prend en considération comme thème d'une entité le premier défini dans l'ordre chronologique.

La coloration permet de faciliter le repérage de ces entités dans les textes ou parties de textes et d'en apprécier la répartition et l'importance. Par exemple, la détection d'éventuelles zones majoritairement d'une même couleur permet de rapidement valider un ensemble d'entités d'un même thème : la cohésion textuelle favorise le regroupement de telles entités dans des zones proches ou contiguës des textes. Elle peut également comme dans l'exemple suivant favoriser l'ajout d'entités.

Gates et son entreprise Microsoft sont partout dans le projet Teledesic de 288 satellites de télécommunications (…); dans le capital de Comcast le troisième cablo-opérateur américain; dans Windows CE, un logiciel système de domotique ; dans WebTV, qui permet l'accès à l'Internet via la télévision, acheté pour 2,5 milliards de francs; dans la télévision, la presse, la création multimédia, etc. Impossible, quasiment, de trouver un domaine touchant aux nouvelles technologies de l'information et de la communication où Bill Gates ne soit présent, d'une manière ou d'une autre.

(1) Vie des réseaux, janvier 1998- *Bill Gates tentaculaire* .X.Celrieu. et P. Courcelles
http://www.regards.fr/archives/1998/199801/199801res02.html

Dans cet exemple, les entités coloriées appartiennent à la liste issue de calculs de MEMLABOR sur le corpus Microsoft (figure 45, p.137). Elles ont été rassemblées sous l'intitulé thématique *Informatique*. Ces entités, qui apparaissent dans un cotexte limité en taille, peuvent, si ce phénomène n'est pas unique dans le corpus d'observation, être validées comme appartenant à ce thème. De plus, l'étude de cette zone peut amener à ajouter d'autres thèmes comme « technologies de l'information » par exemple ou apporter des indications pour la description ultérieure en termes d'attributs associés à ces thèmes (on pourra par exemple retenir l'attribut [Nature du composant informatique : hardware *vs.* software] pour décrire et distinguer les entités *Windows*, *logiciel* d'autres du même thème).

En retour, l'absence de zones majoritairement colorées peut mettre en doute l'appartenance d'une entité à un thème dont elle serait par exemple l'unique représentant dans un texte. Dans un tel cas de figure, soit le nombre d'entités du thème n'est pas suffisant (trop peu d'entités prenant part à l'isotopie peuvent être repérées), soit l'entité en question a été attribuée à un thème de façon erronée (il peut s'agir d'un cas de polysémie par exemple). Dans les deux cas, il faut s'interroger sur la façon dont est abordé le thème en question et sa présence effective au sein des textes : est-il abordé comme il a été envisagé qu'il le soit lors de la définition de la tâche ? Pour savoir dans quel cas de figure l'on se

[105] MulText : http://www.lpl.univ-aix.fr/projects/multext/

trouve, on peut s'appuyer sur des calculs statistiques pour une évaluation grossière du phénomène de redondance. Il convient dans le premier cas de figure, de tenter de détecter de nouvelles entités du thème, dans le second soit de supprimer l'entité de la sélection de départ, soit d'en changer le thème associé. Si des entités proches sont supports d'une même isotopie en rapport avec le thème en question il convient de s'interroger sur la validité de ces entités vis-à-vis de la tâche. Éclairons nos propos par l'étude de l'exemple suivant :

La plate-forme Linux offre aux administrateurs système et aux programmeurs un choix de langages de scripts d'une richesse sans équivalent, qu'il s'agisse d'automatiser les tâches d'administration, de construire des interfaces graphiques, d'effectuer des traitements automatisés sur des fichiers texte ou de « parser » des documents XML. L'auteur explique dans cet ouvrage hors du commun quel langage choisir en fonction de ses besoins, avant de présenter en détail la syntaxe et les techniques de programmation des principaux d'entre eux : Tcl, Tk, Python et Ruby pour leurs notions essentielles ; shell Bash, Sed, Awk et Perl jusqu'à un niveau avancé.

(2) Langages de scripts sous Linux – C. Blaess – Résumé de l'éditeur
http://www.editions-eyrolles.com/php.informatique/Ouvrages/9782212110289.php3

Si le thème de l'informatique est nécessaire à aborder dans la tâche, on peut s'interroger sur l'absence de redondance dans l'extrait proposé en exemple où une seule des entités de ce thème a été repérée et donc coloriée. Si le thème de la tâche est celui du corpus d'observation Microsoft, « Linux » apparaît bien (et cela a été confirmé par l'étude statistique de MEMLABOR) comme une entité à sélectionner. L'absence de zone coloriée dans l'exemple (2) provient du fait que le corpus d'observation utilisé initialement (le corpus Microsoft) n'aborde pas l'informatique avec un point de vue technique mais avec un point de vue économique et judiciaire. Si la tâche nécessite de voir l'informatique également du point de vue technique, l'étude de l'exemple (2) peut amener à ajouter au thème, des entités telles qu'*administrateur, tâches d'administration* ou *interface graphique* par exemple. Dans le cas contraire, le choix du texte pour le corpus d'observation doit être mis en doute. THEMEEDITOR peut générer dans un seul et même fichier l'ensemble des textes coloriés du corpus. Un parcours rapide[106] de ce fichier peut permettre de relever de tels textes où les zones coloriées font défaut pour éventuellement les supprimer de la sélection. Ce parcours permet également d'apprécier l'empan de certaines isotopies à l'échelle du corpus et ainsi d'apprécier une éventuelle homogénéité intertextuelle (à partir des seuls thèmes produits) de textes qui le composent et apprécier par-là même la détermination du global sur le local au niveau de thèmes associés à certaines entités.

Lorsque plusieurs thèmes apparaissent dans des proportions identiques à l'intérieur d'un texte, l'opération est plus délicate : la coloration permet seulement, dans la phase de validation de l'association thème/entité, d'accélérer la lecture des documents en permettant le repérage rapide des

[106] Mais néanmoins peu pratique avec l'interface proposée – voir nos propositions dans ce domaine dans le chapitre 5.

parties à étudier. THEMEEDITOR souffre de ne pas constituer une interface de lecture du corpus car qu'il s'agit toujours de procéder à un parcours linéaire des textes sans représentation graphique intermédiaire entre l'original et le colorié. Le logiciel ne propose pas de vision globale sur l'ensemble des résultats obtenus du corpus entier. THEMEEDITOR ne propose pas non plus de solutions de structuration fine des entités lexicales ce qui limite la portée des évaluations proposées et limite l'étude des thèmes à celle du lexique.

4.2.5 Variantes morphosyntaxiques

Lors des analyses de corpus, il est possible de ne tenir compte que de la forme des entités proposées à partir de la composition manuelle des données issues de MEMLABOR. Cependant, nos premières expériences nous ont amené à constater que les analyses souffrent alors de certains manques. Trop peu d'entités sont repérées alors même qu'elles relèvent du paradigme flexionnel d'entités présentes dans des thèmes de THEMEEDITOR ou dans les tables des dispositifs. Il nous est donc apparu important de proposer la possibilité de traiter les variantes morphosyntaxiques des entités pour maximiser les résultats des analyses automatiques à partir des données fournies par l'utilisateur.

En plus du principe de racinisation (c.f. p. 72), deux choix principaux s'offrent à l'informaticien lorsqu'il s'agit de prendre en compte les différentes formes que peut prendre une entité lexicale lors d'analyses de corpus. Soit l'on utilise des bases de données lexicales associant généralement une forme canonique à un ensemble de flexions [Jacquemin et Zweigenbaum, 2000* : 73]. Soit l'on utilise des règles, des automates ou des transducteurs pour calculer automatiquement les formes ou revenir à une forme canonique à partir d'une forme fléchie ou dérivée – c'est qu'on appelle la lemmatisation [Corbin, 1987 , Dal et Namer, 2000 , Fabre et Jacquemin, 2000]. Lorsque les domaines traités ne sont pas cernés à l'avance, les deux approches peuvent poser problème. Ni les bases de données lexicales, ni les mécanismes de dérivation ou de reconstruction ne peuvent être exhaustifs. Il n'existe pas de base répertoriant tous les mots d'une langue pour la simple raison qu'il peut s'en créer tous les jours. Les systèmes à base de règles de dérivation ont l'avantage de réduire considérablement la quantité de données à stocker pour le fonctionnement du système et de tenir compte des mots nouveaux qui ont des flexions régulières. De plus, on en trouve pour de nombreuses langues. Cependant, les mécanismes proposés sont généralement coûteux en temps de calcul et posent parfois problème pour des langues à exceptions multiples comme le français.

Nous avons opté pour l'utilisation d'une base de données lexicales en laissant cependant le soin à l'utilisateur de ne pas sélectionner toutes les formes proposées automatiquement ou d'en ajouter à sa guise. Utilisant le langage XML aussi bien pour le stockage des données que pour l'annotation des corpus à analyser, l'utilisation d'une telle base permet de minimiser le nombre de calcul pour une enti-

té donnée. Cependant, elle pose problème lors d'un changement de langue puisque les données sont exogènes à la tâche.

La base que nous utilisons est MHATLex. Elle a été élaborée au laboratoire IRIT de l'Université Paul Sabatier de Toulouse[107]. Les ressources de MHATLex sont identiques à celles de BDLex pour le vocabulaire et la facette morphosyntaxique. BDLex contient environ 450 000 formes fléchies pour 50 000 formes canoniques. C'est du point de vue de la prononciation que les deux ressources diffèrent : de manière plus explicite et plus adaptée à la reconnaissance vocale, MHATLex permet de modéliser la prononciation avec ses variabilités libres et contextuelles. Utilisée au sein du GREYC pour des travaux en phonologie, cette base a été transformée au format XML pour les besoins de notre étude par Stéphane Ferrari et Pierre-Sylvain Luquet. Dans ce cadre, les informations phonologiques ne sont pas utilisées. Observons maintenant comment les informations issues de cette base sont intégrées aux fichiers des dispositifs (dans les fichiers de thèmes de THEMEEDITOR le mécanisme est sensiblement le même).

```
1. <table id="disp_La_Bourse_att3-4" attrs="attr3 attr4">
2.     <tablenom>Agents, Activités</tablenom>
3.     <ligne id="disp_La_Bourse_reg3-4ligne0" vals="attr3val0 attr4val0">
4.                 <lexApp lem="petit porteur" />
5.                 <lexApp lem="actionnaire" /> …
7.     </ligne> …
16. </table>
```

Figure 51 – Extrait de la représentation XML d'une table avant appariement.

Du point de vue du stockage, les données correspondant aux entités lexicales proposées pour les dispositifs et leurs formes possibles sont présentes dans des fichiers différents (un fichier par dispositif, un fichier pour les formes possibles, nommé `Dict_Lex`). Tant qu'une entité n'a pas été associée à ces formes possibles, elle est codée comme suit dans le fichier XML du dispositif (Figure 51). Le tag `lexApp` précise que l'entité lexicale est en attente d'appariement. Avant confrontation à MHATLex, un dispositif ne contient que des entités sous la forme de `lexApp`. La valeur de l'attribut `lem` ne correspond pas nécessairement à la forme canonique de l'entité. C'est bien l'entité telle qu'elle a été proposée par l'utilisateur qui est conservée ici.

[107] http://www.irit.fr/ACTIVITES/EQ_IHMPT/ress_ling.v1/accueil01.php - Institut de Recherche en Informatique de Toulouse.

```
<lex id="bdp34418">
 <lemme cat="G">petit</lemme>
       <flexion genre="M" nb="S">petit</flexion>
       <flexion genre="F" nb="S">petite</flexion>
       <flexion genre="F" nb="P">petites</flexion>
       <flexion genre="M" nb="P">petits</flexion>
</lex>
<lex id="bdp35959">
 <lemme cat="N">porteur</lemme>
       <flexion genre="M" nb="S">porteur</flexion>
       <flexion genre="M" nb="P">porteurs</flexion>
       <flexion genre="F" nb="S">porteuse</flexion>
       <flexion genre="F" nb="P">porteuses</flexion>
</lex>
<lex id="bda508">
 <lemme cat="N">actionnaire</lemme>
       <flexion genre="F" nb="S">actionnaire</flexion>
       <flexion genre="M" nb="S">actionnaire</flexion>
       <flexion genre="F" nb="P">actionnaires</flexion>
       <flexion genre="M" nb="P">actionnaires</flexion>
</lex>
```

Figure 52 – Extrait d'un `Dict_Lex`.

L'interrogation automatique de MHATLex à l'aide de requêtes XPath générées automatiquement permet la construction d'un fichier d'association entre une entité (sa forme canonique) et ses formes possibles (figure 52). Les informations d'ordre syntaxique sont conservées dans ce fichier bien qu'elles ne soient pas pour l'instant exploitées par nos programmes (attributs `cat` pour la catégorie, `genre` pour le genre). Conserver des informations n'augmentent pas de manière significative la quantité de ressources. Chaque entité est repérée par un identifiant unique (attribut `id` du tag `lex`).

Si les formes proposées depuis MHATLex sont incorrectes ou ne correspondent pas aux attentes de l'utilisateur, le fichier `Dict_Lex` de la session est modifié ; de nouvelles entrées correspondant aux souhaits de l'utilisateur sont créées. Pour la table proposée en exemple (figure 51), la forme *petites porteuses* n'est pas envisageable en tant que forme possible de *petit porteur*. En effet, dans le corpus utilisé pour la construction de ce dispositif (le corpus *Le Monde sur CD-ROM*), il n'y aucune féminisation du terme, ni même de formes épicènes qui amènerait à croire que l'usage est en train de changer. Le résultat d'une interaction avec l'utilisateur pour modifier cette sélection permet ainsi l'ajout automatique de deux nouvelles entrées au `Dict_Lex` sur le modèle de la figure 53[108].

[108] Les modalités informatiques de cette interaction seront détaillées dans la partie suivante.

```
<lex id="neo18">
 <lemme>petit</lemme>
        <flexion genre="M" nb="S">petit</flexion>
        <flexion genre="M" nb="P">petits</flexion>
</lex>
<lex id="neo19">
 <lemme>porteur</lemme>
        <flexion genre="M" nb="S">porteur</flexion>
        <flexion genre="M" nb="P">porteurs</flexion>
</lex>
```

Figure 53 – Extrait d'un `Dict_Lex` modifié.

Lorsqu'une entrée d'un `Dict_lex` est validée, les objets `LexApp` des fichiers des dispositifs sont remplacés automatiquement par des objets `lexie` (pour *entité lexicale*) sur le modèle de la figure 54. Si l'entité lexicale est complexe, elle correspond alors à un objet de type `clex` composé de plusieurs objets `lexie`.

```
<table id="disp_La_Bourse_att3-4" attrs="attr3 attr4">
       <tablenom>Agents, Activités</tablenom>
       <ligne id="disp_La_Bourse_reg3-4ligne0" vals=" attr3val0 attr4val0">
              <clex lem="petit porteur">
                     <lexie lem="petit"  ref="neo18"/>
                     <lexie lem="porteur" ref="neo19"/>
              </clex>
              <lexie lem="actionnaire" ref="bda508"/>
…
       </ligne>
…
</table>
```

Figure 54 - Extrait de la représentation XML d'une table après appariement.

La possibilité de modifier et d'ajouter des formes graphiques possibles pour une même entité permet l'association de formes non différenciées du point de vue de leur contenu sémantique dans le contexte d'une tâche. C'est essentiellement le cas pour les noms propres : on pourra par exemple considérer *MS* comme une forme possible de *Microsoft* dans un `Dict_lex`. La forme *K7* sera associée à *cassette* au même titre que *cassettes* si le corpus d'observation présente les trois formes et qu'il n'est pas utile de les distinguer, de même nous avons implanté la possibilité d'associer à une forme donnée d'une langue certaines formes d'autres langues. Pour le texte de ce tapuscrit, on peut ainsi envisager d'associer à *segment*, non seulement *segment* et *segments* mais également *token* et *tokens*. Ce mécanisme permet de réduire le nombre d'entrées des `Dict_Lex`, de les personnaliser le plus possible pour la tâche en cours et de réduire le temps de calcul lors des appariements. Nous avons déjà souligné le fait que l'utilisation d'une base lexicale telle que MHATLex était obstacle au multilinguisme et à

nos vœux d'aboutir à une *sémantique légère* : les ressources sont figées, non connues exhaustivement de l'utilisateur, etc. Pour pallier certains de ces inconvénients, Les modifications proposées à l'utilisateur amènent à réduire les données à celles qui sont véritablement utiles lors de la tâche et à personnaliser ces ressources.

Les fonctionnalités exposées dans cette partie, ainsi que celles permettant la création des structures de catégorisation sont proposées à l'interaction à travers le logiciel interactif LUCIABUILDER que nous présenterons dans la partie suivante.

4.3 LUCIABuilder – Logiciel interactif pour la construction de dispositifs.

En TAL, l'élaboration des ressources utilisées pour les logiciels fait rarement l'objet d'une réflexion poussée en termes d'interaction et ceci pour deux raisons principales : soit les ressources sont fournies *a priori* (à partir de processus automatiques ou de données partagées – type ontologies ou thésaurus), soit la constitution des ressources est réservée à des spécialistes ayant des compétences en informatique ou dans un langage formel donné. Par exemple, la plate-forme CONTEXTO de Minel [Minel, 2001] permet à un linguiste de constituer une base de données linguistiques (classes d'indicateurs, règles d'exploration contextuelle, etc.) à partir d'une interface composée de tables et de listes défilantes (figure 55). La plate-forme LinguaStream de Bilhaut [Bilhaut, 2003] propose une interface graphique pour l'agencement de traitements à effectuer sur des corpus. Chaque traitement est représenté par une boîte qu'il suffit de relier à une autre pour élaborer un enchaînement de procédures automatiques (figure 56). Cette interface constitue un effort important dans le domaine pour faciliter le travail du spécialiste, mais il ne s'agit pas ici d'aider à la construction des ressources. LUCIABuilder est un logiciel utilisable aussi bien par des spécialistes que par des novices c'est pour cela que nous avons apporté un soin tout particulier aux principes d'interaction et aux vues sur les données proposées aux utilisateurs.

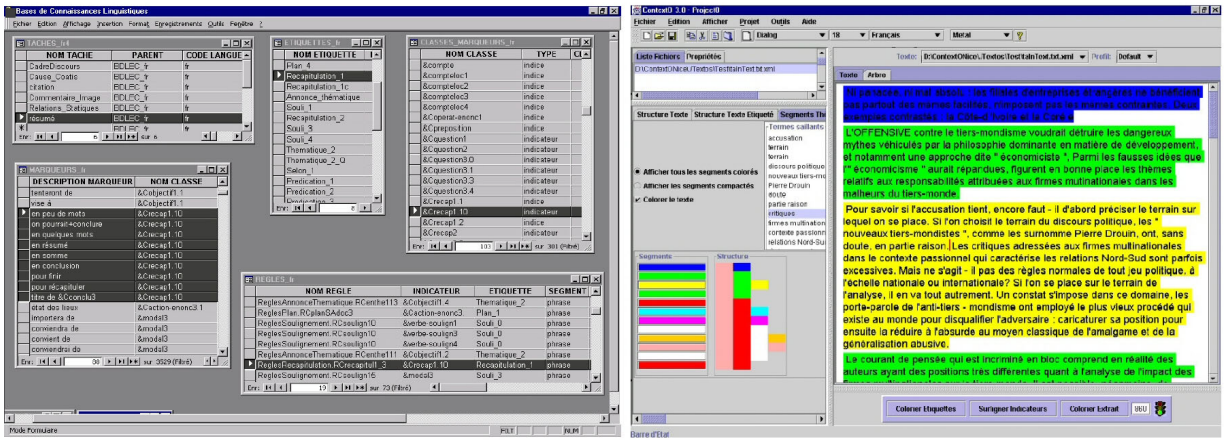

Figure 55 – CONTEXTO : production de ressources et visualisation d'après [Minel, 2001*] et [Ben Hazez *et al.*, 2001].

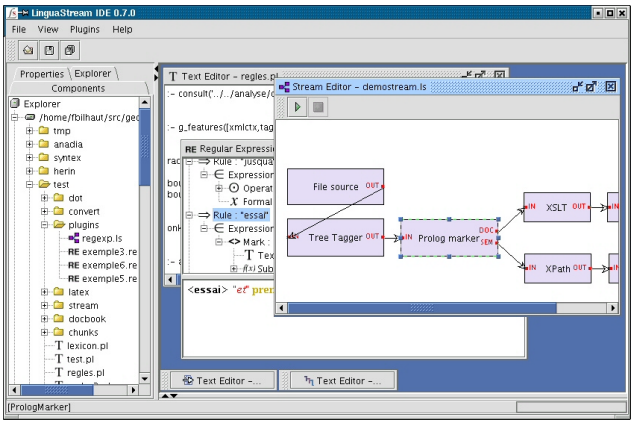

Figure 56 – LinguaStream d'après [Bilhaut, 2003*].

LUCIA place l'utilisateur au cœur même du système. Dans toutes les phases d'utilisation du modèle, aussi bien dans la phase de constitution des ressources que dans les phases d'exploitation des résultats obtenus du calcul automatique, l'utilisateur est l'acteur principal. Le logiciel a donc à charge de l'assister le plus efficacement possible et également de s'adapter à son niveau de compétence et à sa tâche. Les modalités d'interaction sont donc ici d'autant plus importantes qu'elles vont conditionner l'efficacité du système et la capacité de l'utilisateur à en comprendre le fonctionnement. La première application que l'on attend d'un logiciel mettant en œuvre un modèle de représentation lexicale est la gestion des ressources qu'il permet de manipuler : création, révision, *etc.* LUCIABUILDER intègre ces fonctionnalités classiques de gestion de données :

- création et modification de structures :
  - précision des attributs à utiliser ;
  - regroupement de ces attributs en tables et dispositifs ;
  - précision des liens d'héritage de lignes vers tables.

- ajout et modification des données (lexicales) :
    - positionnement des entités lexicales dans les tables ;
    - choix des formes graphiques associées aux entités.

Toutes ces actions ne sont pas nécessairement effectuées dans l'ordre proposé comme le montre le diagramme d'utilisation UML suivant (figure 57).

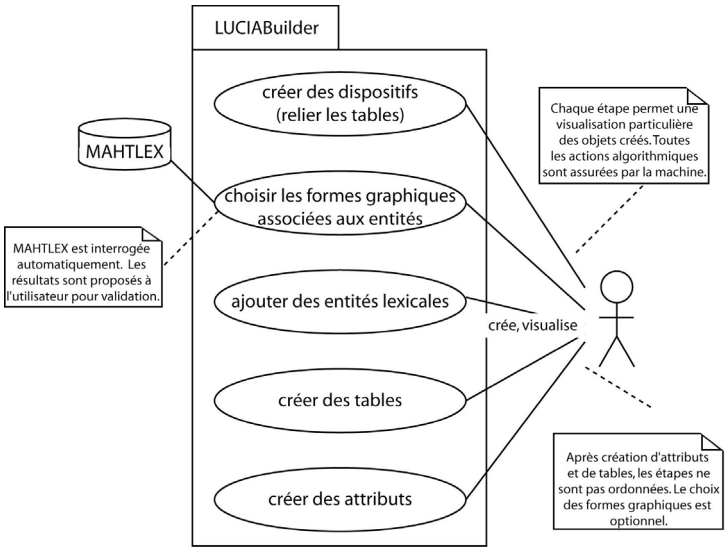

Figure 57 – Cas d'utilisation de LUCIABUILDER**.**

Bien que toutes ces données soient représentées en machine dans le format standard XML, nous avons ressenti le besoin de proposer à l'utilisateur une application entièrement dédiée plutôt qu'un éditeur standard adapté au langage de représentation informatique sous-jacent. L'objectif est de permettre à l'utilisateur de se familiariser avec les concepts qu'il manipule pour l'aider à mieux comprendre le comportement du modèle, sans devoir manipuler des modèles informatiques. De ce fait, LUCIABUILDER offre, en plus des fonctionnalités énoncées, de nombreuses vues différentes sur les structures et les données.

Quelles que soient la tâche et la méthode utilisées pour acquérir les entités lexicales de départ, LUCIABUILDER permet principalement la structuration de lexique(s) en dispositifs à travers une interface graphique composée de 5 panneaux (ou *panels*) principaux d'interaction. Le premier panneau (figure 58) permet de créer les attributs que l'utilisateur juge adéquats pour sa représentation. Les attributs créés peuvent être stockés dans différents ensembles, selon qu'ils sont jugés caractéristiques d'un domaine ou non (il s'agit de différencier les attributs partageables et les attributs partagés). Dans la copie d'écran présentée ci-dessus, l'attribut crée est nommé Direction. Son identifiant, calculé automatiquement est `attr13` et il est stocké dans un fichier nommé `Dict_Attr_Meteo.xml`.

Le second panneau (figure 59) permet le regroupement d'attributs pour composer les tables. L'utilisateur choisit les attributs et le logiciel calcule automatiquement la combinatoire de leurs valeurs. L'ensemble est aussitôt présenté sous la forme d'une table, pour habituer l'utilisateur à cette notion, et lui permettre la saisie des données. Les tables sont stockées dans des fichiers correspondant aux dispositifs LUCIA (et implicitement aux domaines décrits). L'interface force l'utilisateur à exploiter cette notion en présentant la liste des noms des tables d'un même dispositif dans la partie gauche de l'interface.

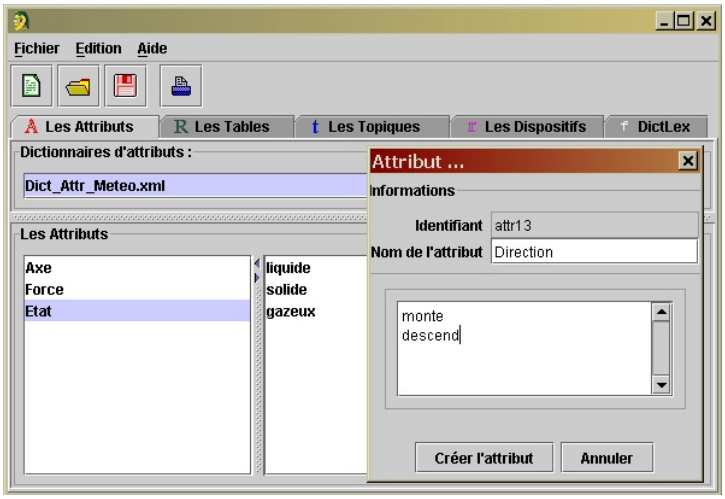

Figure 58 - LUCIABUILDER **: construction des attributs.**

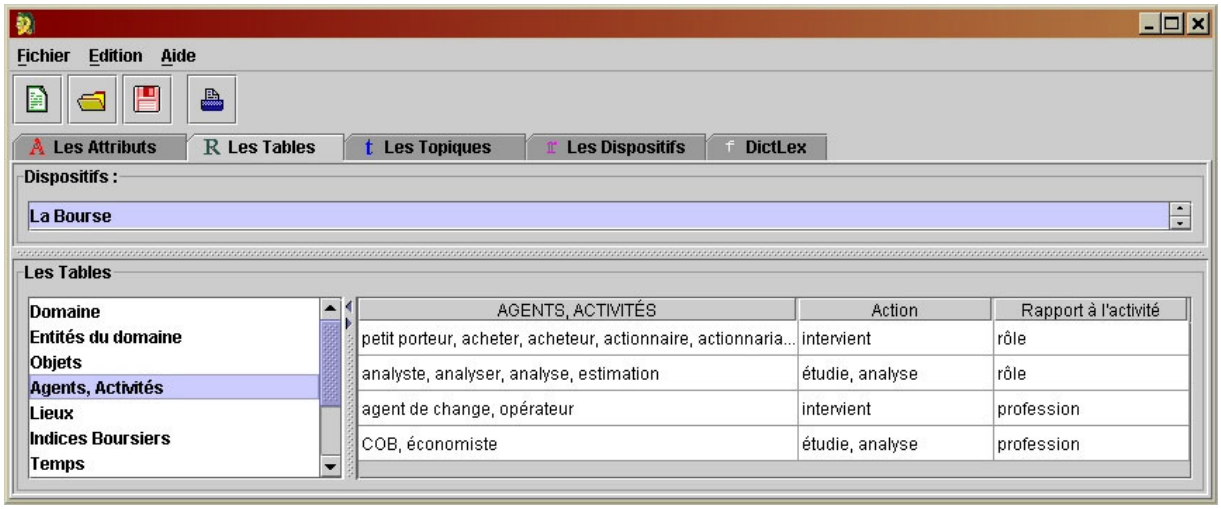

Figure 59 - LUCIABUILDER **: construction des tables.**

Le troisième panneau (figure 60) présente un point de vue différent sur les tables : la topique correspondante. Chaque ligne d'une table y est représentée par un rectangle déplaçable à l'aide de la souris. Il contient le premier mot de la ligne ou, si elle est vide, les valeurs des attributs qui lui correspondent. Les rectangles constituent les sommets d'un graphe dont chaque arête représente, et permet

d'apprécier, les différences mises en jeu entre les entités lexicales des deux lignes associées aux sommets reliés. Les arêtes portent les noms des attributs différenciant les sommets. Le nombre maximal de valeurs d'attributs qui diffèrent entre deux lignes d'une même table est le nombre d'attributs utilisés pour construire la table, et il n'est pas limité. LUCIABUILDER offre de ce fait la possibilité à l'utilisateur de préciser le nombre maximal ou le nombre exact de valeurs d'attributs différenciant les sommets à relier. Ceci permet de ne pas surcharger cette représentation en topiques. Les topiques permettent d'obtenir un point de vue différent sur les informations supportées par les tables, elles montrent en particulier plus nettement le nombre et la nature des différences proposées entre les instances d'une même catégorie. Dans la copie d'écran ci-dessous, pour la table « Agents, Activités » de « La Bourse », l'utilisateur a différencié *analyste* et *petit porteur* par une actualisation différente de l'attribut [Action].

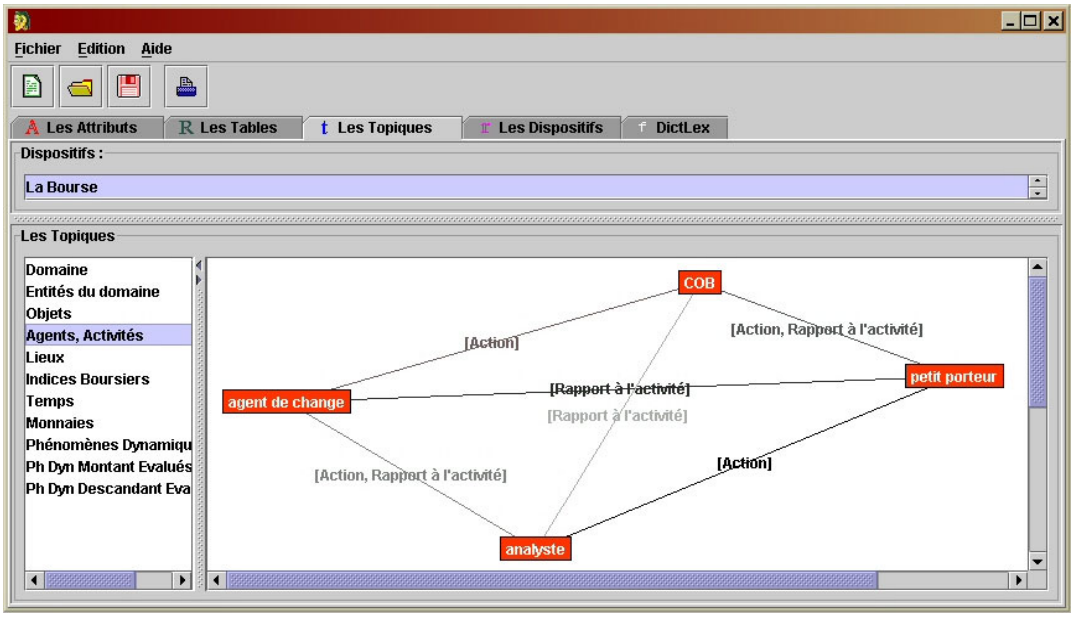

Figure 60 - LUCIABUILDER **: visualisation en topique.**

Le quatrième panneau (Figure 61) permet de créer les liens d'héritage dans un même dispositif, via une vue interactive simplifiée. Les tables y sont représentées par des rectangles déplaçables contenant uniquement leur nom. La première entité lexicale d'une ligne servant de point de départ à un lien d'héritage est affichée sur l'arête correspondante. Si la ligne est vide, [ligne vide] sert d'étiquette à l'arête. L'arrête est orientée : une flèche pointe vers la table héritant.

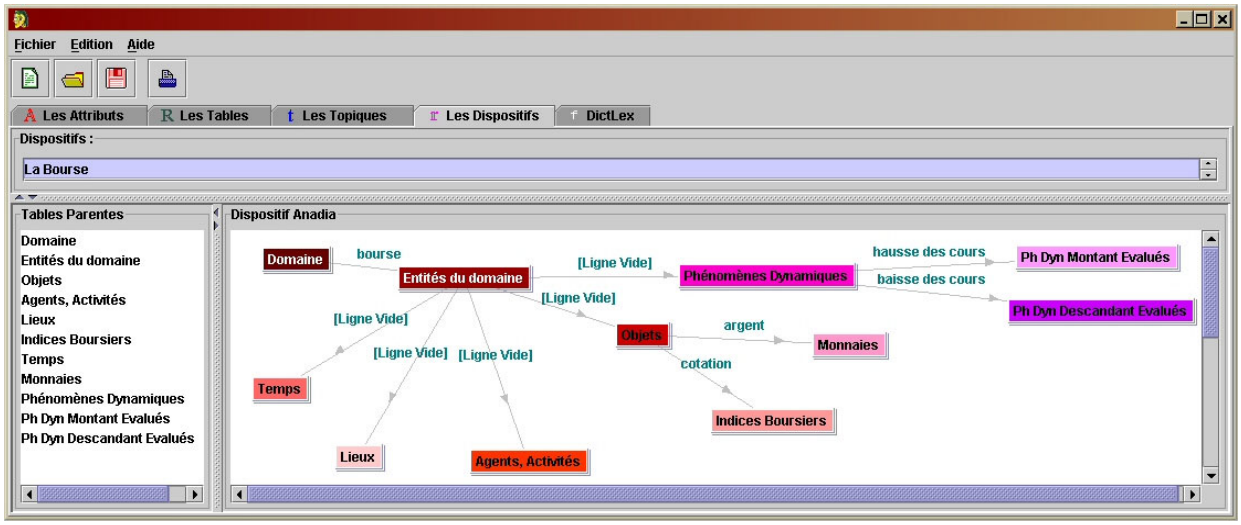

Figure 61 - LUCIABUILDER **: visualisation d'un dispositif**.

Le cinquième panneau (figure 62) permet l'appariement des entités proposées avec les données issues de MHATLex. Si dans le dispositif, des entités lexicales non appariées sont repérées (des `lexApp` dans les fichiers XML), le logiciel propose la création d'un `Dict_Lex` pour la session. Chaque entité est ensuite proposée à la validation (*actionnaire* pour l'exemple ci-dessous). L'utilisateur peut valider les données, les modifier ou créer une nouvelle liste de formes possibles dans une nouvelle fenêtre de dialogue (figure 63). Les aspects flexionnels n'étant pas au cœur du modèle, nous recherchons ici à accélérer cette phase, de manière à ne pas focaliser l'utilisateur sur un problème annexe à celui de la structuration proprement dite.

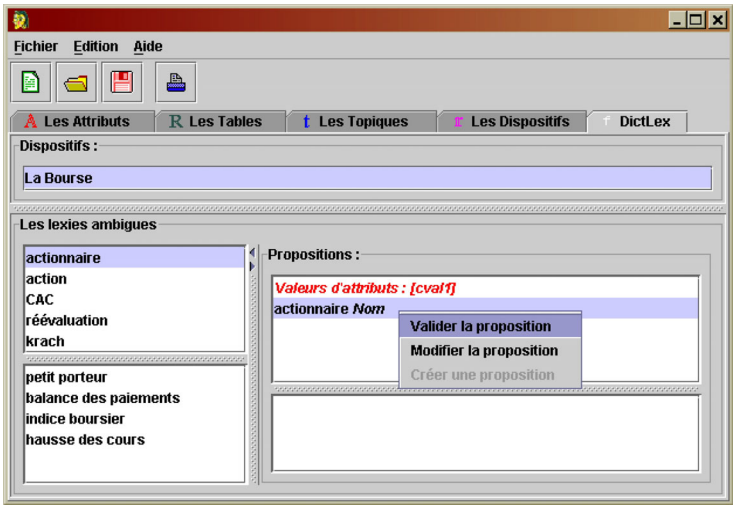

Figure 62 - LUCIABUILDER **: appariement des entités lexicales avec MHATLex.**

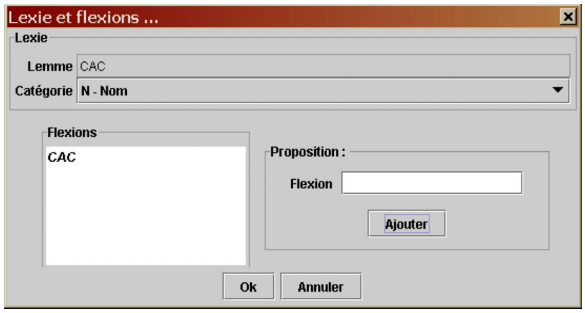

Figure 63 - LUCIABUILDER **: créations d'une nouvelle liste de formes possibles pour une entité.**

Dans les panneaux 3 et 4, l'utilisateur peut affecter une couleur à chaque table d'un dispositif. Ces couleurs sont utilisées dans les traitements ultérieurs, lors des différents affichages de résultats d'analyse et pour l'affichage dans le panel des dispositifs (figure 61). Elles doivent aider l'utilisateur à interpréter les résultats des analyses en les situant par rapport à sa propre structuration. Le chapitre 5 illustre différentes utilisations de cette particularité du modèle. Une fois définis, les fichiers XML correspondant aux dispositifs sont stockés en machine. Selon la tâche pour laquelle il utilise le système, l'utilisateur doit choisir des répertoires distincts. En effet, comme nous l'avons déjà signalé, l'exploitation de ces données sera différente en fonction de cette tâche.

D'autres vues exploitant les couleurs sont proposées dans l'interface. Il s'agit cette fois de familiariser l'utilisateur avec ses propres descriptions plus qu'avec le modèle. En particulier, nous avons exploité l'interactivité du langage SVG[109] pour produire une vue d'un dispositif (figure 64). SVG (*Scalable Vector Graphics*) est un langage de description de représentations graphiques en XML. Les graphiques produits en SVG sont interactifs et dynamiques. Trois principaux types d'objets graphiques sont permis : des formes vectorielles (traits, courbes, etc.), des images et du texte (hyperliens ou non). Les animations peuvent être définies soit à l'intérieur des fichiers SVG, soit dans un langage de script.

Cette vue permet à l'utilisateur d'apprécier la totalité des informations du dispositif, des entrées lexicales figurant sur chaque ligne de chaque table aux liens d'héritages. Les afficheurs SVG possèdent une fonction zoom, qui permet d'obtenir la vue d'ensemble présentée ici autant que la lecture des entrées lexicales des lignes de chaque table.

[109] http://www.w3.org/TR/SVG/

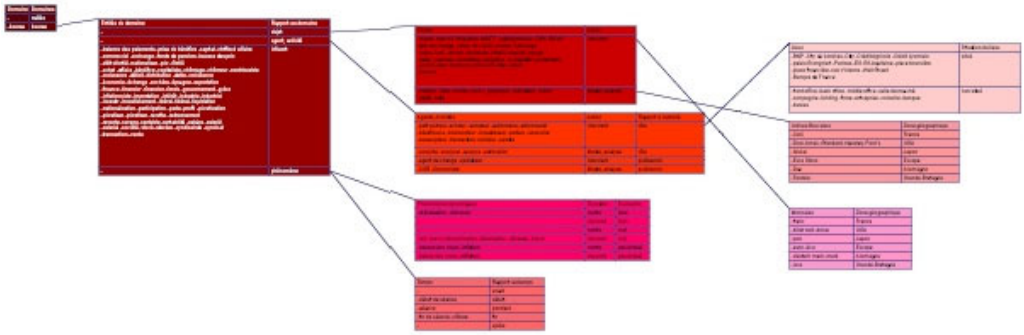

Figure 64 - Vue d'ensemble d'un dispositif complet au format SVG

Le logiciel LUCIABUILDER est utilisable sans connaissances informatiques préalables via son interface. Un utilisateur expérimenté pourra cependant avoir accès depuis l'interface principale aux fichiers XML générés automatiquement (Menu « Edition », sous-menu « Voir le source »). D'un point de vue technique, le logiciel a été implanté en Java (J.D.K. 1.4) et exploite une interface *Swing*. La manipulation et la création des fichiers XML s'effectuent à l'aide d'un *parser*. Le *DOM* des documents est manipulé pour l'interfaçage avec MHATLex. La représentation des dispositifs au format SVG est obtenue automatiquement à l'aide d'une feuille de transformation XSL. Celle-ci est manipulable dans les navigateurs interprétant le SVG avec ou sans recours à un greffon (Internet Explorer, Mozilla, Konqueror, Safari etc.) ou via les logiciels de visualisation dédiés comme ceux fournis par les sociétés Adobe ou Corel[110].

Si toutes les étapes algorithmiques sont bien assurées par le logiciel, la construction d'un dispositif n'en reste pas moins un exercice peu facile. Nous avons vu à travers l'expérience décrite dans le précédent chapitre que les concepts du modèle s'acquéraient aisément mais notre propre expérience nous a amené à nous interroger sur des possibilités d'assistance autres que celles déjà proposées dans LUICIABuilder. L'observation de dispositifs déjà construits nous a permis de relativiser la représentation en tables qui, par rapport au modèle, n'est qu'une représentation possible. Nous avons ainsi mis au jour certaines propriétés des attributs dans un dispositif construit. Ces propriétés sont utilisables par des processus dédiés pour l'assistance à la construction et à la révision des dispositifs. Ces aspects sont développés dans la partie suivante.

[110] http://www.adobe.com/support/downloads/main.html et
http://www.smartgraphics.com/Viewer_prod_info.shtml

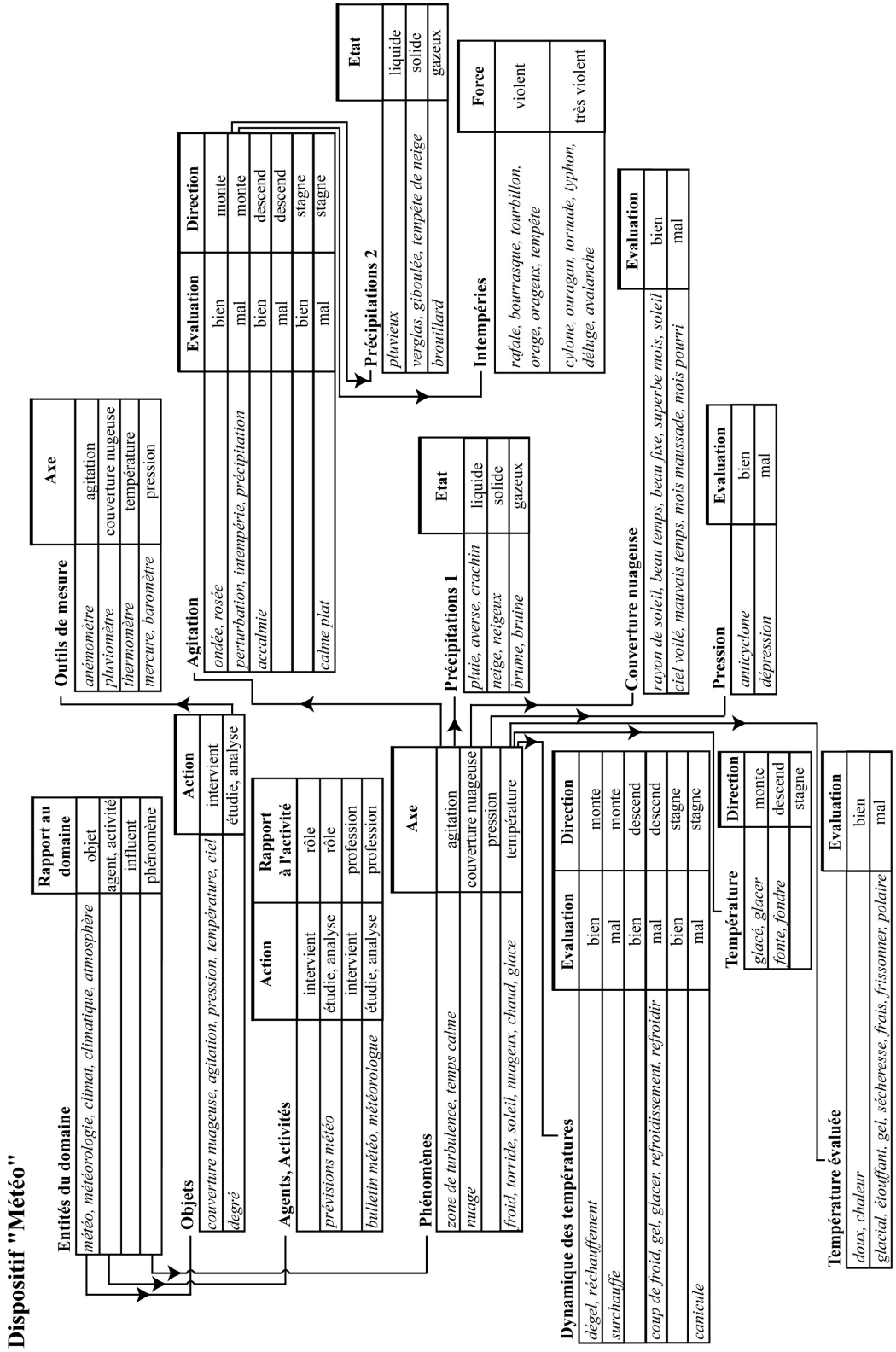

Figure 65 – Dispositif « Météo »

4.4 Propriétés des dispositifs

Dans cette partie, nous présentons un dispositif construit par nos soins (4.4.1). Dans un dispositif, les attributs acquièrent certaines propriétés qui peuvent être exploitées pour l'élaboration de processus automatiques d'assistance à la construction et à la révision des dispositifs (4.4.2). Ces processus sont symétriques dans le sens où ils s'appuient sur la transformation des informations contenues dans le structures LUCIA en fonction de deux représentations possibles : celle en tables et dispositifs et une en listes d'attributs/valeurs (4.4.2.1). Dans ces listes, on peut repérer des relations de subordination et non-subordination entre attributs (4.4.2.2). Ce sont ces relations qui peuvent servir de fondements à l'assistance à la révision ou l'évaluation de certains aspects de la cohérence d'un dispositif (4.4.2.3).

4.4.1 Exemple de dispositif

Le dispositif que nous présentons dans cette partie a été crée en collaboration avec Stéphane Ferrari et Pierre Beust pour une étude sur une métaphore. Ce dispositif (figure 65, p.159) est issu de l'observation du corpus *Le Monde sur CD-ROM* que nous avons déjà présenté. La tâche ayant présidé à la constitution de ce dispositif avait pour but l'étude de l'utilisation des entités lexicales *a priori* considérées comme ayant trait à la météorologie, au temps qu'il fait ou à l'état du ciel dans un cotexte qui portait majoritairement sur la bourse et l'économie. Il a permis de proposer des aides à l'interprétation d'emplois métaphoriques et d'assister l'analyse d'indices de caractérisation d'emplois métaphoriques de termes de la météo dans des textes traitant de la bourse et l'économie.

Nous avons déjà vu que le choix des attributs, des regroupements d'attributs, des liens de catégorisation et des entités instances des types des catégories sont l'expression d'un point de vue sur le domaine décrit et que ces choix ne s'appuient pas nécessairement sur les propriétés du monde, ni sur les propriétés lexicologiques des entités lexicales mais sur la façon dont elles sont interprétées au sein du corpus d'observation par l'auteur du dispositif dans le cadre de sa tâche. Ainsi, Mézaille [Mézaille, 2003] propose la dissimilation des items du taxème //phénomènes atmosphériques// en /occidental/ ('orage', 'bourrasque', 'tempête') vs /exotique/ ('ouragan', 'typhon', 'cyclone') dans l'étude d'un texte de Proust. Cette dissimilation n'apparaît pas en tant que distinction d'ordre sémantique au sein du dispositif proposé. Observons précisément les distinctions et les points communs entre les entités lexicales proposées dans ce dispositif. La différenciation /occidental/ vs. /exotique/ proposée par Mézaille (que l'on pourrait mettre au jour dans une table LUCIA à l'aide d'un attribut [Géographie[111] : occidental *vs.* exotique]) n'a pas été retenue comme pertinente à la vue des emplois des items précités dans le corpus d'observation. Dans le corpus, un seul emploi de *cyclone* est directement cooccurrent

[111] Attribut fortement occidentalo-centré.

d'autres entités pouvant amener à l'interpréter en tant qu'/exotique/ et non /occidental/. Un extrait de l'article correspondant est reproduit en (3)

Les plus pessimistes redoutent maintenant un violent séisme à Tokyo, place jusqu'ici relativement épargnée par le cyclone*. Là-bas, par devoir national, les Japonais empruntent de l'argent pour acheter des actions.*

(3) LE MONDE SUR CD-ROM - Le Monde du 07/12/1987 Page: 14

Pour les autres occurrences (*cyclone* apparaît neuf fois dans le corpus et *ouragan* trois fois), l'évocation d'un quelconque exotisme n'est pas probante – voir les exemples (4) et (5).

*(…) six jours après le lundi noir, le président du CNPF avait fait preuve d'optimisme et de sérénité en déclarant: " L'*ouragan *qui s'est abattu sur les marchés financiers internationaux, et qui est venu d'ailleurs, c'est-à-dire des Etats-Unis, a touché la France mais n'a pas ébranlé les entreprises françaises (…) "*

(4) LE MONDE SUR CD-ROM - Le Monde du 09/01/1988 Page: 24

Des fortunes ont été ainsi englouties. Des particuliers sont ruinés, des organismes au bord de la faillite. Le drame est que le cyclone a rendu inutilisables tous les instruments de mesure.

(5) LE MONDE SUR CD-ROM - Le Monde du 30/10/1987 Page: 34
(L'article aborde une actualité internationale. Il débute par *" Ce n'est plus la rue Vivienne mais le chemin des Dames ", lançait le mercredi 28 octobre sous les lambris à qui voulait l'entendre un ancien qui avait fait 14* .)

En revanche, la lecture des articles où les entités lexicales en question apparaissent, permet d'envisager une différence de signification locale de *cyclone* et *tempête* par exemple du point de vue de la violence des phénomènes décrits ou des conséquences de ces phénomènes – voir les exemples (6), (7), (8), (9) et (10). Cette distinction est exprimée par l'utilisation de l'attribut [Force : violent *vs.* très violent] dans le dispositif « Météo ». Comme on peut le voir dans l'état proposé de ce dispositif, les entités lexicales *bourrasque*, *rafale*, *tourbillon*, *orage*, *orageux* et *tempête* ont été décrites comme pouvant mettre en jeu les éléments de signification suivants : [Force : violent], [Direction : monte], [Evaluation : mal], [Axe : agitation], [Rapport au domaine : phénomène]. Le seul élément de signification permettant de les différencier des autres entités (*cyclone*, *ouragan*, *tornade*, *typhon*, *déluge* et *avalanche*) est l'attribut [Force], actualisé pour ces dernières par [Force : très violent].

Comme un ouragan*, la* tempête *qui s'était abattue, la semaine dernière, sur l'ensemble des marchés mondiaux, s'est déjà éloignée... des cotes boursières.*

(6) LE MONDE SUR CD-ROM - Le Monde du 26/10/1989 Page: 1

L'exploit ne s'arrête pas là puisque, au passage, un autre record est tombé avec près de 3800 milliards de francs de transactions (+ 25%), dont il est vrai les obligations (3400 milliards) ont pris la meilleure part, conséquence inévitable du cyclone *de 1987. (…) Qui l'eût cru? 1987 avait été l'année terrible.*

(7) LE MONDE SUR CD-ROM - Le Monde du 03/01/1989 Page: 17

Le 15 octobre, l'indice général était revenu à la case " départ ", s'inscrivant à un niveau légèrement inférieur à celui du début de l'année. Depuis, la place a pâti de la tempête *générale affectant les marchés financiers.*

(8) LE MONDE SUR CD-ROM Le Monde du 29/10/1987 Page: 34

La tempête *boursière n'a pour l'instant pas ralenti le nombre d'introductions sur le second marché avec Segin et Lhomme notamment. Elle n'a pas non plus freiné les ardeurs de la société Cegid dans son souhait de prendre le contrôle de CCMC,(…)*

(9) LE MONDE SUR CD-ROM Le Monde du 19/10/1987 Page: 14

Pour l'instant, si l'on en croit les discours officiels, il y aurait peu de cadavres victimes de la tempête *récente. A la moindre rumeur sur leurs pertes, agents de change et responsables de banques de trésorerie démentent.*

(10) LE MONDE SUR CD-ROM Le Monde du 03/12/1987 Page162: 35

L'attribut [Evaluation : bien *vs.* mal], utilisé à quatre reprises, est ici directement lié à la tâche pour laquelle on a élaboré le dispositif. Par exemple, son utilisation dans la table « Pression » permet d'exprimer la charge positive exprimée par l'utilisation métaphorique ou non du terme *anticyclone* dans le corpus (11).

À l'instar de la France entière baignée par le redoux, exceptionnel pour la saison, que l'anticyclone des Açores a poussé vers elle avec la complicité du foehn, la Bourse de Paris a eu sa petite poussée de chaleur.

(11) LE MONDE SUR CD-ROM Le Monde du 03/04/1989 Page: 14

Dans le dispositif proposé en figure 65 (p.159), toutes les tables sont reliées entre elles par un ou plusieurs liens. Si l'on dessine le graphe représentant les liens entre les catégories uniquement repérées par leur nom, voici les résultats obtenus (figure 66).

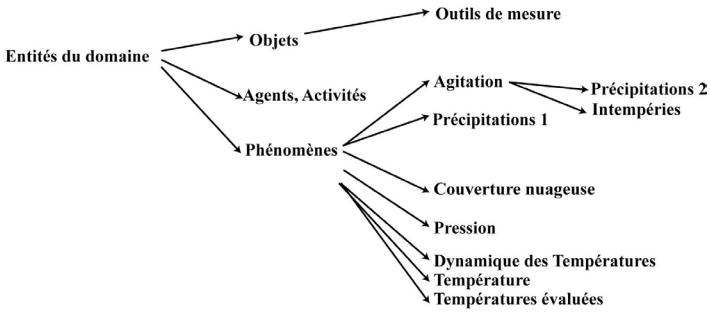

Figure 66 – Schéma du dispositif « Météo »

Ce schéma montre une hiérarchie entre les catégories du dispositif ; on remarquera que les tables d'un même niveau de cette hiérarchie ne présentent pas nécessairement de propriétés communes. Le choix des regroupements d'attributs pour constituer des tables et des liens entre tables sont à la source de la mise en place de cette hiérarchie.

Une entité lexicale peut correspondre dans un dispositif à plusieurs catégories. Deux homographes peuvent par exemple apparaître dans un même dispositif. C'est le cas également pour des em-

plois clairement distingués d'une même entité lexicale (voir l'exemple du *gel* présenté p.118). Le terme *dépression* a un statut particulier au sein du dispositif. S'il apparaît dans huit articles du corpus, les emplois repérés évoquent une *dépression* économique et non atmosphérique. Cependant, dans l'optique de notre étude sur la métaphore, il apparaît que le terme se retrouve dans l'ensemble des huit articles : il s'agit à chaque fois d'un phénomène négatif ; pour le monde de l'économie et de la bourse, la dépression économique fait figure d'épouvantail. Il convient dès lors de distinguer plusieurs types d'attributs. Au niveau des entités lexicales, types de catégories, nous ferons la différence entre les *attributs de catégorie* et les *attributs hérités*. Cette distinction peut faire écho à celle que l'on établit dans la SI entre sèmes génériques et sèmes spécifiques. Les *attributs de catégorie* sont ceux de la catégorie à l'intérieur de laquelle figure l'entité lexicale. Les *attributs hérités* sont associés à une entité lexicale du fait de liens de sous-catégorisation aboutissant à la catégorie de l'entité. Par exemple, et en fonction du dispositif présenté en figure 65 (p.159), à *dépression* correspondent les attributs/valeurs [Évaluation : mal], [Axe : pression], [Rapport au domaine : phénomène]. L'attribut [Évaluation : mal *vs.* mal] est ici un attribut de catégorie tandis que [Axe : pression *vs*. température *vs*. couverture nuageuse *vs*. agitation] et [Rapport au domaine : phénomène *vs*. agent, activité *vs*. objet *vs*. influent] sont des attributs hérités. La distinction entre des types d'attributs n'est pas primordiale lors de l'élaboration des dispositifs. Du point de vue informatique et lors des analyses de corpus (c.f. chapitre 5), cette distinction peut cependant s'avérer utile. Elle peut en particulier justifier l'insertion d'entités lexicales du domaine du dispositif, alors qu'elles apparaissent vraisemblablement dans le corpus d'observation comme porteuses de significations différentes. C'est le cas de *dépression* dans le dispositif « Météo » présenté ci-dessus. Nous verrons dans le prochain chapitre que nous faisons également une différence (plus importante pour le modèle) au niveau inter-dispositif et trans-dispositifs, entre les *attributs partageables* (ceux l'on peut utiliser dans plusieurs dispositifs, dans plusieurs domaine de tâche) et les *attributs propres* (qui sont *a priori* non partageables). Du point de vue de la SI, cette différence peut être comparée à celle qui oppose les sèmes potentiellement afférents et les sèmes inhérents – nous donnerons des informations plus complètes sur les attributs propres et partageables dans le chapitre suivant (partie 5.3.1 p.211). Contrairement à une tâche de lexicologie, il ne s'agit pas de demander à l'utilisateur de choisir la nature de ses attributs (de catégorie ou propre) car c'est la place des attributs dans le dispositif qui leur confère cette nature. De plus, au sein d'un dispositif les attributs acquièrent d'autres propriétés exploitables pour des processus d'assistance à l'élaboration et à la révision de telles structures. C'est ce que nous allons examiner dans la partie suivante.

4.4.2 Symétrie du processus : des attributs aux dispositifs

La construction des ressources LUCIA peut débuter soit, comme nous l'avons présenté, par la constitution de tables et leur mise en relation au sein d'un dispositif, soit par la mise en relation entre des listes d'attributs/valeurs et des groupes d'entités lexicales rassemblées à l'aide de THEMEEDITOR

ou simplement à la main. Cette dernière méthode permet de faire abstraction de la représentation en tables et dispositif du modèle et peut ainsi, être plus aisée pour des utilisateurs novices ; dans l'expérience décrite dans le chapitre précédent, nous avons pu voir que certains utilisateurs proposaient initialement ce type de regroupement (c.f. 3.4 p. 104). Elle donne un autre point de vue sur le modèle et participe ainsi à l'adaptation des principes d'interaction proposés aux usagers. Les propriétés entre attributs analysées par le logiciel permettent également une validation automatique des dispositifs. On pourra ainsi détecter les incohérences issues des contraintes du modèle et suggérer des modifications en transformant un dispositif en liste d'attributs/valeurs que l'on rétablira ensuite en dispositif.

Les manipulations automatiques des ressources que nous proposerons dans cette partie, tendent à minimiser le nombre de lignes vides dans les tables. Il s'agit de produire automatiquement des indications qui seront proposées à l'utilisateur qui pourra ou non les confirmer en fonction de ses besoins. La minimisation des lignes vides dans les tables ne constitue pas une contrainte du modèle. Les lignes vides peuvent s'avérer utiles pour certaines structurations (nous avons en présenter dans les exemples de tables de la figure 22 p. 93 et de la figure 29 p. 100). Les transformations assistées que nous proposons dans cette partie, permettent simplement de proposer aux utilisateurs des indices de détection de relations entre attributs et valeurs d'attributs susceptibles de ne pas correspondre aux exigences (fussent-elles minimales) de la catégorisation selon le modèle LUCIA. Ces indices, calculés automatiquement par le logiciel, ont également pour but de proposer des vues différentes sur les ressources construites pour apporter un certain recul par rapport aux données fournies et en faciliter ainsi l'évaluation comme cela est mis en place dans LUCIABUILDER.

4.4.2.1 Construction de listes d'attributs/valeurs

À partir d'attributs construits pour l'occasion ou récupérés d'autres dispositifs, il est possible de dresser une liste d'attributs/valeurs pour un groupe d'entités lexicales rassemblé en thème dans THEMEEDITOR ou à la main. Par exemple, si l'on a à disposition les deux attributs : Rôle [intervient *vs.* étudie, analyse] et Axe [vent *vs.* précipitations *vs.* température *vs.* pression] et que l'on souhaite décrire les entités lexicales : *couverture nuageuse, agitation, pression, température, ciel, degré, anémomètre, pluviomètre, thermomètre, mercure* et *baromètre* pour une tâche proche de celle ayant présidé à l'élaboration du dispositif « Météo » proposé en figure 65, on peut soumettre la liste d'associations suivante (figure 67).

couverture nuageuse, agitation, pression, température, ciel	**Rôle**[intervient]
degré	**Rôle**[étudie, analyse]
anémomètre	**Rôle**[étudie, analyse], **Axe**[vent]
pluviomètre	**Rôle**[étudie, analyse], **Axe**[précipitations]
thermomètre	**Rôle**[étudie, analyse], **Axe**[température]
mercure, baromètre	**Rôle**[étudie, analyse], **Axe**[pression]

Figure 67 – Liste d'associations entre entités lexicales et attributs/valeurs.

Ce type de listes peut être utilisé par le logiciel pour proposer un dispositif en fonction des propriétés de *non-subordination* et de *subordination* entre les attributs ainsi agencés (c.f. 4.4.2.2). Pour être valable, un groupe d'attributs/valeurs associé à une ou plusieurs entités lexicales dans un dispositif doit répondre à la contrainte suivante.

(1) Contrainte structurelle des groupes d'attributs/valeurs

Un groupe d'attributs/valeurs associé à une ou plusieurs entités lexicales est constitué d'attributs valués où chaque attribut apparaît au plus une fois dans le même groupe.

En ce qui concerne la réalisation informatique de cette définition, les groupes d'attributs/valeurs présentant plusieurs fois le même attribut avec une valeur identique sont modifiés pour ne le faire apparaître qu'une fois ; les entités lexicales qui sont associées plusieurs fois au même attribut avec des valeurs différentes sont soumises à l'utilisateur qui peut les éliminer ou modifier ces propositions.

Par le truchement des liens de sous-catégorisation, les listes de groupes d'attributs/valeurs peuvent être élaborées automatiquement depuis un dispositif. Par exemple, on peut, à partir de l'extrait de dispositif proposé en figure 68, construire la liste de la figure 67.

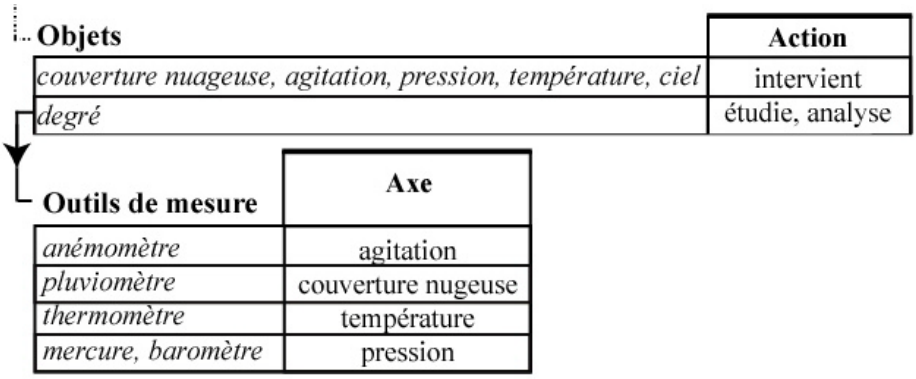

Figure 68 – Extrait d'un dispositif en rapport avec la météorologie

4.4.2.2 Propriétés des attributs et relations entre attributs dans des dispositifs construits

Dans [Beust, 1998* : 98], l'approche qui consiste à définir des concepts dans les tables de catégorisation ANADIA amène à définir la relation d'indépendance de deux attributs en fonction du nombre de lignes renseignées : *on sait que deux attributs sont indépendants si toutes les places de la table formée par leur combinatoire sont sélectionnées*[112]*, c'est-à-dire, si il n'y a pas de places vides.*

[112] Ce qui signifie que toutes les lignes de la table sont renseignées. Cette condition est une condition suffisante mais pas nécessaire.

Dans le cas contraire, le doute quant à la dépendance des deux attributs est possible. Dans la plupart des cas, cette dépendance consiste en une subordination des attributs. Un attribut A est subordonné à un attribut B si A est pertinent seulement quand B a certaines valeurs. La pertinence est définie comme suit : *un attribut est* pertinent *pour une catégorie si tous les objets de la catégorie ont la propriété dénotée par l'attribut mais qu'ils n'ont pas tous la même valeur pour cette propriété.* Ce qui, ramené au modèle LUCIA reviendrait à dire qu'un attribut est pertinent pour une catégorie donnée si toutes les catégories de ligne présentent bien un élément de signification, valeur de l'attribut, et que toutes les instances des catégories des lignes ne sont pas caractérisées par une même valeur de cet attribut.

Pour assister l'utilisateur dans la validation de la cohérence de ses représentations, nous proposons de redéfinir la notion de *non-subordination*[113] entre attributs à la lumière des contraintes de constructions d'un ensemble de catégories LUCIA ; dans ce cas, nous ferons abstraction de la représentation en tables de cet ensemble : c'est au niveau des groupes d'attributs/valeurs d'un dispositif déjà construit que sont définies ces contraintes.

(2) Définition : Non-subordination entre attributs

Etant donnés deux attributs A_i et A_j avec V_{Ai} l'ensemble des valeurs de l'attribut A_i et V_{Aj} l'ensemble des valeurs de l'attribut A_j. Les attributs A_i et A_j ne sont pas subordonnés au sein d'une liste de groupes d'attributs/valeurs si et seulement si :

- A_i et A_j apparaissent au moins dans deux catégories distinctes C_1 et C_2,
- il existe deux valeurs distinctes v_{Ai1} et v_{Ai2} appartenant à V_{Ai} et deux valeurs distinctes v_{Aj1} et v_{Aj2} appartenant à V_{Aj} tel que :
  - v_{Ai1} et v_{Aj1} appartiennent à C_1
  - v_{Ai2} et v_{Aj2} appartiennent à C_2

Notation :

La non-subordination entre deux attributs A_i et A_j est notée $A_i \sim A_j$.

Notes :

La non-subordination entre deux attributs A_i et A_j peut être valuée par le nombre absolu de co-occurrences différentes de valeurs distinctes de ces deux attributs dans les groupes d'attributs/valeurs considérés. On note ainsi $A_i \sim A_j\,(n)$ où n est l'entier correspondant à ce nombre de co-occurrences.

Propriétés :

(A) La non-subordination entre attributs est symétrique.

[113] La subordination est une relation de dépendance à part entière. La dépendance *stricto sensu* entre attributs n'est pas une relation intéressante dans le cas de la construction de dispositifs à partir de listes de groupes d'attributs/valeurs. Par exemple, la relation de dépendance entre les attributs A et $\neg$A ne saurait avoir de répercutions dans un tel cadre.

Démonstration :
Si $A_i \sim A_j$ alors il existe C_1 et C_2, v_{Ai1}, v_{Ai2}, v_{Aj1} et v_{Aj2} qui satisfont aux propriétés de la définition pour $A_j \sim A_i$ donc $A_j \sim A_i$.

(B) La non-subordination entre attributs n'est pas transitive.

Démonstration :
Soit $A_i \sim A_j$ et $A_j \sim A_k$, si $A_i \sim A_k$ alors il existe deux catégories C_1 et C_2 où A_i et A_k apparaissent conjointement or les relations $A_i \sim A_j$ et $A_j \sim A_k$ ne permettent pas de conclure au minimum à cette particularité.

À partir de ces définitions et propriétés, on peut élaborer une contrainte par défaut pour la transformation d'une liste de groupes d'attributs/valeurs :

Recommandation (3) : Regroupement en table de deux attributs non subordonnés

Si l'on désire minimiser les lignes vides des tables, deux attributs non subordonnés A_i et A_j avec $A_i \sim A_j$ *(n)*, peuvent être regroupés au sein d'une même table LUCIA si $((\text{card}(V_{Ai}) \times \text{card}(V_{Aj}))/2) = n$.

Cette contrainte se lit sur les topiques des tables des dispositifs (figure 69). Elle tend à minimiser les types non instanciés dans les tables (les places vides dans les colonnes de gauche des tables). En effet, $\text{card}(V_{Ai}) \times \text{card}(V_{Aj})$ représente le nombre de types renseignés d'une catégorie regroupant uniquement les attributs A_i et A_j et donc le nombre maximal de sommets de la topique correspondante si l'on ne considère que les types instanciés de la catégorie. On fixe à $((\text{card}(V_{Ai}) \times \text{card}(V_{Aj}))/2)$ le nombre minimum de types instanciés et donc le nombre minimum de sommets de la topique pour la table considérée, assurant ainsi à cette dernière une connexité partielle minimale.

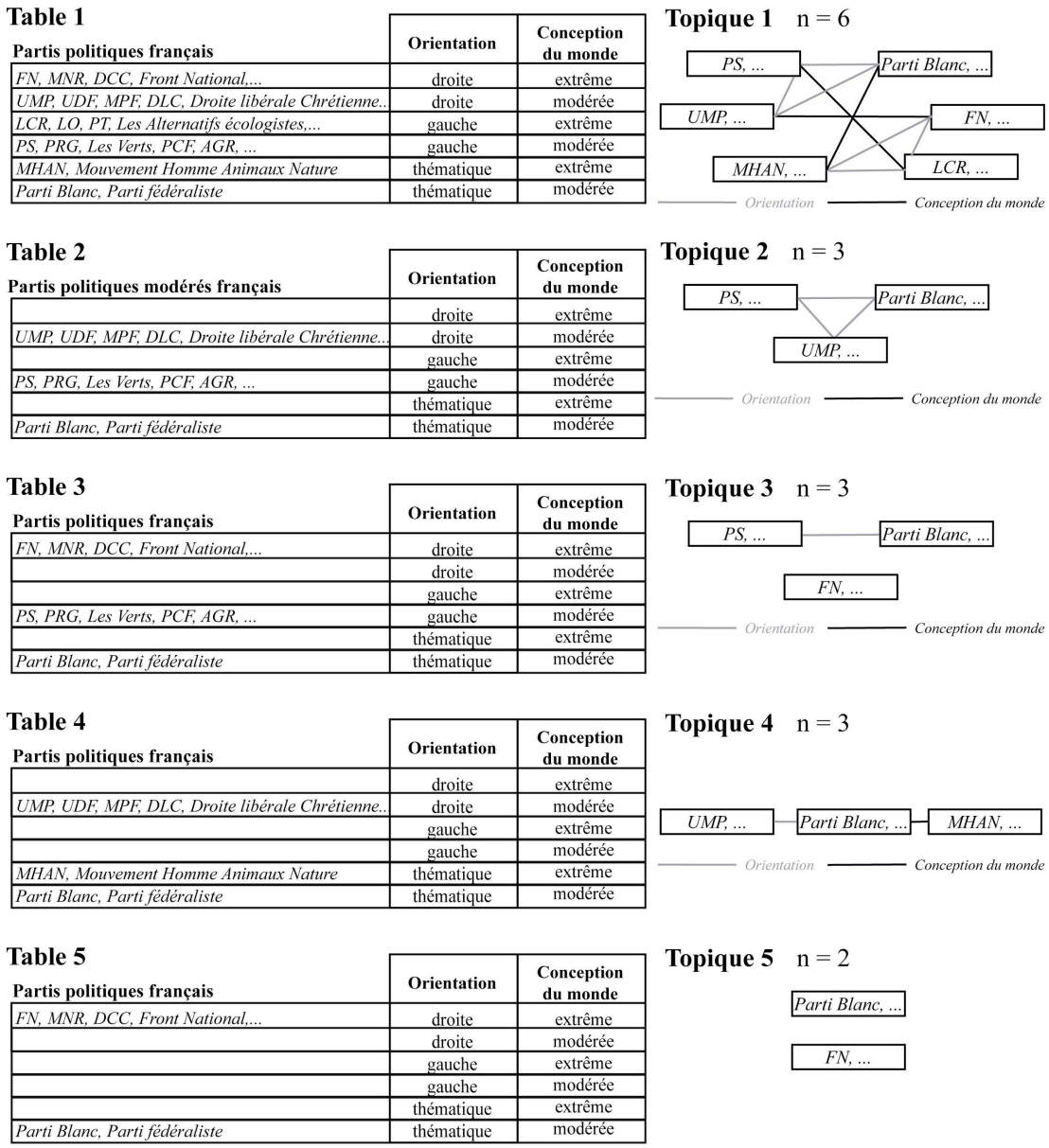

Figure 69 – Tables et topiques LUCIA **décrivant des partis politiques français.**

Les tables proposées en exemple dans la figure 69 ont été construites à partir de l'étude des pages Internet du site « France Politique[114] » dans le but de proposer un dispositif permettant de repérer les orientations politiques de partis politiques français. Les entités lexicales des tables ont été associées à des éléments de signification représentés par les attributs [Orientation : droite *vs.* gauche v*s.* thématique] et [Conception du monde : extrême *vs.* modérée]. Les associations proposées découlent de l'étude de ces pages et reflètent ainsi notre point de vue ; certaines configurations ne sont proposées

[114] http://francepolitique.free.fr - Ce portail francophone de la vie politique en France et en Europe propose des informations précises sur plus d'une vingtaine de partis politiques français (orientations, représentativité, etc.).

qu'à titre d'exemple pour exemplifier les propriétés des attributs mis en jeu lors de l'utilisation dans des dispositifs. Les tables correspondent à différents états de la catégorisation et donc à différentes listes de groupes d'attributs/valeurs pour les entités lexicales retenues.

Le produit de la cardinalité des valeurs des attributs utilisés dans les tables de l'exemple est égal à 6 - card($V_{Orientation}$) x card($V_{Conception\ du\ monde}$) = 3x2 = 6. Pour la table 1, les deux attributs sont en relation de dépendance telle que : *Conception du monde ~ Orientation (6)*. La formule ((card(V_{Ai}) x card(V_{Aj}))/2) = n correspond donc à 3 = 6 et elle est donc vraie. La topique correspondante à un trait près (Topique 1) est totalement connexe. Pour les tables 2, 3 et 4, les deux attributs sont en relation de dépendance tel que : *Conception du monde ~ Orientation (3)*. La formule correspond donc à 3 = 3 et est donc vraie. Les topiques correspondantes à un trait près (topiques 2, 3 et 4) sont au moins partiellement connexes. Pour la table 5, les deux attributs sont en relation de dépendance tel que : *Conception du monde ~ Orientation (2)*. La formule correspond donc à 3 = 2 et est donc fausse. La topique correspondante à un trait près (topique 5) n'est pas connexe.

La table 2 associe systématiquement aux entités lexicales de la catégorie l'attribut/valeur [Conception du monde : modérée]. Une seule valeur de cet attribut est en fait utilisée. Dans ce cas, il peut être intéressant de s'interroger sur l'utilité de la présence de cet attribut pour caractériser cette catégorie puisqu'il ne distingue pas les types effectivement instanciés. Seule la distinction ayant trait à l'orientation des partis politiques a été retenue pour la structuration. Dans la table 3, [Orientation : droite] est systématiquement associée à [Conception du monde : extrême][115], dans ce cas, il est préconisé de s'interroger sur l'utilité de la valeur 'droite' de l'attribut [Orientation] puisqu'elle n'est pas distinguée de [Conception du monde : extrême] pour les types instanciés de la catégorie. De même, dans la table 5, la valeur 'gauche' de l'attribut [Orientation] n'est jamais utilisée.

Dans le cas de la table 3, la distinction entre les deux attributs mis en jeu est discutable. On peut parler d'une certaine dépendance entre ces derniers. Comme nous le verrons dans les parties suivantes, cette dépendance ne correspond pas à une relation de subordination. La suppression de l'attribut [Conception du monde] ne modifie pas le fait que les trois groupes d'entités lexicales sont distingués entre eux. Nous l'avons déjà signalé : le choix du regroupement de ce type d'attributs peut dépendre directement d'une heuristique d'analyse particulière de la part de l'auteur de cette association. On ne peut donc taxer d'incohérence, ni voir aucune malformation, lorsque ce phénomène se produit dans un processus de construction de dispositif. Nous ne proposons donc pas de définition pour ce type de dépendance.

[115] Il aurait pu bien entendu en être autrement dans un exemple analogue associant systématiquement [Orientation : gauche] à [Conception du monde : extrême] – encore une fois, les tables de la figure 69 ne sont proposées ici que pour exemplifier les relations entre attributs.

Liée à la non-subordination, la notion de subordination, nous permet de proscrire le regroupement de certains attributs en se basant comme précédemment, sur les co-occurrences de ces derniers au sein de l'ensemble de groupes d'attributs/valeurs d'un dispositif.

Définition (3) : Subordination entre attributs

Étant donnés deux attributs A_i et A_j avec V_{Ai} l'ensemble des valeurs de l'attribut A_i et V_{Aj} l'ensemble des valeurs de l'attribut A_j et soit **C** l'ensemble des catégories LUCIA où apparaissent conjointement A_i et A_j, on considère que A_j est subordonné à A_i si et seulement si :

- **C** est non vide,
- quelle que soit C_k appartenant à **C**, il existe un v_{Ai} unique tel que $A_i v_{Ai}$ appartient à C_k.

La subordination entre deux attributs A_i et A_j par la valeur v_{Ai} de A_i est notée $A_i\, v_{Ai} > A_j$.

Propriétés :

Les notions de non-subordination et de subordination ont été définies séparément, il faut donc montrer qu'elles sont contraires.

(A) Deux attributs subordonnés ne sont pas non-surbordonnés.

Démonstration :
Si $A_i\, v_{Ai} > A_j$, C_k appartenant à **C,** la liste non vide de groupes attributs/valeurs présentant A_i et A_j, il existe un unique v_{Ai} appartenant à V_{Ai} l'ensemble des valeurs de l'attribut A_i tel que $A_i v_{Ai}$ appartient à C_k. v_{Ai} étant unique, il n'existe pas deux catégories distinctes C_1 et C_2 appartenant à **C** tel que v_{Ai1} appartiennent à C_1 et v_{Ai2} appartiennent à C_2 où v_{Ai1} est distinct de v_{Ai2}. A_i et A_j ne sont donc pas non subordonnés.

(B) La propriété de subordination entre attributs n'est pas symétrique.

Démonstration :
Si $A_i\, v_{Ai} > A_j$ et $A_j\, v_{Aj} > A_i$ alors A_i et A_j apparaissent au moins dans deux catégories distinctes C_1 et C_2 de la liste de groupes attributs/valeurs et il existe deux valeurs distinctes v_{Aj} et v_{Ai} appartenant respectivement à V_{Ai} et à V_{Ai} telles que v_{Ai} et v_{Aj} appartiennent à C_1 et v_{Ai} et v_{Aj} appartiennent à C_2. A_i et A_j ne sont donc pas non-subordonnés.

(C) La propriété de s entre attributs n'est pas transitive.

Démonstration :
Soit $A_i\, v_{Ai} > A_j$ et $A_j\, v_{Aj} > A_k$, si $A_i\, v_{Ai'} > A_k$ l'ensemble **C** des catégories LUCIA présentant A_i et A_k est non vide, or les relations $A_i\, v_{Ai} > A_j$ et $A_j\, v_{Aj} > A_k$ ne permettent pas de conclure au minimum à cette particularité.

La relation de subordination permet de définir une règle de détection de modification possible d'une partie d'un dispositif.

Recommandation (4) : Mise en dispositif d'attributs subordonnés

Lorsque deux attributs A_i et A_j sont subordonnés avec $A_i\, v_{Ai} > A_j$, ces deux attributs peuvent être placés dans deux tables distinctes telles que A_i est présent dans une table arborant un lien de sous-catégorisation avec la table regroupant l'attribut A_j et la ligne dont est issu ce lien actualise la valeur v_{Ai} de A_i.

Cette particularité permet, tout comme celle relative à la non-subordination entre attributs, de minimiser les types non instanciés des catégories. La notion de subordination telle que nous la définissons, peut être mise en relation avec celle proposée dans [Beust, 1998* : 101] où il affirme qu'un *(…) attribut A est* subordonné *à un attribut B si A est* pertinent *seulement quand B a certaines valeurs* et que *mettre dans un même registre* [regrouper au sein d'une même table] *deux attributs subordonnés amène un nombre important de places vides.*

Table 1
Partis politiques modérés français

	Orientation	**Conception du monde**
	droite	extrême
UMP, UDF, MPF, DLC, Droite libérale Chrétienne...	droite	modérée
	gauche	extrême
PS, PRG, Les Verts, PCF, AGR, ...	gauche	modérée
	thématique	extrême
Parti Blanc, Parti fédéraliste	thématique	modérée

Table 2a
Visions politiques

	Conception du monde
	extrême
	modérée

Table 2b
Partis politiques modérés français

	Orientation
UMP, UDF, MPF, DLC, Droite libérale Chrétienne...	droite
PS, PRG, Les Verts, PCF, AGR, ...	gauche
Parti Blanc, Parti fédéraliste	thématique

Figure 70 – Subordination entre attributs : tables décrivant des partis politiques français modérés.

Dans l'exemple proposé en figure 70, les deux attributs *Conception du monde* et *Orientation* sont en relation de subordination (Définition (3)). Rassemblés en une même table comme pour la table 1, le nombre de lignes vides est important (3 sur 6 possibles). Les tables 2a et 2b proposent les mêmes distinctions et descriptions des entités lexicales de la table 1 et minimisant les places vides au sein du dispositif construit. Cependant, elles ne reflètent pas exactement le même point de vue sur ces entités, et n'offrent pas les mêmes possibilités d'évolution sans modification profonde. C'est là les limites du tout-automatique : l'usager est seul à pouvoir décider de l'opportunité des modifications de ce type.

4.4.2.3 Assistance à la révision de dispositifs

Les relations définies dans les parties précédentes sont exploitées pour le repérage d'incohérences manifestes (incompatibles avec les contraintes du modèle) ou de possibilités de changement d'association d'attributs pour les catégories formées. Pour observer l'effectivité de la contrainte structurelle (1), il est possible de transformer un dispositif en liste de groupes d'attributs/valeurs puis d'inverser le processus. Si une incohérence est suspectée, prévenir l'utilisateur va lui permettre de modifier ses propositions et lui donner un certain recul quant aux données fournies. L'objectif est d'exploiter certaines propriétés algébriques des dispositifs pour assister leur réalisation. D'une certaine manière, c'est le même objectif que les topiques ; on propose à l'usager des représentations différentes de ses propositions pour mieux l'aider à les évaluer. Un dispositif non-conforme au modèle est proposé en figure 71. Ces données ont été construites à dessein pour exposer les principes

d'assistance à l'utilisateur pour la construction des dispositifs. Des identifiants ont été ajoutées au schéma pour repérer les attributs et leurs valeurs.

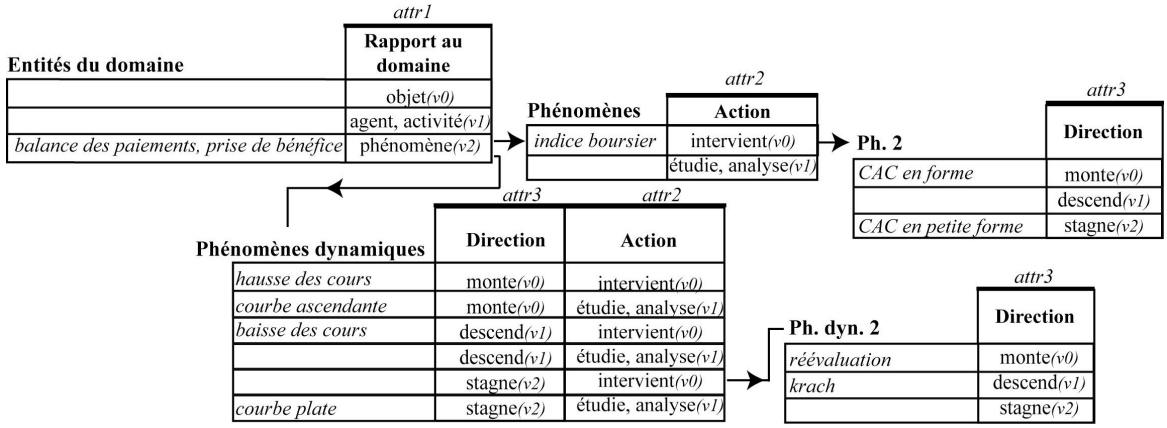

Figure 71 – Dispositif en rapport avec la bourse non-conforme au modèle.

À partir de ce dispositif, on construit automatiquement la liste des groupes d'attributs /valeurs correspondants (figure 72). Cette liste permet de repérer deux incohérences manifestes par rapport à la définition (1) (en gras sur la figure). Les groupes incohérents sont associés deux fois à l'attribut [Direction] (attr3) avec des valeurs différentes. Nous avons déjà abordé ce sujet dans le chapitre précédent : ce type d'attribut amène parfois à des incohérences à cause de la présence d'une valeur qui nie les deux autres. Stagner nie le fait de monter ou de descendre. L'utilisateur, une fois averti, a la possibilité de modifier ses propositions : il peut soit réviser cet attribut ou supprimer le lien de sous-catégorisation entre les tables incriminées.

balance des paiements	[attr1val2]
indice boursier	[attr1val2, attr2val0]
CAC en forme	[attr1val2, attr2val0, attr3val0]
CAC en petite forme	[attr1val2, attr2val0, attr3val2]
hausse des cours	[attr1val2, attr2val0, attr3val0]
courbe ascendante	[attr1val2, attr2val1, attr3val0]
baisse des cours	[attr1val2, attr2val0, attr3val1]
courbe plate	[attr1val2, attr2val1, attr3val2]
réévaluation	**[attr1val2, attr2val0, *attr3val2, attr3val0]***
krach	**[attr1val2, attr2val0, *attr3val2, attr3val1]***

Figure 72 – Liste des groupes d'attributs/valeurs correspondant à la figure 71.

Une fois les groupes non-conformes ôtés de la liste, on construit la matrice de co-occurrences des valeurs d'attributs (figure 73). La représentation proposée ici place une croix lorsque les valeurs sont co-occurrentes dans un groupe. La matrice étant symétrique (la co-occurrence est bien sûr une relation symétrique), seule une moitié a été reproduite.

	attr1val0	attr1val1	attr1val2	attr2val0	attr2val1	attr3val0	attr3val1	attr3val2
attr1val0								
attr1val1								
attr1val2								
attr2val0	/	/	×					
attr2val1	/	/	/					
attr3val0	/	/	×	×	×			
attr3val1	/	/	×	×	/			
attr3val2	/	/	×	×	×			

Figure 73 – Matrice de co-occurrences.

	attr1	attr3
attr2	6 *(n=1)*	6 *(n=5)*

card(Vattr$_i$) x card(Vattr$_j$)

Figure 74 – Produits de la cardinalité des ensembles de valeurs pour les attributs non subordonnés.

En fonction des recommandations (3) et (4), on peut proposer l'organisation suivante :

- attr2 et attr3 peuvent être regroupés dans même table ((card(V_{attr2}) x card(V_{attr3j}))/2) = *n* correspond à 6/2<5 ce qui est vérifié – recommandation (3). (attr1 et attr2 doivent de préférence apparaître dans des tables différentes).
- attr1 et attr3 doivent plutôt appartenir à deux tables distinctes avec un lien de sous-catégorisation partant de la catégorie formée par attr1 (le type correspondant à attr1val2 puisque attr1(val2)>attr3) – recommandation (4) .

À partir de la configuration de départ, la suppression des entités lexicales dont les groupes d'attributs/valeurs ne sont pas conformes et le suivi des recommandations proposées, amène au nouveau dispositif suivant (figure 75).

	attr1		*attr3*	*attr2*
Entités du domaine	**Rapport au domaine**	**...**	**Direction**	**Action**
	objet*(v0)*	*hausse des cours, CAC en forme*	monte*(v0)*	intervient*(v0)*
	agent, activité*(v1)*	*courbe ascendante*	monte*(v0)*	étudie, analyse*(v1)*
balance des paiements, prise de bénéfice	phénomène*(v2)*	*baisse des cours, CAC en petite forme*	descend*(v1)*	intervient*(v0)*
			descend*(v1)*	étudie, analyse*(v1)*
			stagne*(v2)*	intervient*(v0)*
		courbe plate	stagne*(v2)*	étudie, analyse*(v1)*

Figure 75 – Dispositif en rapport avec la bourse conforme au modèle.

L'exemple proposé dans cette partie est très simple. La mise en œuvre de ces principes dans un cas d'utilisation réel amène à détecter un certain nombre de cas où les configurations possibles relèvent de l'utilisateur en fonction de son projet. Cependant, cette approche des relations entre attributs et valeurs d'attributs permet d'amorcer un processus de validation et d'objectivation des données

fournies. C'est ce même souci de changement de point de vue sur les données qui a présidé à la réalisation des outils informatiques implémentant le modèle, LUCIABuilder et les solutions logicielles pour la lecture et l'évaluation des résultats d'analyse qui seront l'objet du prochain chapitre.

Pour clore ce chapitre, nous proposerons dans la prochaine partie, un protocole de construction de dispositifs. Une fois encore, il s'agit d'un moyen d'assistance à l'utilisateur, lui proposant des pistes pour amorcer la réalisation d'un tel support d'informations.

4.5 Protocole de construction d'un dispositif

Pour proposer un protocole de construction des dispositifs LUCIA, il est difficile de fournir une ligne de conduite stricte, car les facteurs qui influencent cette tâche sont multiples et découlent de la situation de la tâche documentaire à réaliser. Étant données les libertés permises par le modèle et par les implantations informatiques le mettant en œuvre, ce protocole n'est pas une *norme* mais un simple jeu de recommandations génériques. Le protocole proposé dans cette partie est celui que nous avons suivi lors de nos propres expérimentations, il reprend en fait points par points les parties de ce chapitre. Il est proposé ici à titre indicatif dans la mesure où nous sommes conscients des différences qui peuvent exister entre les vœux initiaux présidant à la réalisation d'outils informatiques et les usages effectifs qui en sont fait. Les instruments et les théories informatiques ont souvent une « vie » propre, relativement indépendante des théories dont ils se réclament [Bruillard, 2003 : 232]. C'est l'un des atouts du modèle LUCIA : nous pensons que sa souplesse et son ouverture peuvent lui assurer une certaine pérennité.

1°) **Définition de la tâche**

La première étape est celle de définition de la tâche par l'usager. D'un point de vue informatique, il est recommandé de créer un répertoire relatif à cette tâche sur le modèle suivant (figure 76).

Le répertoire « Corpus d'observation » est facultatif puisque, comme nous l'avons vu en 4.2.3 et 4.2.4, les logiciels THEMEEDITOR et MEMLABOR ne modifient pas les fichiers du corpus et peuvent accéder aux documents via des chemins (références absolues ou relatives). Les résultats obtenus de ces deux logiciels peuvent être stockés respectivement dans les répertoires « Résultats de MemLabor » et « Thème de ThemeEditor » qui contiendront les fichiers XML issus des interactions avec les logiciels en question (listes Zipf et ensembles d'entités regroupées en thèmes). Les ressources de LUCIABuilder, les fichiers XML de la session, peuvent être stockées dans deux sous-répertoires distincts attendu qu'un même ensemble d'attributs (présent dans un `Dict_Attr`) peut être utilisé pour plusieurs dispositifs et plusieurs sessions à la fois. Cette proposition est également facultative car les fichiers XML d'une session sont constitués de références vers les fichiers à prendre en compte. Ainsi, un même `Dict_Attr` pourra être utilisé pour plusieurs tâches différentes et être placé dans un autre endroit dans

l'arborescence des fichiers de l'utilisateur. Il faudra alors préciser le répertoire d'où devront être puisées les informations utiles aux analyses. Enfin, les résultats obtenus de ces analyses seront stockées dans un dernier répertoire.

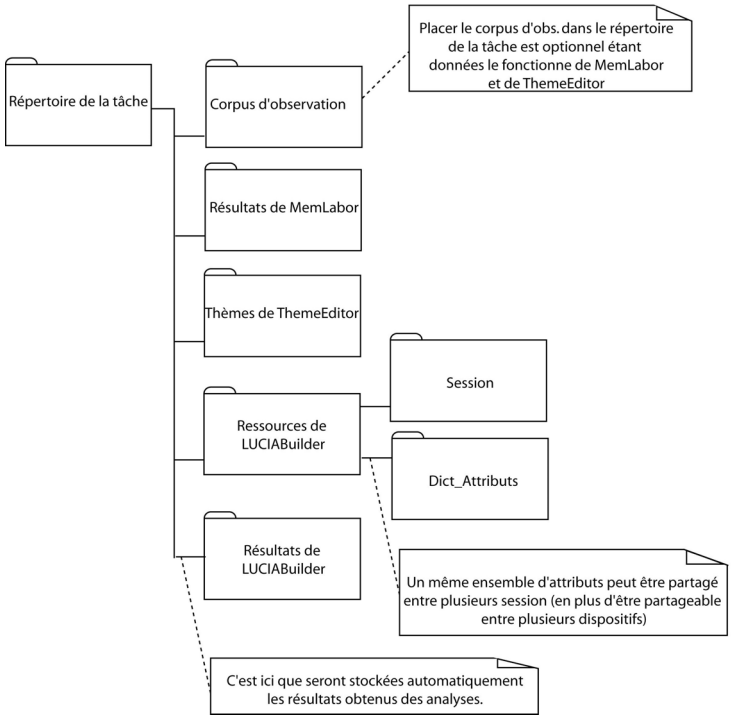

Figure 76 – Arborescence de répertoire recommandée

2°) **Constitution du corpus d'observation, acquisition, pré-structuration**

La constitution du corpus d'observation a été abordée en 4.2 : c'est une étape facultative mais fortement recommandée. MEMLABOR assiste l'usager pour l'acquisition d'entités lexicales depuis un corpus. Les entités lexicales extraites du corpus seront catégorisées en fonction des connaissances de l'utilisateur sur le domaine à traiter, en fonction de la tâche à réaliser et en fonction de l'observation des significations amenées par ces dernières dans le corpus d'observation. La structuration en thèmes à l'aide de THEMEEDITOR constitue une pré-structuration des ressources qui soulage le travail de l'étape suivante : les thèmes ainsi formés pourront correspondre dans les dispositifs à des catégories.

3°) **Structuration et description des entités lexicales**

Pour les utilisateurs novices, nous avons vu lors de l'expérience décrite dans le chapitre 3 que les catégories prises en considération dans un premier temps étaient majoritairement de nature ontologique. Pour débuter la structuration à l'aide de LUICIABuilder, on peut par exemple s'inspirer des « catégories ontologiques de bases » proposée par Jackendoff, sans pour autant souscrire à la théorie des constituants conceptuels et des « Structures Sémantiques » [Jackendoff, 1992]. Ses catégories sont : *state, event, path, place, manner, purpose, thing, property* et *time* (état, évènement, chemin, lo-

calisation, manière, but, choses, propriétés et temps). Selon Jackendoff, de telles catégories peuvent être décomposées en sous-catégories « conceptuelles ». « Choses et entités » peut être par exemple décomposée en « humain », « animal » et « objet ». Avec LUCIA, on peut sans distinction hiérarchique considérer à un même niveau les catégories et sous-catégories proposées. Les valeurs d'attributs de nature ontologique que nous avons utilisées dans les dispositifs pour notre étude sur la métaphore sont les suivantes : phénomène, agent, activité, objet et influent - cette dernière rassemblant toutes les entités lexicales intéressantes pour la tâche mais dont l'association avec le domaine traité n'était pas évidente pour les membres du projet. Ces catégories sont représentées par un seul attribut au sein des dispositifs. Dans celui qui a trait à la bourse et à l'économie (chapitre 5, p.189), construit pour cette même expérience, les deux catégories de lieux et de temps regroupant respectivement des entités lexicales telles que *BNP*, *city de Londres*, *back office*, *rue Vivienne*, *salle des marchés (…)* et *début de séance*, *fin de séance*, *clôture*, *(…)* sont des sous-catégories issues des valeurs d'attributs [objets] et [phénomènes]. Elles ne sont donc pas en adéquation avec celles proposées par Jackendoff pour qui la catégorie *temps* est de niveau supérieur à celle d'*objet* par exemple. Leurs relations concordent simplement avec les interprétations faites des entités en question dans le corpus d'observation pour la tâche en cours. Les dispositifs LUCIA ne définissent pas des concepts ; ils structurent sémantiquement du lexique. Le choix de catégories qui appartiennent à l'ontologie traditionnelle pour amorcer une structuration, n'implique aucun préexistence du concept par rapport au terme. Ainsi, la liste de Jackendoff n'est ni exhaustive, ni figée, ni obligatoire : elle peut servir de base pour une approche simple de la première étape de structuration. Nous nous abstiendrons de prendre position sur une quelconque « naturalité » de ces catégories. Nous constatons tout au plus que l'expérience décrite dans le chapitre précédent amène à croire qu'elles sont facilement accessibles à des non-spécialistes de la description lexicale. Cependant, les catégories de nature ontologique peuvent s'avérer totalement inadaptées si l'organisation lexicale est faite en vue d'une analyse linguistique ou littéraire. L'exemple de l'étude de *l'Assommoir* de Zola (chapitre 3, p.118) nous montre que les attributs (et donc les catégories de tables) utiles à de telles analyses se soustraient largement au consensus ontologique.

4°) **Structuration**

À partir des thèmes formés dans THEMEEDITOR, on peut adopter la démarche suivante :

- pour chaque thème créé, rassembler des entités lexicales qui partagent potentiellement des éléments de significations ;
- pour chaque groupe d'entités lexicales ainsi formé, créer les attributs correspondant aux éléments de significations envisagés.

Si certains de ces attributs ont déjà été utilisés pour une autre tâche ou pour un autre dispositif, il faut importer dans la session le `Dict_Attr` dans lequel on les trouve et utiliser le panel adéquat de

LUICIABuilder pour les tables à construire. L'utilisation d'attributs déjà existants ou partageables entre plusieurs dispositifs est préconisée pour des tâches où il est intéressant d'étudier des récurrences d'attributs trans-domaines. C'est par exemple le cas pour l'étude sur la métaphore (c.f. [Perlerin et Ferrari, 2003a] et la partie 5.3 du chapitre 5). Lorsqu'il s'agit de veille documentaire, les récurrences relatives à un même domaine (celui de la recherche) pourront suffire pour procéder au filtrage et au classement des résultats obtenus. Il n'est donc pas indispensable d'utiliser des attributs déjà utilisés pour d'autres dispositifs que celui de la tâche en cours.

Il faut ensuite rassembler les attributs pour créer les tables. Les places vides des tables ainsi formées peuvent être complétées par des entités lexicales non répertoriées après les manipulations à l'aide de MEMLABOR et THEMEEDITOR[116].

Une fois un certain de nombre de tables proposées, on peut créer les liens de sous-catégorisation entre tables. Si une table de sous-catégorisation paraît utile à la tâche alors qu'elle n'a pas été construite, on peut la créer en suivant les indications données précédemment. Les liens de sous-catégorisation entre tables permettent de concevoir des dispositifs où les entités lexicales d'un domaine sont organisées des plus générales aux plus déterminées pour la tâche en cours. Ceci permet également d'apprécier la cohérence d'un dispositif par rapport aux objectifs initialement fixés.

On réitère ces opérations jusqu'à ce que le dispositif paraisse satisfaisant à l'utilisateur. Les premiers résultats d'analyse obtenus à partir de ces ressources permettent d'en évaluer certains aspects quantitatifs et qualitatifs.

Si lorsqu'un dispositif paraît satisfaisant, une table (donc une catégorie) présente plus de lignes vides que de lignes renseignées, il convient de s'interroger :

- sur la pertinence des valeurs des attributs utilisés : ont-elles les propriétés énoncées par le modèle ?
- sur la pertinence de choix des attributs dans le cas où plusieurs attributs sont utilisés pour la catégorie : existe-il des relations entre eux qui ne permettraient pas de les distinguer ?

Ces interrogations sont facilitées par la représentation en topique qui, contrairement à celle en table, montre le nombre de différences mises en jeu entre les lignes d'une même table. Le repérage de la non-pertinence d'une distinction entre types est ainsi facilité.

[116] Contrairement au modèle ANADIA [Beust, 1998* : 100-102], les dispositifs LUCIA ne souffrent cependant pas de la présence de tables majoritairement vides si elles sont à la fois en adéquation avec les contraintes du modèle et à la fois signifiantes et utiles pour la tâche et l'utilisateur. La table « Agents, activités » du dispositif « météo » (figure 65 p.159) n'a par exemple que deux lignes renseignées sur quatre possibles (ce qui favorise en l'occurrence la redondance des attributs [Action] et [Rapport à l'activité] au sein du dispositif dans son ensemble).

4.6 Conclusion

Dans ce chapitre, nous avons présenté différents aspects de l'élaboration des ressources pour le modèle LUCIA. La tâche pour laquelle le modèle est utilisé influence toutes les étapes de la construction des dispositifs : de l'acquisition d'entités lexicales depuis un corpus d'observation à la structuration à l'aide du logiciel LUCIABuilder. La dernière partie de ce chapitre nous a permis de relativiser la représentation en tables du modèle et de proposer des mécanismes supplémentaires d'assistance automatique à la construction et la révision de dispositifs.

La pertinence et la justesse des descriptions obtenues avec le modèle LUCIA sont évaluées par l'utilisateur. Leur cohérence par rapport aux contraintes du modèle peut être assistée par le logiciel qui repère d'éventuelles inadéquations et suggère certaines modifications (c.f. 4.4.2). Dans l'exemple du dispositif en rapport avec la météorologie (figure 65 p.159), on peut questionner la description de l'entité *tempête de neige* : une « baisse de température » peut tout autant décrire cette entité lexicale qu'une « montée en agitation », ce qui d'une certaine manière remet en question le choix du jeu d'opposition des « axes » *agitation* et *température*. De telles remises en question font partie du processus de structuration. Une science n'a pas pour but de produire des interprétables à volonté ; elle doit se contenter d'offrir des arguments, voire des outils pour instituer ces nouveaux interprétables comme objets de pratiques socialement partageables [Kanellos, 2003 : 307]. Dans l'optique d'une *instrumentation du sens* telle que nous l'avons défendue dans [Perlerin et Beust, 2003*], nous nous attachons à rendre nos propositions utilisables par des tiers, c'est pourquoi nous avons été amené à proposer un protocole de construction des dispositifs.

Nous avons souligné à plusieurs reprises la souplesse et l'ouverture du modèle. LUCIA peut être utilisé avec ou sans les *stop-lists* ou la base de données lexicales MHATLEX. Tout comme PASTEL de Tanguy [Tanguy et Thlivitis, 1996*], l'utilisateur de LUCIA n'est pas contraint par des formes figées de la langue et choisit lui-même le domaine sémantique intéressant pour son application. Un corollaire important à ce type d'approche est que l'usager peut entrer toutes les informations qu'il souhaite dans les dispositifs. Dans PASTEL, ce sont des contraintes formelles fortes en forme de « requêtes d'éclaircissement » et structuration en fonction des principes de la SI demandées par le machine qui poussent l'utilisateur dans des retranchements interprétatifs et l'empêchent d'être incohérent dans ces descriptions. Dans LUCIA, ce sont à la fois les contraintes du modèle et les nécessités de la tâche courante qui contraignent l'utilisateur. Le rôle d'assistance du logiciel concerne la structuration des connaissances de l'utilisateur. *On a souvent défini l'intelligence comme l'art de faire des rapprochements. Peut-être que l'on peut envisager des ordinateurs intelligents qui feraient pour nous des rapprochements et nous laisseraient la liberté de confirmer ou d'infirmer. Car la difficulté de l'ordinateur (…), c'est bien qu'il croule sous les informations et trop d'information n'est pas de l'information (…)* [Rastier *et al.*, 1995]. Lors des analyses, les processus automatiques feront bien des *rapproche-*

ments à partir des données fournies par l'usager et s'il y a de l'intelligence dans ces processus, elle proviendra avant tout de l'interaction. Dans le chapitre suivant, nous présentons des exemples de ces analyses. Nous verrons comment elles sont utilisées par les logiciels pour assister l'utilisateur dans des tâches documentaires.