Chapitre 3

Le modèle LUCIA

Dans le chapitre précédent, nous avons présenté les fondements théoriques de LUCIA. Nous savons désormais que ce modèle est un support de catégorisation différentielle permettant à un utilisateur de décrire et organiser des éléments de signification associés à des entités lexicales. La figure 7 présente cette étape de construction des ressources du système. L'utilisateur va être assisté par les logiciels à partir de ses connaissances et/ou d'un corpus d'observation pour produire un dispositif LUCIA relatif à une tâche documentaire. Les dispositifs sont constitués d'entités lexicales associées à des éléments de signification selon un modèle de catégorisation précis que nous détaillerons dans ce chapitre.

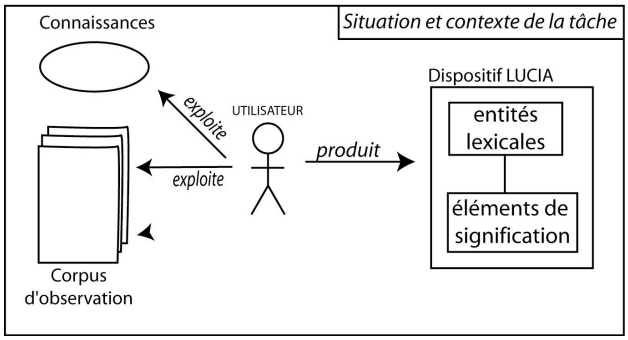

Figure 7 – Construction des ressources.

La construction des ressources est un préalable indispensable à l'utilisation du système (étape 1 sur la figure 8). Les étapes suivantes, i.e. l'utilisation de ces ressources, peuvent différer selon les spécificités des tâches pour lesquelles le système est utilisé. Les informations représentées dans le dispositif sont utilisées par les logiciels pour analyser des documents inconnus[55] par recherche des entités lexicales et projections des éléments de signification associés (étape 2 sur la figure 8). Cette projection peut servir à pressentir des informations relatives au contenu de documents inconnus. À ce stade, les éléments de signification des dispositifs sont donc des éléments de significations *potentiels* puisque leur validité en contexte ne peut être au final qu'évaluer par l'utilisateur. Les logiciels peuvent assister cette évaluation par la construction d'un certain nombre d'interfaces de lecture des ensembles de do-

[55] Ou simplement à analyser s'il s'agit de documents que l'on désire voir au travers des interfaces de visualisation et d'interaction que LUCIA permet de construire pour mieux étudier certains phénomènes.

cuments et des documents (étape 3 sur la figure 8) qui permettront à l'utilisateur de décider d'exploiter ou non les documents retenus.

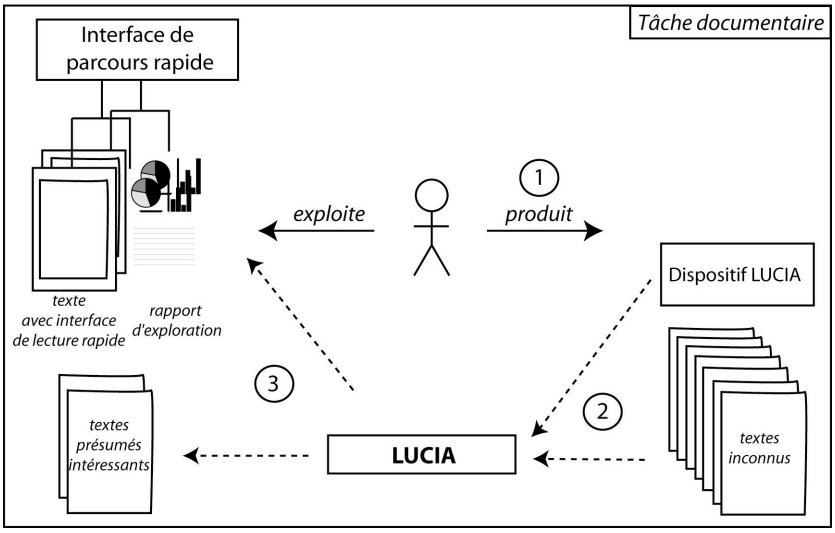

Figure 8 – Schéma global du système.

Dans ce chapitre relatif au modèle LUCIA en tant que modèle de catégorisation et d'association d'éléments de signification à des entités lexicales, nous débutons par la présentation de ces entités en précisant leur nature et les moyens envisagés pour les manipuler d'un point de vue informatique (3.1). Nous apportons ensuite des informations précises sur les éléments du modèle de catégorisation, i.e. sur les attributs, les tables et les dispositifs (3.2). Du fait de notre optique applicative, nous étudions ensemble ce qui relève de la conception théorique de ces éléments et ce qui concerne leur implantation effective. Dans la partie suivante (3.3), nous montrons en quoi le modèle de catégorisation et de représentation lexicale permet l'expression d'un point de vue et même d'un jugement par rapport aux entités lexicales retenues. Pour concevoir des outils accessibles aussi bien à des spécialistes de la langue qu'à des novices, nous avons été conduit à mener une première expérience sur la possibilité d'élaborer de telles structures. Nous en rendons compte dans la partie 3.4. Comme dans le chapitre précédent, nous illustrons nos propos à l'aide des deux tâches que nous avons retenues pour l'évaluation de nos propositions : la veille documentaire et l'analyse d'un fait de langue.

3.1 Entrées lexicales du système

De nombreuses propriétés des textes influencent les interprétations que l'on peut en faire. Aussi bien la syntaxe et la grammaire que les informations connues sur l'auteur, la date et les circonstances de la rédaction peuvent constituer des facteurs d'interprétation. Dans le chapitre précédent, nous avons pu voir que pour le TAL, la priorité était la plupart du temps donnée à la syntaxe - l'influence de Chomsky en la matière n'y est pas étrangère. En renonçant à la vision compositionnelle du sens inhérentes aux approches avant tout fondées sur la syntaxe, nous donnons dans nos travaux la primauté à la sémantique ; ce ne seront alors plus les phrases mais les textes et les ensembles de textes qui constitueront notre principal niveau d'analyse. Mais les textes sont des objets complexes ; les appréhender dans leur globalité et dans toutes leurs dimensions n'est pas compatible avec nos objectifs, puisque cela suppose soit la prise de conscience et l'explication d'opérations compliquées, soit la formalisation de ces opérations qui ne semblent pas être toutes déterministes. Le grain de départ de nos analyses seront les signes situés dans les textes signifiants pour l'utilisateur par rapport à sa tâche. Il pourra s'agir de noms propres et communs (sous forme complète, abrégée), d'adjectifs, de verbes et d'adverbes ou toutes combinaisons syntagmatiques redondantes et remarquables de ces éléments ; nous admettrons la présence de mots grammaticaux dans ces combinaisons. Les entités lexicales du système pourront également correspondre à des acronymes. Par exemple, *société*, *back office*, *rue Vivienne*, *chiffre d'affaire*, *CAC 40* et *réchauffement* pourront être utilisés comme entrées lexicales de LUCIA. En SI, où l'on considère dans un principe herméneutique que le global influence le local, les textes sont appréhendés dans leur globalité à la lumière des influences sémantiques entre leurs composants. Ces composants sont les morphèmes, qui sont les signes minimaux, indécomposables dans un état synchronique donné. Par exemple, *rétropropulseurs* compte cinq morphèmes. Du fait des références théoriques à la SI, il sera nécessaire de justifier dans cette partie le choix des entrées lexicales comme mots ou suite de mots et pas comme des morphèmes. Cette justification sera l'objet de la partie 3.1.1. En 3.1.2, nous exposerons les contraintes opératoires inhérentes à la manipulation informatique de tels signes. Ceci nous permettra dans la section suivante (3.1.3) de dresser la liste des critères de sélection de ces entités qui pourront être extraites d'un corpus d'observation ou provenir simplement des connaissances de l'utilisateur.

3.1.1 Lexique de mots *versus* lexique de morphèmes

L'analyse morphématique se heurte au caractère non compositionnel et non prédictif de la construction d'un signifié, d'un mot en fonction des signifiés de ses morphèmes. L'exemple de Corbin [Corbin, 1988] est explicite à cet égard : il est difficile de trouver un contexte où *pommade* aurait trait à une « préparation à base de pomme » selon le modèle d'*orangeade* malgré la présence du morphème-suffixe *–ade* (notons que cette vision simple de la compositionnalité n'est pas celle défendue par Corbin). Touratier [Touratier, 2003] tente de renouveler les recherches sur le thème en proposant différentes approches pour l'analyse morphématique des signifiants. L'auteur soumet des analyses descriptives sur une dizaine de langues. Ces analyses donnent un aperçu des problèmes auxquels le linguiste doit faire face lorsqu'il essaie de fonder la description des langues sur le morphème (et non plus sur le mot). L'objectif est principalement de montrer comment l'analyse en morphèmes est un biais intéressant pour la description grammaticale. Malheureusement, du point de vue de la construction des signifiés, l'auteur s'abstient de toute systématisation des phénomènes décrits : d'une part, les analyses proposées se limitent généralement à de courtes expressions ou énoncés examinés hors contexte et d'autre part de nombreux morphèmes apparaissent comme « sans signifiant » propre et ne sont donc pas analysés à la lumière de leur participation au signifié du mot dans lesquels ils apparaissent. Dans les travaux de Valette [Valette, 2003*] relatifs à la détection et l'interprétation automatique de contenus illicites sur Internet (sites racistes ou révisionnistes par exemple), l'auteur étudie la redondance de morphèmes dans des textes pour en évaluer le contenu. L'auteur a repéré dans son corpus certains morphèmes qu'il qualifie de péjoratifs (comme « -ouill- ») ou de vulgaires (comme « foutr- »). Il constitue ainsi un dictionnaire des principaux morphèmes utilisés dans la construction des lexies racistes pour au final, détecter non plus les lexies elles-mêmes mais les combinaisons de morphèmes leur correspondant. Le genre des sites racistes est généralement pamphlétaire : les morphèmes sont exploités pour distinguer les sites racistes des sites anti-racistes pour lesquels le taux de recouvrement des occurrences de certains mots est important. Le cadre d'étude de Valette est fixe ; l'approche morphématique proposée nécessite un important travail de repérage et de classification préalable dans un corpus particulier. Ce travail relève d'un spécialiste. L'approche semble prometteuse mais elle se heurte à la mise en corrélation systématique d'un « fond sémantique » avec un morphème donné, ce qui limite la généricité des données. S'intéresser au morphème dans un cadre computationnel informatique amène souvent à rejoindre les techniques de *stemming* (ou racinisation) utilisées couramment pour la recherche automatique d'information. Le *stemming* consiste à associer un ensemble de termes à une pseudo-racine, comme par exemple, *européen*, *Europe*, *européanisation (…)* à la racine *europ* pour simplifier les textes en particulier lors des phases de recherche d'occurrences de termes. Si cette technique, souvent non linguistiquement fondée dans ses réalisations concrètes, augmente par exemple le rappel de certains systèmes documentaires [Loupy, 2000*] où l'approche du contenu des textes reste approximative, elle peut s'avérer source d'ambiguïté et de pertes d'informations dans le cadre

d'analyses plus précises du contenu telles que nous les proposons. Ceci a entre autres été souligné dans [Grabar et Zweigenbaum, 2000].

La SI n'envisage comme lexique pouvant être rapporté à la langue que celui des morphèmes : les lexiques des lexies et des phraséologies relèvent selon Rastier, de normes déterminées par le discours et par le genre dont tout texte relève [Rastier, 2001a* : 154]. La détermination partielle du discours et du genre peut être un des enjeux des tâches que nous proposons d'assister par l'ordinateur. Si ce n'est pas le cas, alors le discours et le genre sont stabilisés : celui du corpus d'observation sera celui des documents à analyser. Dans ces conditions, les combinaisons sélectionnées pour correspondre à nos entités lexicales apparaissent comme les seuls paliers minimaux manipulables, surtout lorsqu'il s'agit de proposer des instrumentations qui sont destinées non seulement à la linguistique informatisée mais aussi à l'informatique linguistique (au sens de *computational linguistics*). Si dans le premier cadre applicatif, le morphème semble pouvoir être l'objet de base de la description, il n'a pas pour autant été abordé comme tel dans la plupart des travaux informatiques antérieurs relatifs à la Sémantique Interprétative (voir [Beust, 1998*] , [Tanguy, 1997a] et [Thlivitis, 1998*]). La sémantique de la construction des signifiés en fonction des signifiés de leurs morphèmes semble dépendre de normes difficiles à traiter dans une optique computationnelle dans la limite des connaissances actuelles dans ce domaine, en particulier parce qu'il n'existe pas de dictionnaires, de grammaires ou de règles sémantiques des constructions morphémiques. L'approche morphématique nous semble d'autant plus en inadéquation avec notre façon de concevoir les tâches documentaires que la complexité des phénomènes tend à multiplier les ressources nécessaires aux systèmes sans pour autant garantir une plus grande effectivité qu'une approche purement lexicale. Elle semble donc incompatible avec une *sémantique légère*. Enfin, dans le cadre de l'informatique linguistique où les utilisateurs potentiels ne sont pas des spécialistes de la langue, le morphème semble non seulement très difficile à décrire en terme de significations mais aussi peu facile d'accès en tant que tel. Le choix d'une approche morphématique pourra cependant prendre forme à travers l'assimilation de plusieurs entités lexicales en terme d'éléments de signification potentiels associés. Par exemple, on pourra dans le modèle, choisir de ne pas distinguer *chant*, *chanson* et *chanter* si cela est conforme aux objectifs de la tâche.

3.1.2 Détermination des entités

Les entités lexicales du système sont soit des mots, soit des expressions composées. Leur utilisation dans un système informatique nécessite un certain nombre de choix opératoires qui sont l'objet de la partie 3.1. Dans cette partie, nous débuterons par l'examen de plusieurs caractéristiques communes à ces entités : la synonymie, le statut grammatical et la langue. Ensuite, nous considérerons deux points propres aux expressions composées : la permutation de leurs composants et la possibilité d'insérer des entités entre les éléments.

Dans les dispositifs, il est possible de ne pas distinguer des entités telles que *pattern-matching* et *recherche de motifs* ou *acheteur*, *acheter* et *acquéreur*. De même, les différentes formes graphiques d'une même entité (comme *CAC 40* et *CAC40*), ces formes abrégées ou acronymiques (*K7* pour *cassette* ou *MS* pour *Microsoft*) pourront être considérées comme identiques du point de vue des éléments de signification utiles pour la tâche qui leur est associée. D'une manière générale, la synonymie telle qu'on la définit couramment, n'est pas utilisée en tant que telle dans le système. Le principe de catégorisation que nous mettons en place s'appuie sur les différences et les points communs en terme d'éléments de signification associés aux entités par un utilisateur, à partir d'un corpus d'observation ou de ses propres connaissances. Si des attributs identiques sont retenus pour caractériser plusieurs entités et qu'ils ont la même valeur, il ne sera fait aucune distinction entre elles du fait de leur caractère synonymique couramment admis ou non, de leur langue ou de leur statut grammatical. Ce cas de figure correspond à la présence de plusieurs entités dans une même ligne de table. Si les pratiques du domaine d'intérêt de la tâche en font état, il est possible de placer sur la même ligne d'une table, des représentants de langues différentes et de statuts grammaticaux différents. Leurs significations sont alors jugées localement (dans la situation) identiques ou proches. Plus simplement, il est possible qu'il ne soit pas utile de les distinguer pour la tâche.

En ce qui concerne les flexions d'une entité, il sera proposé à l'utilisateur de fixer lui-même les limites du paradigme flexionnel à prendre en considération. Cette précision n'est pas sans conséquence car par exemple, *boulangère* ne signifie pas toujours *femme qui fait le pain* mais indique souvent qu'il s'agit de la *femme du boulanger*, de même que la féminisation d'un nom comme *péripatéticien* peut faire glisser une tâche de considérations aristotéliciennes à d'autres moins philosophiques. L'exemple de la boulangère peut être modéré selon l'évolution diachronique de la langue qui autorise de nos jours la féminisation de nombreuses professions. Il s'agit là d'un argument supplémentaire pour considérer la signification d'un terme comme dépendante des pratiques sociales et de l'évolution sociale des usages dans une langue. L'utilisation de données fixées *a priori* comme des dictionnaires par exemple permet difficilement la prise en considération de cette dimension diachronique de la signification. En permettant la construction des ressources par l'utilisateur, nous allouons au système la capacité à saisir les phénomènes de langue au moment où ils évoluent. Dans les parties de ce tapuscrit relatives à l'implémentation du système, nous verrons que la tâche de détermination des flexions d'une entité sera assistée par l'utilisation d'une base lexicale. Les différentes formes d'une entité donnée appartenant à cette base seront proposées automatiquement : à charge pour l'utilisateur de ne sélectionner que celles qui sont conformes à ses attentes (pour aller plus vite, il pourra, dans la plupart des cas, valider les propositions de logiciels).

La linéarité des entrées lexicales composées est considérée comme fixée, ce qui signifie en terme de stockage et d'appariement lors des analyses, que l'ordre des termes des entités complexes se-

ra inchangé. Aucune permutation n'est envisagée dans l'ordre des éléments d'une entité : seules les combinaisons syntagmatiques présentes dans les dispositifs seront prises en considération. Il est néanmoins possible de tenir compte de différentes formes morphosyntaxiques ou graphiques d'un même élément au sein d'une entité complexe. Dans ces circonstances, l'insertion d'un élément au sein d'une entité lexicale complexe dans un texte à analyser rend cette entité inaccessible aux traitements puisqu'elle ne sera pas reconnue. Nous verrons dans la partie 3.1.3 relative aux critères de sélection des entités que le critère de récurrence des entrées lexicales au sein des textes est primordial. Si différentes configurations syntagmatiques qui relèvent d'une même entité sont redondantes dans le corpus d'observation ou particulièrement significatives pour le lecteur/utilisateur, elles pourront être placées dans le système. Si l'utilisateur ne parvient pas à les différencier en terme de signification, ces configurations pourront alors être placées, comme nous le montrerons plus tard, dans une même ligne d'une même table. En revanche, si ces configurations ont des significations jugées utiles à différencier, elles pourront être distinguées dans les tables de catégorisation produites à cette fin. *Chiffre d'affaire* et *affaire de chiffres* relèvent par exemple de ce type de distinction. La permutation syntagmatique des éléments de ces entités amène généralement à les interpréter différemment en contexte. Si les critères de sélection sont satisfaits par des ceux entités (c.f. partie suivante), elles pourront être placées dans un dispositif LUCIA. Au cours des analyses, c'est l'ordre des éléments qui apparaît dans le dispositif qui prévaudra : il n'y aura pas de confusion possible entre les deux entités même si elles sont finalement composées des mêmes mots. Toutes ces considérations à propos des entrées lexicales de LUCIA, nous permettent, du point de vue de l'implémentation informatique, de limiter à la fois les ressources et les traitements ; nous nous fondons principalement sur la récurrence (et la différence) pour assurer le fonctionnement des analyses automatiques. Ces prises de position sont elles-aussi liées à notre approche d'une *sémantique légère* défendue dans le premier chapitre. Cependant ce choix ne nous permet pas de prendre en considération l'insertion d'un adjectif ou d'un adverbe entre les éléments d'une lexie complexe (utiliser des expressions régulières pour reconnaître les termes composés en cas d'insertion aurait considérablement alourdi les algorithmes). Nos calculs ne souffrent pas dans des proportions suffisantes de cette imprécision pour envisager ce phénomène - en particulier, parce que ils se basent avant tout sur la récurrence. Dans le cadre des logiciels d'étude permettant d'évaluer l'intérêt de nos techniques, ces cas n'ont donc pas été envisagés.

3.1.3 Critères de sélection

Nous verrons dans le chapitre 4, relatif aux phases de constitution des ressources dans une pratique donnée, comment la tâche documentaire peut être définie et comment la construction des dispositifs LUCIA peut être amorcée par l'étude d'un corpus d'observation à partir duquel seront sélectionnées les entités lexicales. Nous pouvons d'ores et déjà dégager un certain nombre de critères pour que des entités lexicales puissent présenter un intérêt pour une tâche documentaire donnée. Nous utili-

sons la veille documentaire et l'analyse d'un fait de langue (une métaphore conceptuelle) comme exemples.

Les critères de sélection des entités lexicales sont :

- *leur signifiance pour le lecteur dans le cadre de sa tâche.* L'utilisateur exprime, à travers la création des ressources, à la fois un point de vue et un besoin. Par exemple, dans un cadre de veille documentaire, les entités lexicales sélectionnées doivent correspondre soit à la façon dont le lecteur/utilisateur parle de son sujet d'intérêt, soit à la façon dont il souhaite le voir aborder. Dans le cadre de l'étude d'une métaphore conceptuelle, elles doivent relever, comme nous le verrons plus tard, du domaine cible ou du domaine source. La signifiance pour le lecteur apparaît donc comme le critère primordial pour la sélection des entités.
- *leur récurrence au sein des textes étudiés et/ou attendus.* Les méthodes d'analyse distributionnelle, fondées sur la mesure statistique des occurrences, sont un moyen efficace pour détecter les thèmes des textes [Rastier, 1995]. Cependant, le palier lexical ne constitue que la première étape du processus permettant d'accéder à la dimension sémantique de ces thèmes. La récurrence d'une entité lexicale au sein d'un corpus thématique est un indice potentiel de l'appartenance de cette entité au contenu thématique des textes du corpus et elle peut être repérée comme l'un des supports d'une isotopie générique permettant de découvrir ce thème. En retour, la présence redondante de cette entité lexicale au sein d'un texte peut être un indice de la présence de ce thème particulier dans ce texte. Ces deux aspects amènent donc à considérer la récurrence dans un corpus étudié comme un critère pour la sélection des entrées lexicales du système. Ce critère fait également référence au figement qui est utilisé en linguistique pour désigner la lexie complexe et le syntagme lexicalisé, voire l'expression idiomatique et la locution[56] et justifie nos propositions de traitement quant à la permutation des éléments d'une expression ou l'insertion d'un élément dans une entité lexicale complexe.
- *la capacité de l'utilisateur à les décrire en terme d'éléments de signification.* L'isotopie est le principe fondamental de la SI pour l'explicitation d'une interprétation. La cohésion d'un texte (son intelligibilité) dépend des isotopies [Rastier, 1987* : 156] et [Beust, 1998* : 138] - l'interprétation consiste alors à relever les isotopies d'un texte et donc à mettre en évidence sa cohésion. Pour qu'une entité lexicale puisse être placée dans les structures LUCIA, elle doit donc pouvoir faire l'objet d'une description en terme d'attributs et de valeurs d'attributs. Les analyses proposées se fondent sur le principe de l'isotopie, qui correspond à la récurrence syntagmatique d'un élément de signification. Il sera donc possible de sélec-

[56] On pourra trouver une liste détaillée des termes linguistiques relatifs au figement dans [Kocourek, 1982] cité dans [Mejri, 2000].

tionner une entité peu redondante dans un corpus d'observation par exemple si elle peut être décrite à l'aide d'un élément de signification utile à la tâche (voir critère suivant) ; il faudra cependant pourvoir évaluer l'intérêt de cette insertion à l'aune de la quantité de données véritablement nécessaire à la tâche. Les résultats des analyses pourront apporter des indices pour cette évaluation.

Le critère primordial pour la sélection des entités est celui de leur signifiance pour l'utilisateur dans le cadre de sa tâche, les autres critères sont plutôt des gages de l'utilisabilité d'une entité par le système car il s'avère peu aisé de les utiliser pour une sélection initiale à partir d'un corpus d'observation.

Dans cette partie, nous avons vu quelles pouvaient être les entrées lexicales du modèle. Malgré l'influence de la SI sur nos travaux, nous avons montré pourquoi elles correspondaient non pas à des morphèmes mais des groupements stables de morphèmes (des lexies au sens de la SI). Nous avons également montré comment ces entités lexicales seront manipulées dans le système et quels étaient les critères utiles à l'utilisateur pour les sélectionner depuis des textes. Dans les parties suivantes, nous allons décrire les moyens mis en œuvre pour les décrire en terme d'éléments de signification et par-là même pour les placer dans des structures catégorielles. Avant cela, nous préciserons comment il est possible d'exprimer un point de vue voire un jugement sur les textes initialement analysés pour l'extraction, à travers l'association entités lexicales / attributs.

3.2 Modèle de catégorisation et de description

Dans cette partie, nous présentons en détails les éléments du modèle de catégorisation envisagé comme support de descriptions d'entités lexicales. Dans le premier chapitre, nous avons introduit les termes désignant ces éléments : les entités lexicales sont décrites à l'aide d'*attributs* (3.2.1), agencés au sein de *tables* (3.2.2) et elles-mêmes regroupées au sein de *dispositifs* (3.2.3). Les concepts relatifs à ces termes feront chacun l'objet d'une partie ce qui nous permettra à la fois d'en présenter une définition et les propriétés. Nous aborderons également leur nature théorique et les détails de leur conception informatique. En effet, notre approche applicative nous invite à considérer ensemble la conceptualisation théorique du modèle de son élaboration logicielle.

3.2.1 Les attributs

3.2.1.1 Définition, propriétés

Dans le chapitre 2, nous avons présenté notre approche comme inspirée de la sémantique componentielle : dans les dispositifs LUCIA, les éléments de signification associés à une entité lexi-

cale dans le contexte d'une tâche sont décrits par composition de valeurs d'attributs. Nous avons également déjà affirmé qu'étant donnés nos objectifs, les contraintes de rigueur et d'effectivité inhérentes à une modélisation informatique, nous ont amenés à nous écarter un tant soit peu d'une théorie linguistique descriptive et à adopter une démarche opportuniste par rapport aux propositions de la SI. Nous avons ainsi rejoint les propositions de redéfinition du sème oppositionnel de Beust. Dans le modèle LUCIA, l'utilisateur est sollicité pour exprimer des relations d'ordre sémantique qui seront utilisées par les logiciels pour l'assister dans une tâche documentaire. Ces relations peuvent donc être considérées comme une compétence de la machine qui a charge, non seulement d'en faciliter l'expression et d'en assurer la cohérence, mais également de les exploiter au mieux lors d'analyses de contenu de textes. Comme Beust, nous sommes amenés à définir un modèle des éléments de signification qui dépasse la simple entité sans contenu accessible pour les logiciels : c'est ce que nous appelons les attributs.

L'utilisation du mot attribut provient de la définition linguistique du terme rappelée en ces termes par Jacques Poitou : *On appelle attribut une propriété qui a un nombre limité de valeurs qui s'excluent mutuellement*[57]. Le rôle des attributs dans LUCIA est double : ils portent des informations qui ont orienté une interprétation humaine et ils structurent les représentations utilisées par les logiciels. Les attributs/valeurs du modèle peuvent s'apparenter aux sèmes, en particulier ceux de la Sémantique Interprétative mais présentent cependant quelques différences essentiellement dues à leur statut computationnel. Voici quelques assertions relatives à leur définition, leur forme, leur nombre et leur zone de validité exposées de façon contrastive (voire différentielle) vis-à-vis d'autres propositions analogues ou proches :

Définition des attributs

Dans la SI, le sème est défini comme l'extrémité d'une relation fonctionnelle binaire entre signifiés de morphèmes [Rastier, 1987*]. Dans notre approche computationnelle, le sème n'est plus considéré comme une entité autosuffisante issue de l'interprétation mais plutôt comme une entité relationnelle, structurante au niveau des représentations. Ce qui paraît pertinent alors pour fonder une représentation en machine, ce n'est pas uniquement le trait sémantique en lui-même, mais également ce à quoi on peut l'opposer. Dans LUCIA, les attributs sont à la fois des éléments de signification pouvant participer à l'interprétation d'un texte et des propriétés sémantiques associées à des entités lexicales conçues comme relation entre des significations [Nicolle *et al.*, 2002* : 43]. L'attribut est alors avant tout un élément de différenciation et de mise en corrélation entre signifiés au sein d'un projet de description, d'un projet de construction de dispositif.

[57] http://perso.univ-lyon2.fr/~poitou/Morpho_Lexico/6_sem-lex.html

Lors des analyses informatiques, les attributs seront considérés comme des éléments potentiels de significations. La potentialité sera confirmée ou infirmée par l'utilisateur aidé en cela par les résultats des analyses automatiques fondées sur la récurrence des attributs et des valeurs d'attributs au sein de textes inconnus de l'utilisateur.

Forme des attributs

L'expression formelle d'un sème est une paraphrase méta-linguistique [Rastier, 1987*]. Les attributs LUCIA sont également exprimés par des paraphrases considérées comme méta-linguistiques. Cependant, ils se présentent sous la forme d'un couple : [domaine d'interprétation : jeu de valeurs] où le domaine d'interprétation (au sens de [Beust, 1998*]) s'exprime à l'aide d'une paraphrase métalinguistique et où les valeurs s'excluent mutuellement et sont interprétables selon le contexte proposé. Le domaine d'interprétation représente le domaine dans lequel l'opposition entre les valeurs prend sens. La relation qui unit un attribut à ses valeurs est définitoire. Une valeur exprimée par la même paraphrase dans deux attributs différents ne sera pas la même si le domaine d'interprétation est différent. Par exemple, le mal et le bien de l'attribut [Morale : bien *vs.* mal] et celui de l'attribut [Santé et bien-être : bien *vs.* mal] ne sont pas les mêmes – nous avons pu le montrer dans [Nicolle *et al.*, 2002* : 48].

Un attribut est signifiant pour l'utilisateur qui le crée (ou l'utilise). La forme [D : v1 *vs.* v2 *vs.* …] (avec D pour domaine d'interprétation et v pour valeur) pourrait même être envisagée avec d'autres moyens d'expression que le linguistique : pour certains attributs évaluatifs, on pourrait faire appel à des émoticônes (c.f. p.251). Ce faisant, nous nous heurtons à un problème inhérent au sème : les moyens de son expression relèvent de la langue si bien que le serpent risque de se mordre la queue. L'idée a été maintes fois développée. Confronté à cette problématique, Greimas [Greimas, 1974 : 13] utilisait une comparaison à caractère judiciaire : *l'objet de l'étude se confond avec les instruments de cette étude : l'accusé est en même temps son juge d'instruction*. Ces propos faisaient eux-même échos à ceux de Wittgenstein, plus pessimiste encore, qui affirmait que *la langue est une cage dont on ne peut sortir* [Wittgenstein, 1922]. Pour les attributs, leur caractère fondamental est leur signifiance pour l'utilisateur. D'une certaine manière, ils sont également signifiants pour les logiciels car ils peuvent être manipulés (forcément de façon symbolique) et distingués. On reproche au modèle du sème sa circularité métalinguistique ; or pour écarter cette critique, il suffit d'établir une distinction entre la métalangue, qui n'est pas analysée par le modèle, et la langue, qui est à la fois source de production des structures et objets des analyses. Les simples identités et différences de deux chaînes de caractères (les valeurs d'attributs par exemple) ne seront pourtant pas suffisantes pour effectuer des analyses automatiques : le modèle informatique utilisé doit tenir compte du lien entre attributs et valeurs. Nous verrons plus loin qu'il s'agit là d'une des justifications de l'utilisation d'un langage de programmation orienté objet pour l'implantation du système correspondant et qu'ainsi, la modélisation informatique de LUCIA s'inspire, non de la logique ou des mathématiques comme cela a été souvent le cas dans le do-

maine, mais de la modélisation objet et des diagrammes de structure UML[58] (*Unified Modeling Language*).

Les attributs se caractérisent par leurs valeurs. Comme le préconise la définition de Poitou donnée p. 78, les valeurs des attributs doivent s'exclurent mutuellement. Dans le modèle, il est possible de construire des attributs qui n'entrent pas strictement dans le cadre de cette définition. Cependant, les attributs qui contiennent une valeur neutre ou qui nient une ou plusieurs de ses autres valeurs, doivent être utilisés avec précaution. Nous verrons dans la partie 4.4 du chapitre 4 (p.160) qu'ils peuvent être à l'origine d'incohérences au sein des structurations en dispositifs. Par exemple, des attributs tels que [Évaluation : bien *vs.* mal *vs.* pas évalué] ou [Direction : monte *vs.* descend *vs.* stagne] sont envisageables dans le cadre de LUCIA mais peuvent amener à des incohérences d'association au sein d'un même dispositif. Nous le rappelons ici : les attributs ont un double rôle. Ils sont à la fois structurants au niveau de la mise en place des tables d'un dispositif, et ils participent également à la caractérisation sémantique des éléments des tables pour lesquelles ils ont été choisis. Si un attribut est choisi pour une table donnée, toutes ses valeurs doivent alors avoir une pertinence pour la caractérisation des éléments de cette table. Précisons dès maintenant, que les tables formées à partir d'un ou plusieurs attributs auront le statut de catégorie, et que l'assertion précédente correspond au fait que si un attribut est utilisé pour exprimer une catégorie, toutes ses valeurs doivent avoir une pertinence pour tous les éléments de cette catégorie.

Nombre des attributs

Dans la SI, le nombre des sèmes est indéfini et non infini. Ce nombre dépend de l'organisation des signifiés du fait de leur statut d'extrémité de relations entre des signifiés. Dans le cadre de LUCIA, le nombre d'attributs est dépendant du projet de catégorisation de l'utilisateur et de sa capacité à décrire les entités lexicales qu'il a retenues pour sa tâche à un instant donné. Ce nombre est non borné. Un nombre très important d'attributs (de l'ordre du millier par exemple) n'est pas envisageable pour un dispositif : l'interaction et la présentation en grille de catégorisation est un frein, assumé et apprécié en tant que tel, à la multiplication des attributs pour un projet donné. Les attributs doivent être en nombre suffisant pour structurer les entités lexicales préalablement sélectionnées et répondre aux besoins de l'utilisateur quant aux détails de description d'un dispositif. Dans l'absolu, il est impossible de dresser une liste exhaustive des attributs utilisables au sein d'un dispositif comme il est impossible de dresser une liste exhaustive des sèmes en langue. La combinatoire des sèmes, problème algorithmique de l'approche componentielle, est restreinte par notre approche interactionniste de la sémantique des textes. C'est la fonction de mémoire de l'usager qui va limiter la prolifération des attributs et donc éviter l'explosion combinatoire. Pour un non-spécialiste, le temps consacré à la tâche

[58] http://www.omg.org/gettingstarted/what_is_uml.htm et http://www.uml.org

de construction des ressources LUCIA sera d'autant mieux accepté qu'il rendra par la suite d'éminents services automatiques. Il n'est pas nécessaire d'avoir beaucoup d'attributs pour obtenir d'emblée des résultats satisfaisants : nous nous attacherons à le démontrer dans le chapitre 5. Pour un spécialiste, la structuration en tables impose la composition de catégories représentées par les tables, et donc le regroupement d'entités lexicales, ce qui constitue une limitation importante à la multiplication infinie des attributs.

Zone de validité des attributs

Les attributs, comme les sèmes de la SI, ne sont pas des universaux. Ils n'ont pas le statut des noèmes, ces unités minimales de sens conceptuel définies par Pottier. Certains exemples proposés plus tard dans cette partie montreront que leur validité d'une langue à l'autre n'est pas postulée *a priori*. Un attribut donné n'est pas un outil à portée généraliste : il est choisi par un lecteur/utilisateur interprétant dans le cadre d'une tâche précise. Il n'y a aucune garantie qu'un autre interprétant pratique la même dénomination et en fasse une utilisation équivalente, même si les tâches sont les mêmes – seules des pratiques culturelles et sociales précises pourraient permettre cette identité. Du point de vue de leur portée, nous verrons en détail dans le chapitre 5 que pour l'étude sur la métaphore les empans principaux d'étude est la phrase et la paragraphe et que dans le cadre de la veille document, ces empans sont la zone textuelle (repérée comme telle par nos logiciels) et la document dans son ensemble.

Notre modèle est avant tout un modèle informatique, un modèle destiné à une implantation. Dans cette optique, la conception théorique doit s'articuler avec la conception computationnelle. Dans la partie suivante, nous verrons comment nous avons envisagé l'implantation des attributs, en particulier sur la base d'une modélisation objet. Rappelons une nouvelle fois, que les attributs LUCIA sont assemblés en tables pour définir des catégories au sein d'un dispositif et que les tables sont regroupées en dispositifs.

3.2.1.2 Conception et implantation

D'un point de vue pratique, nous avons fait le choix des langages Java et XML. Ces deux langages nous permettent d'obtenir une portabilité optimale pour nos développements logiciels. XML est utilisé ici pour ses capacités de structuration des données textuelles manipulées. Les représentations informatiques (attributs, tables, dispositifs…) sont toutes assurées par ce langage. Java en tant que langage objet, est adapté aux structures construites inspirées de UML et à la manipulation de structures XML grâce surtout à la présence de *packages* de classes permettant soit la manipulation des DOM (*Document Object Models*) soit une réification en instances de classes Java à l'aide d'analyseurs syntaxiques dédiés (*XML parsers*). Dans cette partie, nous présentons les détails de la conception informatique et de l'implantation des attributs.

Pour être réifiés, les attributs doivent pouvoir être mis en relation avec les valeurs qui les caractérisent et le domaine d'interprétation de ces valeurs. Ayant fait le choix de l'utilisation d'une programmation orientée objet et de la formalisation UML pour la conception informatique du modèle, nous présenterons d'abord le schéma UML de conception correspondant (figure 9). Ici, un attribut correspond à un domaine d'interprétation et un seul, et à au moins deux valeurs. Le domaine d'interprétation est défini par l'opposition de valeurs et est nommé pour des commodités de manipulation. Les relations qui unissent un domaine d'interprétation à un attribut et une valeur à un attribut doivent donc être des relations de composition ; les valeurs et un domaine d'interprétation ne peuvent exister indépendamment d'un attribut. Leur durée de vie est incluse dans celle de l'attribut.

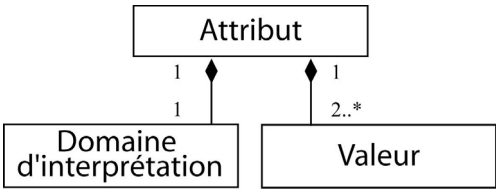

Figure 9 – Conception des attributs LUCIA

Pour le stockage en XML, les attributs sont repérés par un identifiant unique codé sous la forme `attr`$_i$ où `i` est un entier (figure 10, ligne 1). Le domaine d'interprétation d'un attribut, exprimé à l'aide d'une paraphrase ou d'une entité lexicale, est représenté par une simple chaîne de caractères. Du point de vue de la manipulation des données et de l'implantation, il est appelé *nom* de l'attribut (codé en XML à l'aide d'une balise `attrnom`, figure 10, ligne 2). Chacune des valeurs d'un attribut est repérée par un identifiant formé par la concaténation de l'identifiant de l'attribut correspondant et de la chaîne `val`$_j$ où j est un entier. La valeur elle-même est codée à l'aide d'une chaîne de caractères (figure 10, lignes 2 et 3). Les identifiants sont utilisées principalement pour ne pas surcharger les fichiers XML : un attribut peut apparaître dans plusieurs dispositifs, il sera ainsi possible de regrouper certains attributs dans des fichiers autres que celui d'un dispositif et de n'y faire référence que par l'intermédiaire des identifiants. Les identifiants permettent également de faciliter le traitement des fichiers XML à partir de leur *DOM* lors des analyses. La représentation XML d'un attribut comprend alors trois types de balises sur le modèle suivant :

```
1. <attr id="attr_i">
2. <attrnom>DOMAINE D'INTERPRETATION</attrnom>
3. <val id="attr_ival_j">VALEUR j</val>
4.  <val id="attr_ival_j'">VALEUR j'</val>
5. …
6. </attr>
```

Figure 10 – Représentation XML d'un attribut

La réification du code XML présenté ci-dessus donne lieu à la création d'une instance de la classe `Attribut` selon le diagramme de classes suivant (figure 11).

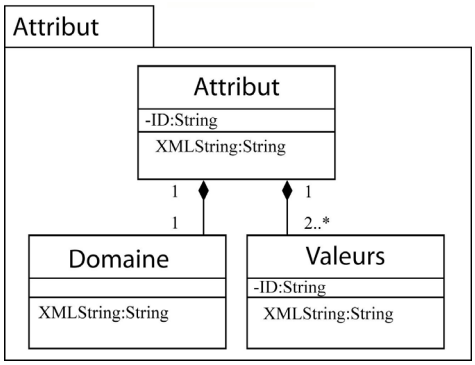

Figure 11 – Diagramme de classe pour les attributs.

Les trois classes présentées implantent toutes une interface assurant leur retranscription en XML (méthode `XMLString`) lorsque l'on n'accède pas directement au *DOM*. L'utilisation des identifiants présentés ci-dessus (`ID:String`), ainsi que la relation de composition entre les classes `Attribut`, `Domaine` (correspondant au domaine d'interprétation d'un attribut) et `Valeurs` instrumentalisent la relation stricte d'appartenance unique d'un domaine d'interprétation et de valeurs à un attribut. Les attributs sont définis dans un *package* éponyme parce qu'il peut être intéressant de créer des instances de ces classes dans un autre contexte que celui de LUCIA : leur implantation correspond en effet à la définition générique de Poitou (citée p.78). Nous nous attacherons dans le chapitre suivant à présenter le logiciel `LUCIABuilder` qui permet la construction interactive des dispositifs et donc en particulier, la construction des attributs en fonction des classes présentées ici. La partie suivante, présente les tables et une représentation graphique possible des tables : les topiques.

3.2.2 Les tables, les topiques

3.2.2.1 Présentation

À partir d'un ou plusieurs attributs, on peut créer une table LUCIA. Une table correspond alors à une catégorie. Les instances de cette catégorie seront les entités lexicales associées à chacune des combinaisons de valeurs de la table.

La relation entre les trois éléments attributs/valeurs, table et instance de catégorie correspondant à une ligne de table est circulaire. Un ou plusieurs attributs/valeurs sont associés à toutes les instances d'une catégorie ; en retour, les instances en question ont comme point commun d'appartenir à la même catégorie et donc d'être associées aux mêmes attributs/valeurs. Les propriétés caractérisant les catégories sont exprimées à l'aide d'attributs/valeurs. Pour décrire une catégorie, on peut utiliser plusieurs attributs. La combinatoire des valeurs de plusieurs attributs ou la liste des valeurs d'un seul attribut peut être représentée sous la forme d'une grille de catégorisation (une table). Dans les deux exemples proposés en figure 12 p.84 , on a créé deux catégories C_1 et C_2. La catégorie C_1 a été pro-

duite à l'aide des attributs A_1 et A_2 ayant respectivement 2 et 3 valeurs. Cette grille de catégorisation distingue six lignes, éléments de la catégorie C_1 : les six lignes sont obtenues à partir de la combinaison des valeurs des deux attributs utilisés. Les points communs des lignes de la catégorie C_1 sont caractérisés par les attributs A_1 et A_2. Leurs différences sont exprimées à l'aide des valeurs d'attributs correspondants (A1v1 et A2v1 pour la ligne c1 par exemple alors que la ligne c2 correspond à A1v2 et A2v1). La catégorie C_2 présente deux lignes différentes (c'1 et c'2) différenciées par les valeurs A3v1 et A3v2. Chaque ligne d'une table correspond elle-même à une catégorie.

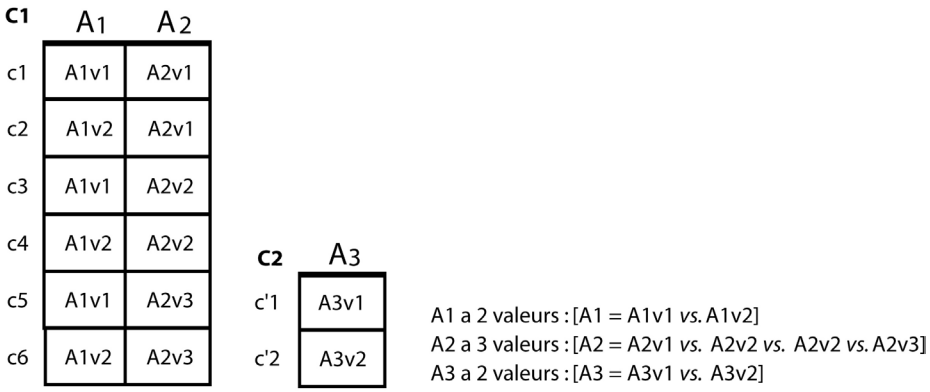

Figure 12 – Deux grilles représentant deux catégories C_1 et C_2.

Dans le chapitre 2, nous avons déjà vu que la méthode de catégorisation mise en place dans LUCIA emprunte beaucoup aux principes d'ANADIA. Le principe de catégorisation d'ANADIA ne s'appuie pas sur les propriétés intrinsèques des choses mais sur les différences qui existent entre elles dans la langue. Pour nous, ces différences (et ces points communs) seront celles proposées par l'utilisateur en fonction de ses connaissances, de ses pratiques langagières et éventuellement à partir de l'examen de ce qui fait sens pour lui et ce qui est intéressant pour lui dans un corpus d'observation. Le sème selon le modèle oppositionnel de Beust est utilisé dans ANADIA comme critère définitoire des éléments de la catégorie à laquelle ils participent. Par exemple, le sème [Nature de composant : software *vs.* hardware] permet de définir une catégorie que l'on peut nommer //Composants informatiques// et que l'on peut représenter sous la forme d'une table comme dans la figure suivante (figure 13, d'après [Beust, 1998* : 176]). La notation des noms des catégories (entre //) fait référence à la notion de *taxème* qui correspond pour la SI, à une classe de sémèmes minimale en langue.

//Composants informatiques//	**Nature de composant**
	hardware
	software

Figure 13 – Catégorie ANADIA, représentation en table

Dans les travaux de Beust, toutes les places d'une table sont la représentation d'un concept, *dans le sens d'objet mental résultant d'une opération algébrique (il ne nécessite pas obligatoirement*

l'existence d'objets) [*ibid.* : 94]. Dans ce cadre, les notions de concept et de catégorie sont distinguées comme suit : *pour qu'une catégorie soit attestée, on doit pouvoir donner des exemples d'objets tombant sous son concept*. Ainsi, la table présentée en figure 13 offre 2 places, qui déterminent 2 concepts et donc 2 catégories potentielles. On peut transformer ces concepts en catégories en associant par exemple 'système d'exploitation' à la valeur [software] de l'attribut et 'carte mère' à la valeur [hardware]. Beust s'appuie sur le fait que *rien en peut être représenté en langue qui n'ait auparavant été décrit en contexte* [Rastier, 1987* : 62] *in* [Beust, 1998* : 148] pour utiliser les principes que nous venons de décrire en vue de l'élaboration d'un modèle *de langue* dont les implantations ont pour but d'améliorer les systèmes de dialogues homme/machine. Les valeurs d'attributs sont des conditions nécessaires (mais pas suffisantes) pour différencier les concepts.

LUCIA reprend le principe de catégorisation proposé dans ANADIA, cependant les objectifs applicatifs n'étant pas les mêmes, certains concepts changent d'un modèle à l'autre. Pour LUCIA, l'acteur principal du processus de catégorisation est le lecteur/utilisateur dans la situation d'une tâche documentaire. Il est assisté via un certain nombre d'interfaces dédiées. Le processus de catégorisation a pour but de produire un ou plusieurs ensembles de tables (les dispositifs) à l'intérieur desquels les logiciels pourront puiser des informations relevant de deux dimensions distinctes et complémentaires. La première relève du lexique : les instances des catégories exprimées dans les tables sont des entités lexicales (considérées dans la représentation comme des *sémèmes*). La seconde représente le point de vue interprétatif de l'utilisateur sur les entités en question. Ce point de vue est exprimé en particulier à l'aide de couples attributs/valeurs qui sont l'expression d'éléments de significations associés aux entités dans le contexte de la tâche. L'association d'un ou plusieurs attributs/valeurs à un ensemble d'entités lexicales permet d'instituer cet ensemble en catégorie. Réciproquement, une catégorie donnée est déterminée par un ou plusieurs attributs. Contrairement à ANADIA, les catégories produites à l'aide de LUCIA peuvent être décrites à l'aide d'attributs dont les valeurs sont thymiques ou évaluatives ; cette possibilité n'avait pas été envisagée dans l'utilisation d'ANADIA proposée par Beust puisqu'il s'agissait essentiellement d'obtenir des représentations utilisables pour rendre plus naturelles les interactions hommes/machines en dialogue parlé et donc objectiver des significations dans un contexte donné. Ces traits ne doivent pas selon Rastier, être différenciés des traits descriptifs dans une théorie sémantique [Rastier et Malrieu, 2000] ; en effet, la SI rejette l'opposition ancienne entre dénotation et

connotation[59]. Il n'en reste pas moins une réticence à considérer comme *concept* de simples associations d'attributs à un ensemble des entités lexicales dans le cadre de LUCIA. Par exemple, un attribut comme [Évaluation : bien *vs.* mal] qui permettrait à un utilisateur de distinguer et de décrire des pays du monde ne saurait donner lieu à la définition d'un concept défini par l'une ou l'autre de ces valeurs (nous reviendrons sur cet exemple dans la partie 3.3). Nous nous soustrayons ainsi à une certaine validité lexicologique ou encyclopédique en privilégiant le point de vue – éventuellement singulier – de l'utilisateur pour mieux l'assister de façon personnalisée dans une activité langagière, dans une tâche documentaire. De même, les concepts de la SI ne seront pas nécessairement présents *stricto*-sensu au sein des représentations. Ainsi dans LUCIA, nous parlerons de catégories pour désigner des regroupements d'attributs que l'on peut représenter sous la forme de tables. Ces catégories pourront être l'équivalent de taxèmes qui traduiront une norme accessible à l'utilisateur mais dont la véritable dimension pourra lui échapper (l'exemple précédent relatifs à certains pays tend vers une norme apparemment très locale, dans le sens idiolectale ou tout du moins culturellement et historiquement très marquée). Les lignes de ces tables, seront-elles aussi des catégories dont les valeurs d'attributs seront l'équivalent pour leurs instances d'un sémantème (partiel) et sans validité autre que celle que lui conférera l'utilisateur (en SI, le sémantème est l'ensemble des sèmes spécifiques d'un sémème). Pour pouvoir distinguer ces deux types de catégories dans nos propos, nous nous référerons à la représentation en table en les désignant respectivement par table et ligne de table.

Pour un projet de catégorisation donné, l'étape de création des tables et donc des catégories est centrale. C'est à l'intérieur des tables que l'on met en valeur les éléments de signification (les attributs/valeurs) retenus comme caractéristiques d'une catégorie au sein d'un dispositif. La structure en tables permet de prendre en considération simultanément tous les attributs jugés pertinents pour une catégorie donnée, sans pour autant être pertinents pour la catégorie supérieure : le problème du choix de l'attribut à chaque étape de discrimination est ainsi évité. Le cas où une catégorie est partitionnée par les valeurs d'un seul attribut n'est pas le seul envisagé ici : c'est un retour non systématique aux arbres de catégorisation de Porphyre. Les exemples proposés ultérieurement (figure 14, p.84 par exemple) montrent qu'une catégorie peut être construite à partir de plusieurs attributs. La table correspondante est alors obtenue à partir de la combinaison des valeurs des attributs choisis. Cette combinaison correspond à un produit cartésien. Pour une même ligne d'une table, on peut avoir plusieurs entités

[59] Rastier précise alors que « Les réactions émotionnelles à un texte ne sont pas liées directement aux traits thymiques ou évaluatifs qui y sont déployés. Heureusement, sans quoi la propagande serait l'arme absolue. » [Rastier et Malrieu, 2000*]. Si l'on conçoit la dénotation comme un ensemble de propriétés que l'on peut inférer des propriétés inhérentes ou fonctionnelles d'un référent, et la connotation comme un ensemble des propriétés qui peuvent être associées à une entité lexicale par un locuteur ou un groupe de locuteurs, la distinction entre les deux, de toute façon problématique et critiquable, n'est pas de mise pour les attributs LUCIA. Ils peuvent en effet être utilisés pour mettre en lumière une propriété d'un référent invocable dans le contexte de la tâche à partir d'une entité lexicale donnée comme ils peuvent relever d'un effet de sens perçu par le lecteur en fonction de sa propre expérience.

lexicales, instances de ces catégories si l'auteur de la table n'a pas jugé utile de les distinguer. Dès lors, elles sont potentiellement porteuses des mêmes éléments de signification représentés par les valeurs d'attributs correspondant à la ligne en question. Elles peuvent cependant être distinguées à tout moment par la création éventuelle d'une sous-catégorie (voir 3.2.3). Les entités lexicales d'une même ligne sont des entités partageant des éléments potentiels de significations dans le contexte de la tâche. Dans certains cas, elles peuvent donc être :

- des entités considérées comme synonymiques dans le contexte de la tâche (*épargne* et *bas de laine* dans le domaine de l'économie par exemple) ;
- des entités relevant d'un même paradigme dérivationnel dont la distinction n'est pas justifiée pour la tâche du point de vue de leur contenu sémantique (*analyste* et *analyser*) ;
- des représentants de langues différentes utilisés (quasi-)indifféremment dans une langue donnée (*brainstorming* et *réunion de réflexion*) ;
- voire des entités à contenu sémantique identique pour des graphies différentes (par exemple *cassette* et *K7*). Pour cette dernière catégorie, nous préconisons, si l'emploi est systématique dans le domaine de la tâche, de considérer les graphies « exotiques » au même titre que les flexions (c.f. chapitre 4 partie 4.2.5 p.147).

Dans l'exemple de la figure 14, nous proposons une table nommée « Agents, Activité ». Le nom des tables permet de nommer les catégories produites au sein d'un dispositif. Chaque nom est propre à une catégorie au sein d'un dispositif donné. Il n'a originellement de signification au sein de la catégorisation que pour le lecteur/utilisateur, mais il peut bien entendu faire l'objet d'une négociation avec un ou plusieurs tiers. Dans cette table à deux attributs bivalués, toutes les lignes ont été renseignées et on y trouve aussi bien des verbes (a*nalyser*, *acheter*), que des noms (*actionnaire*, *achat*) et un acronyme (*COB*).

Agents, Activités	**Action**	**Rapport à l'activité**
petit porteur, acheter, acheteur, actionnaire, achat	intervient	rôle
analyste, analyser, analyse, estimation	étudie, analyse	rôle
agent de change, opérateur	intervient	profession
COB, économiste	étudie, analyse	profession

Figure 14 – Table LUCIA **« Agents et Activités » du domaine de la Bourse et de l'Économie.**

Le rapport entre les instances des catégories représentées par les lignes d'une table et les valeurs des attributs sélectionnés pour cette table n'a pas de caractère définitoire. Il est le reflet d'un point de vue issu d'une observation ou de connaissances sur un domaine. C'est une mise en lumière de caractéristiques pour un projet de catégorisation. Les critères utilisés peuvent être non seulement de nature ontologique mais aussi d'orientation plus culturelle, sociale et praxéologique. L'exemple de la

table de la figure 14 permet d'illustrer cette propriété du modèle : des entités lexicales du domaine de la Bourse et l'Économie ont été rassemblées au sein d'une catégorie en fonction de deux caractéristiques ; celle qui permet d'évoquer un rôle ou une profession du point de vue de l'activité à laquelle elles peuvent faire référence et celle qui autorise une action d'intervention, ou une action d'étude ou d'analyse dans le domaine.

La création de tables à l'aide du regroupement d'attributs amène parfois à l'apparition de lignes vides dans les tables. Ces lignes vides ne sont pas un obstacle à l'utilisabilité des dispositifs. Elles permettent de placer dans les tables des entités lexicales non repérées dans le corpus mais que l'utilisateur envisage comme adéquates à sa tâche. Une combinaison de valeurs d'attributs non encore associée à une instance peut être renseignées en fonction des connaissances du domaine ou laissée telle quelle.

Les relations différentielles entre les instances des lignes d'une même table sont quantifiables par le nombre de différences qui les séparent. La représentation en graphe de ces relations est appelée *topique* (cette dénomination s'inspire de [Coursil *et al.*, 2000*]). On peut construire les topiques à *n*-traits près, où *n* représente le nombre de valeurs d'attributs différentes entre ligne d'une même table. Dans une topique, chaque ligne est représentée par le sommet d'un graphe relié aux autres par les relations de différences au sein de la catégorie. Les figure 15 et figure 16 présentent les topiques à 1 trait près et à 2 traits près que l'on peut obtenir de la table présentée en figure **14.**

Figure 15 – Topique à un trait près de la table présentée en figure 14.

Figure 16 – Topique à deux traits près de la table présentée en figure 14.

Les topiques permettent de représenter les structures différentielles engendrées au sein d'une catégorie par les attributs considérés. Leur structure provient du nombre et de la nature des attributs à l'origine de la table. Dans les exemples ci-dessus, nous sommes en présence de deux attributs bivalués. La topique à un trait près de la figure 15 fait apparaître les instances des lignes de la table comme différenciées deux à deux par un seul attribut, tandis que la topique à deux traits prés de la figure 16 montre qu'il n'existe que deux relations différentielles entre lignes faisant intervenir les deux attributs.

Nous verrons dans le chapitre 4 comment nous exploitons les topiques pour l'assistance à la construction des dispositifs par un utilisateur.

Dans la partie suivante, nous exposons les détails de conception computationnelle et d'implantation des tables.

3.2.2.2 Conception et implantation

En suivant les indications du modèle, les tables doivent être créées par les logiciels à partir des attributs qui les composent. Une table est créée à partir d'un ou plusieurs attributs et la combinaison de leur valeurs forme des lignes auxquelles correspondront des entités lexicales. Le schéma simplifié de conception d'une table comprend de ce fait une `Ligne` à laquelle on pourra faire correspondre plus tard des entités lexicales. La relation qui unit un ou plusieurs attributs à une ligne est une relation d'agrégation car un même attribut peut être utilisé dans plusieurs tables donc il peut apparaître dans des lignes différentes. En revanche, une table est composée de lignes qui n'appartiennent qu'à elle (figure 17).

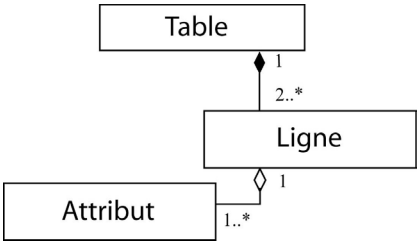

Figure 17 - Conception des tables LUCIA

Le code XML correspondant à la table LUCIA de la figure 14 est présenté dans la figure suivante (figure 18). La balise `tablenom` permet de nommer la table et d'en faciliter ainsi la manipulation.

```
1. <table id="disp_La_Bourse_att3-4" attrs="attr3 attr4">
2. <tablenom>Agents, Activités</tablenom>
3. <ligne id="disp_La_Bourse_tab3-4ligne0" vals=" attr3val0 attr4val0">
4.     <!– ici sont placées les entités lexicales de la première ligne -->
5. </ligne>
6. <ligne id="disp_La_Bourse_tab3-4ligne1" vals=" attr3val0 attr4val1">
7.     <!– ici sont placées les entités lexicales de la seconde ligne -->
       9. </ligne>
       …
16. </table>
```

Figure 18 – Extrait de la représentation XML d'une table.

Toutes les tables ont un identifiant au sein d'un dispositif donné (attribut `id` figure 18 - lignes 1) pour en faciliter la manipulation à partir du *DOM* et éviter de multiplier les informations redondan-

tes au sein des fichiers analysés. Par exemple, si le degré de détail désiré par l'utilisateur est celui de la table, on pourra mettre en correspondance une entité repérée dans un texte simplement avec l'identifiant de la table correspondante. Les identifiants des attributs qui composent la table apparaissent comme les arguments de l'attribut `attrs` de la première ligne de définition d'une table (figure 18 - ligne 1) pour cette même raison. Chaque ligne possède également un identifiant (attribut `id` figure 18 - lignes 3 et 6). Les entités lexicales, instances des types, sont codées en XML sous trois formats possibles selon qu'elles sont composées de plusieurs mots ou d'un seul et selon que l'utilisateur en a précisé les formes graphiques possibles. Par souci de clarté, nous les avons omises dans l'exemple donné en figure 18, nous les détaillerons dans les parties suivantes.

Du point de vue de l'implantation, la relation d'agrégation entre un ou plusieurs attributs et une ligne de table a été abandonnée au profit d'une relation plus directe et plus facilement manipulable entre les valeurs d'attributs et les lignes selon le schéma de classes suivant (figure 19). Le concept d'une table LUCIA ne s'en trouve pas modifié pour autant puisqu'une valeur donnée ne peut correspondre qu'à un seul attribut. Le mécanisme des identifiants et de la réification des instances de ces classes à partir d'XML est identique à celui des attributs. Les classes sont rassemblées dans un *package* `Lucia` de niveau supérieur au *package* `Attribut` puisque les tables ne semblent pas pouvoir faire l'objet d'instanciations en dehors de l'utilisation du système LUCIA.

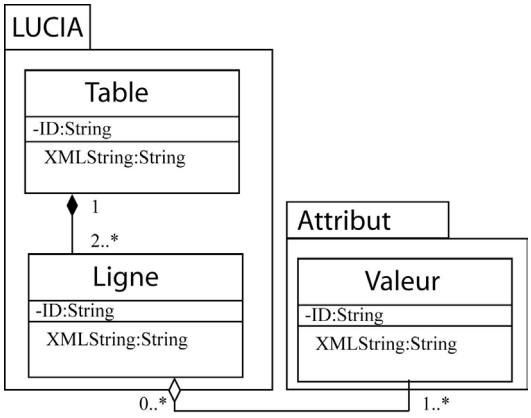

Figure 19 - Diagramme de classe pour les tables.

Dans la partie suivante, nous présentons les dispositifs qui sont des ensembles de tables d'un même projet de catégorisation et de description lexicale.

3.2.3 Les dispositifs

3.2.3.1 Présentation

À une ligne d'une table LUCIA, on peut associer une ou plusieurs entités lexicales instances de cette catégorie. Ce processus de catégorisation est récursif : chaque catégorie représentée par une ligne de table peut être elle-même sous-catégorisée comme sur le modèle de la figure 20[60]. Dans cet exemple, on voit que la catégorie c4 a été sous-catégorisée par la catégorie C2. Les valeurs d'attributs A1v2 et A2v2 sont alors associées à toutes les instances des catégories de C2 (c'1, c'2, c'3 et c'4). La validité d'une sous-catégorisation est dépendante des possibilités admises par la pratique langagière envisagée puisque le lien de sous-catégorisation préfigure également un héritage d'attributs/valeurs symbolisé par une ligne orientée d'une ligne vers une table. L'héritage concerne toutes les lignes de la table d'arrivée.

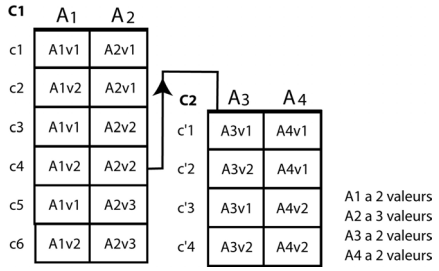

Figure 20 - Deux grilles de catégorisation reliées par un lien de sous-catégorisation.

Du fait de l'héritage, si la catégorie C2 est considérée comme un taxème, alors la catégorie c4 pourra être considérée comme le classème des instances de la catégorie C2 puisque les valeurs A1v2 et A2v2 pourront être considérées alors au même titre que des sèmes (plus) génériques pour les instances des catégories c'1, c'2, c'3 et c'4 de C2. C'est la tâche pour laquelle la catégorisation est produite qui permet de déterminer quand s'arrête le processus de sous-catégorisation et quels sont les liens de sous-catégorisations admissibles. La sous-catégorisation peut, par exemple, avoir lieu après une première utilisation de tables donnant lieu à des résultats partiellement satisfaisants. La combinatoire des sèmes – problématique dans l'approche componentielle informatisée – est ici résolue en partie du fait des structures manipulées et du processus de catégorisation proposée . À une ligne donnée représentant une catégorie (instanciée par des entités lexicales ou non), on ne peut associer que les attributs/valeurs pris en considération dans des tables de catégorisation.

[60] Dans le modèle, la sous-catégorie d'une ligne de table est une table. Une même ligne de table peut être sous-catégorisée en plusieurs tables. Le terme est donc employée différemment que dans la plupart des modèles de catégorisation (par exemple les projets de catégorisations thématiques comme l'Open Directory Projet - http://www.aef-dmoz.org/) où les sous-catégories sont des catégories « contenues dans une autre ».

Un dispositif LUCIA est constitué de tables et de leurs relations de sous-catégorisation. Certaines de ces tables peuvent être reliées par un lien de sous-catégorisation, elles forment alors un arbre de grilles de catégorisation où chaque niveau est une sous-catégorisation d'une ligne d'une table de niveau supérieur. Certaines tables d'un dispositif peuvent ne présenter aucun lien de sous-catégorisation (qu'elles en soient l'aboutissement ou le point départ) : les catégories d'un même dispositif ne partagent pas nécessairement des attributs communs. Un dispositif ne correspond donc pas nécessairement à un arbre, au sens de la théorie des graphes, mais à une forêt (figure 21). Nos premières communications sur le modèle ont pu faire état de la possibilité de produire des cycles au sein des dispositifs, par exemple dans [Nicolle *et al.*, 2002*]. Cette éventualité a été abandonnée puisqu'elle n'était pas au final viable d'un point de vue computationnel. La partie 4.4.2 du chapitre 4 nous permettra de voir pourquoi une telle configuration apporterait trop d'indécidabilité lors des phases d'analyses automatiques et serait d'autre part beaucoup plus compliquée à concevoir pour l'utilisateur. Les grilles de même niveau dans un arbre de catégorisation n'ont pas *a priori* de propriétés communes particulières : les attributs utilisés à ce niveau peuvent être différents. La présence d'une table à l'intérieur d'un dispositif traduit le fait que les instances des catégories correspondantes sont en rapport avec le domaine d'intérêt décrit dans le dispositif et devront être exploitées en tant que telles au cours des analyses. Nous verrons lors la présentation de la conceptualisation computationnelle des dispositifs comment cela est rendu opératoire dans les logiciels.

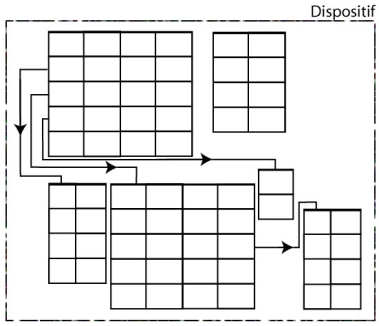

Figure 21 – Forêt de grilles et grilles de catégorisation : un dispositif.

Les dispositifs sont des ensembles de tables en rapport avec un même domaine de tâche défini en partie par un corpus d'observation utilisable pour soutenir le processus de catégorisation lexicale. La pertinence et la justesse des descriptions obtenues avec le modèle LUCIA sont à la discrétion de l'utilisateur. L'évaluation de leur cohérence du point de vue des contraintes du modèle est assistée par les logiciels qui repèrent d'éventuelles inadéquations et suggèrent certaines modifications (c.f. partie 4.4 – chapitre 4). Dans l'exemple de dispositif proposé en figure 22, on peut questionner la description proposée sous divers aspects. De telles mises en question font partie du modèle centré sur l'utilisateur ; son utilisation s'inscrit dans un cycle itératif qui peut intégrer un processus de révision des ressources. Les ressources qui exemplifient l'ensemble de ce tapuscrit suivent cette même démar-

che : elles sont susceptibles d'être révisées après utilisation et ne sont que des instantanés sortis d'un cycle d'expérimentations.

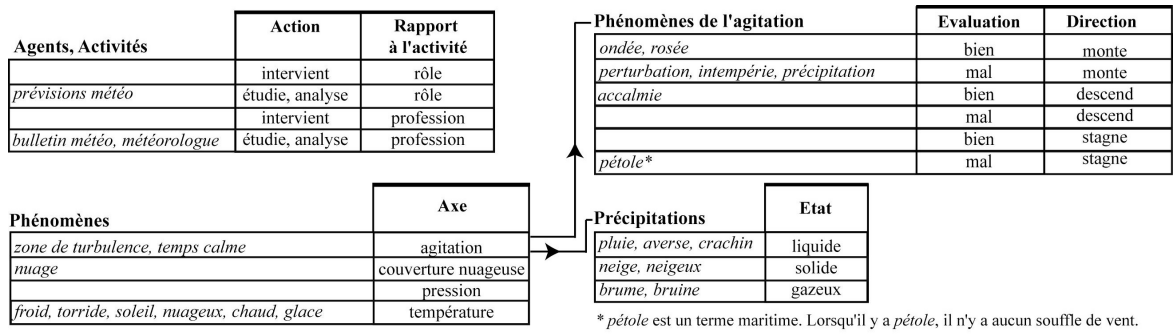

Figure 22 – Un dispositif en rapport avec la météorologie

Le dispositif présenté en figure 22 est en rapport avec le domaine de la météorologie. La configuration proposée correspond à l'une des étapes d'élaboration d'un dispositif utilisé dans le cadre d'une étude sur une métaphore conceptuelle. Nous présenterons en détails cette étude dans les prochains chapitres. Nous sommes ici en présence de quatre tables dont deux correspondent à des sous-catégorisation de la table « Phénomènes ». Cette dernière permet de catégoriser des entités lexicales ayant trait à des phénomènes météorologiques que l'on a choisi de réunir à l'aide de l'attribut [Axe] et de distinguer à l'aide des valeurs [agitation *vs.* couverture nuageuse *vs.* pression *vs.* température]. Le domaine d'interprétation « Axe » permet de préciser que les valeurs *agitation* , *couverture nuageuse*, *pression* et *température* sont envisagées relativement à une représentation spatiale permettant, comme c'est le cas dans la table « Phénomènes de l'agitation » de préciser une dynamique des phénomènes. Les liens d'héritages proposés permettent par exemple d'associer à *ondée* et *rosée* instances de la catégorie formée par les valeurs [Évaluation : bien] et [Direction : monte] de la table « Phénomènes de l'agitation » à la valeur [Axe : agitation]. Cette description peut être paraphrasée comme suit : *ondée* est susceptible d'avoir une évaluation positive et de correspondre à un phénomène qui monte sur l'axe météorologique des agitations – c'est en tout cas ainsi qu'il a été repéré dans le corpus d'observation utilisé pour la construction de ce dispositif. Le nom d'un dispositif circonscrit le domaine qu'il décrit. Ainsi, l'exemple proposé en figure 22 peut être nommé « Météorologie » ou « La météo ». Nous reviendrons sur les modalités de cette dénomination dans le prochain chapitre à propos de la construction de dispositifs pour des tâches documentaires définies.

Un corrélat important de cette organisation en tables au sein d'un dispositif, est que toute entité lexicale n'est pas interdéfinie avec toute autre, mais que les différences et les points communs ne sont pertinents que dans la délimitation opérée par les tables et le dispositif. Il s'agit à la fois d'une économie descriptive pour réduire la quantité de ressources nécessaires à la tâche mais également un moyen de prévenir une dérive purement sémasiologique soulignée dans [Pincemin, 1999b*] qui amè-

nerait à définir des entités les unes par rapport aux autres alors qu'elles n'apparaissent pas dans les mêmes contextes.

Dans les chapitres suivants, nous donnerons plusieurs exemples de dispositifs ainsi qu'un protocole de construction en fonction d'une tâche documentaire donnée. Dans la partie suivante, nous exposons les détails de conception et d'implantation des dispositifs.

3.2.3.2 Conception et implantation

Dans le modèle, une table LUCIA appartient obligatoirement à un dispositif et inversement, un dispositif est au minimum constitué d'une table. La conception de l'implantation d'un dispositif correspond donc au diagramme suivant (figure 23) : un dispositif correspond à un ensemble de tables et à un nom.

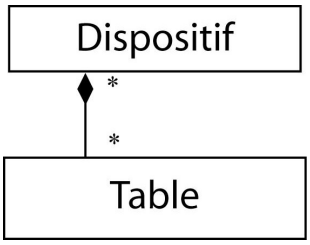

Figure 23 – Conception des dispositifs LUCIA.

La représentation XML d'un dispositif fonctionne sur les mêmes principes que pour les autres objets du modèle. Les dispositifs sont repérés par des identifiants (`id`) et englobent les tables (`table`) qui leurs sont associées (figure 24). Les identifiants servent ici aussi à prévoir les degrés de détails utiles à l'utilisateur dans sa tâche – nous avons à ce stade trois degrés de détails pour une entité repérée dans un texte : ligne, table et dispositif. En revanche, les tables sont présentées entièrement dans les fichiers XML des dispositifs car même si une même combinaison d'attribut peut apparaître dans plusieurs dispositifs, la notion de domaine de tâche associée à un dispositif implique que les entités lexicales des tables correspondantes ont peu de chance d'être identiques d'un dispositif à l'autre. Les attributs ne sont pas reproduits dans le code d'un dispositif. Ils sont stockés dans des fichiers à part appelés dictionnaires d'attributs (`dictarr`) (figure 25 – les attributs sont empruntés à la figure 22). et simplement présents par l'intermédiaire de leurs identifiants puisque certains peuvent être partagés entre plusieurs tables et entre plusieurs dispositifs. Le caractère partageable de certains attributs sera exploité au cours des analyses comme nous le verrons dans le chapitre 5.

```
 1. <?xml version="1.0" encoding="iso-8859-1" ?>
 2. <disp id="disp_La_météo">
 3. <dispnom>La meteo</dispnom>
 4.  <table id="disp_La_météo_att3-4" attrs="attr3 attr4">
 5.  <tablenom>Agents, Activités</tablenom>
 6.  <ligne id="disp_La_ météo_tab3-4ligne0" vals=" attr3val0 attr4val0">
 7.      <!— ici sont placées les entités lexicales du premier type -->
 8.  </ligne>
…
xx.  </table>
 XXX.</disp>
```

Figure 24 - Extrait de la représentation XML d'un dispositif.

```
1. <?xml version="1.0" encoding="iso-8859-1" ?>
2. <dictattr>
3.     <attr id="attr1">
4.           <attrnom>Rapport à l'activité</attrnom>
5.           <val id="attr1val0">rôle</val>
6.           <val id="attr1val1">profession</val>
7.     </attr>
…
XX.</dictattr>
```

Figure 25 - Extrait de la représentation XML d'un dictionnaire d'attributs.

La réification du code XML représentant un dispositif donne lieu à la création d'une instance de la classe dispositif selon le diagramme de classes suivant (figure 26). La classe `dispositif` et celle qui correspond à son nom sont présentes au sein du même *package* pour les mêmes raisons que celles exposées pour les autres éléments du modèle.

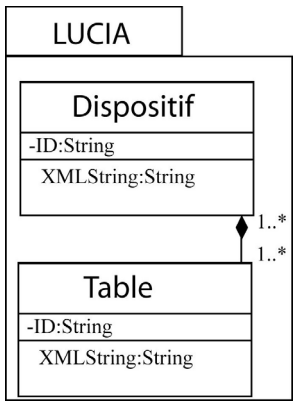

Figure 26 - Diagramme de classe pour les dispositifs.

La réification des informations contenues dans les fichiers XML est assurée par la classe `Parser`, sous-classe du `DefaultHandler` du `SAXParser`, appartenant au *package* `LUCIA` englobant lui-

même le sous-*package* `Attribut`. Dans ce même *package*, une classe `Session` permet de regrouper plusieurs dispositifs communs à une même tâche. La session fait également référence aux dictionnaires d'attributs utilisés dans les dispositifs en question puisque c'est relativement à une même tâche que l'on pourra décider si des attributs sont partageables ou non entre plusieurs dispositifs. Le diagramme suivant (figure 27) donne une vue d'ensemble des classes utilisées pour la création ou l'utilisation des éléments du modèle de catégorisation et de représentation lexicale qui sont toutes rassemblées au sein du *package* `LUCIA`.

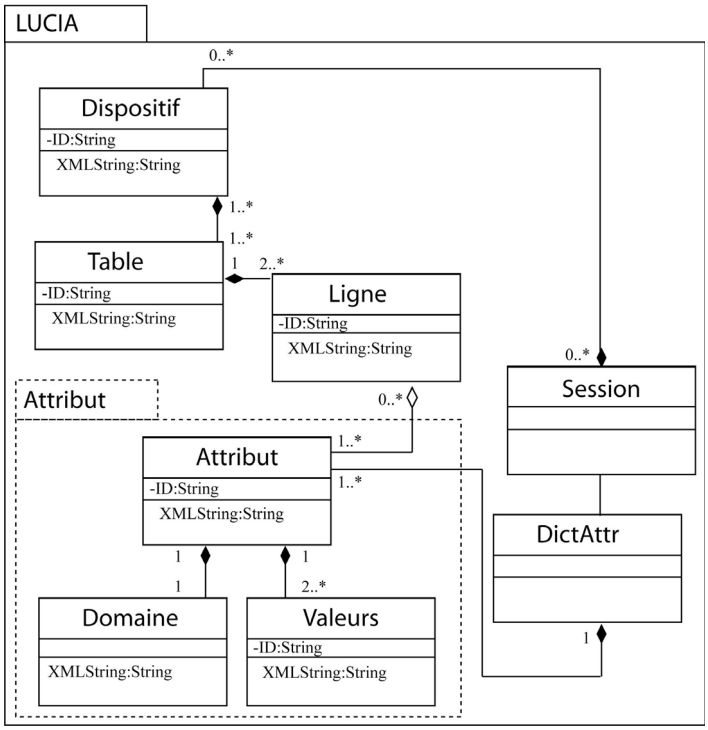

Figure 27 - *Package* LUCIA.

Dans les parties précédentes, nous avons présenté le modèle conceptuel et informatique LUCIA en tant que modèle de catégorisation et de représentation lexicale. Dans la partie suivante, nous allons voir en quoi nos choix de conceptualisation font de LUCIA un modèle de représentation des mots dans le discours, permettant l'expression d'un point de vue de l'auteur des descriptions et même certaines dimensions relevant du jugement.

3.3 LUCIA, un modèle de représentation des mots dans le discours

Depuis le début de ce tapuscrit, nous avons affirmé que les structures LUCIA sont l'expression d'un point de vue de leur auteur vis-à-vis de sa tâche et sur le corpus d'observation qu'il a

éventuellement utilisé pour appuyer ses descriptions. Dans cette partie, nous nous attachons à éclairer la notion de point de vue à la lumière de nos propositions. Nous verrons également que le modèle de catégorisation et de description lexicale LUCIA permet l'expression d'un jugement qui sera exploité par les logiciels au même titre que les autres informations relevant de l'observation du matériau linguistique ou des connaissances de l'utilisateur.

3.3.1 Point de vue et jugement de l'utilisateur

La notion de point de vue apparaît dans de nombreux champs de la linguistique. Chez Genette par exemple [Genette, 1972 : 212-213], il s'agit du phénomène de changement de focalisation dans un texte narratif. Ce changement est interprété comme une « variation de point de vue » de la part du narrateur. Ducrot introduit le point de vue dans la théorie énonciative polyphonique [Ducrot, 1984] et aborde ainsi l'*énonciateur* qui désigne le personnage dont le récit exprime les opinions, sans que ce personnage soit nécessairement confondu avec le narrateur [Moeschler et Reboul, 1994 : 426]. Comme il est souligné dans [Norén, 2000], cette vision du point de vue relève exclusivement de l'analyse littéraire. Eco [Eco, 1992] aborde le point de vue en fonction de l'interprétation. Dans son roman *Le Nom de la Rose*[61], l'auteur écrit : « *Un roman est une machine à générer de l'interprétation. (...) Et je définirais l'effet poétique comme la capacité, exhibée par un texte, de générer des lectures toujours différentes, sans que jamais on en épuise les possibilités.* ». Pour lui, une interprétation qui fait sens doit produire quelque chose de nouveau, et cependant, veiller à ne jamais, en quelques sortes, proposer rien qui soit indifférent à l'interprété, faute de quoi elle devient *surinterprétation*. Pour le système LUCIA, la liberté est donnée à l'utilisateur de placer dans les structures ce qui pourrait relevait selon Eco d'une *surinterprétation* dans le sens où il pourra fournir au systèmes des informations soutenues par un corpus d'observation, mais également plus personnelles en rapport avec sa propre expérience de la tâche en cours et du domaine d'intérêt de cette tâche hors d'un support textuel. Nous nous attachons en effet à proposer un système utilisable dans le cadre d'une tâche donnée. Pour cette tâche, l'analyse d'un corpus d'observation n'est pas nécessairement suffisante pour rendre compte de tous les aspects intéressants pour l'usager. Dans d'autres tâches que celles que nous proposons, l'interprétation linguistique est primordiale. C'est le cas de la traduction par exemple. En revanche, pour une tâche de veille documentaire par exemple, c'est l'interprétation de l'usager qui prévaut.

L'activité d'interprétation, c'est-à-dire la construction d'un sens par un sujet interprétant d'un texte est un processus influencé par les conditions d'utilisation de ce texte. Nous avons vu que l'intérêt suscité par un même texte dans une tâche documentaire varie en fonction des conditions qui président à sa lecture ou à sa recherche. Les conditions de l'interprétation ne relèvent donc pas toutes du linguistique. On considère parfois les connaissances « encyclopédiques » ou culturelles au même titre que les

[61] U. Eco. *Le nom de la Rose*, Paris : LGF, 2002. traduit de l'italien par Jean-Noël Schifano.

compétences idéologiques ou les déterminations psychosociales de l'interprétant en les regroupant dans un cadre non-linguistique [Kerbart-Orecchioni, 1990]. Il est donc impossible de trouver *le* sens d'un texte, tout au plus, pourrait-on parler *des* sens d'un texte puisque les compétences susmentionnées peuvent varier d'une lecteur à un autre[62]. Comme il a été souligné dans [Brouillette, 1995], les travaux de Rastier ont fait suite à une certaine polarisation des théories du langage voulant que l'on affirme soit l'objectivité sans nuance du sens, soit sa subjectivité absolue. La SI propose de considérer le sens comme étant dépendant du texte, mais aussi d'une situation de communication comprenant un émetteur, un récepteur et un ensemble de conditions (genre textuel, pratique sociale, etc.) : elle s'attache à prendre en considération l'ordre herméneutique et considère donc le sens non pas comme donné mais à construire. La SI envisage cependant une possibilité de consensus minimal entre récepteurs dans une situation donnée à partir des thèmes génériques correspondants aux taxèmes ou niveaux supérieurs que son le domaine et la dimension. Cependant, ce qui relève du spécifique, i.e. ce qui n'est lié à aucune classe sémantique déterminée, doit être repéré indépendamment d'une lexicalisation précise par la découverte des isotopies. L'auteur illustre ces propos avec un regroupement de sèmes tirés de *l'Assommoir* de Zola (1840-1902) : 'chaud'', 'visqueux' , 'jaune' et 'néfaste' se répètent tout au long du texte, sans aucune constance lexicale, par des mots comme *jus*, *pipi*, *sauce*, *morve*, *beurre*, *bedon*, *cuivre*, *huile*, *lune*, *goutte*, etc. [Rastier, 1989 : 58]. Il démontre ainsi que le sens n'est pas donné mais à construire : c'est l'acte d'interprétation. Nous voyons la poindre une différence fondamentale entre nos attributs et les sèmes de la SI.

Nous l'avons vu à travers la figure 8 p.70, l'association des attributs aux entités lexicales est le fait de l'utilisateur du système. Ainsi, l'expression de son point de vue sur la tâche et le domaine lexical de cette tâche s'effectue au cours de toutes les étapes de la construction des ressources :

- *À travers le choix des entités lexicales* : même si, comme nous le verrons en détail dans le prochain chapitre, ce choix est assisté à travers des principes simples de calcul d'occurrences et de cooccurrences de termes, le dernier mot est donné à l'utilisateur. Les entités retenues (et les entités délaissées) représentent donc l'intérêt particulier que peut avoir l'usager par rapport au domaine de sa tâche. Pour reprendre l'exemple de Rastier sur *l'Assommoir*, il est possible qu'un lecteur ne soit, dans le cadre d'une étude linguistique, intéressé que par la présence des sèmes 'chaud' et 'visqueux' et ne retienne donc pas tous les termes précités. Dans un cadre de veille documentaire, selon que l'on est intéressé par tel ou tel aspect d'un sujet précis (par exemple la biocorrosion des matériaux dans son aspect bactérien ou mycologique), on ne retiendra pas les mêmes termes à décrire et donc les mêmes attributs à partir de la lecture d'un même texte. C'est bien l'expression d'un besoin particu-

[62] Si l'on parle *des* sens d'un texte, cela peut sous-entendre que l'on peut les énumérer. Dans nos propos, il faut comprendre qu'il serait possible d'en expliquer plusieurs (au moins deux par exemple).

lier, l'expression d'un point de vue sur une tâche qu'il est possible ici de partager avec les logiciels.

- *À travers le choix des attributs à associer aux entités lexicales et les regroupements en tables au sein d'un dispositif* : en ne distinguant pas explicitement les attributs qui relèveraient de sèmes spécifiques ou génériques, l'association d'attributs avec une entité lexicale au sein d'une table LUCIA permet la focalisation sur un aspect précis, et sur des éléments de signification particuliers que peuvent supporter ces entités en contexte. Une lecture « médicale » de *l'Assommoir* amènerait probablement à porter un intérêt particulier aux lexies *pipi*, *morve*, *bedon* et *goutte* et donc d'envisager pour la structuration des attributs en rapport avec leur capacité à avoir trait à des manifestations ou des caractéristiques physiques d'un individu. Dans un cadre de veille documentaire sur la biocorrosion, on pourrait envisager de regrouper les termes *micro-algue*, *moisissure*, *champignon* et *bactérie* à l'aide des attributs [Type de cellule : eucaryote *vs.* procaryote] et [Fonctionnement : photosynthèse *vs.* pas de photosynthèse]. Le choix de ces attributs permet de mettre l'accent sur deux propriétés biologiques particulières des éléments auxquels peuvent référer ces termes dans un tel contexte. En revanche, il est possible de n'envisager l'utilisation que d'un seul de ces deux attributs. L'utilisation du seul attribut [Type de cellule : eucaryote *vs.* procaryote] amènerait ainsi à ne pas distinguer dans la catégorisation les termes *micro-algue, moisissure* et *champignon* sur le modèle de la figure suivante (figure 28).

Micro-organismes facteurs de risque	Type de cellule	Fonctionnement
micro-algue	eucaryote	photosynthèse
	procaryote	photosynthèse
moisissure, champignon, mycologique	eucaryote	pas de photosynthèse
bactérie, bactérien, bactériologique	procaryote	pas de photosynthèse

Micro-organismes facteurs de risque	Type de cellule
micro-algue, moisissure, champignon	eucaryote
bactérie	procaryote

Figure 28 – Deux tables LUCIA en rapport avec les micro-organismes facteurs de risque de la biocorrosion.

- *À travers l'organisation des tables au sein d'un dispositif* : les deux étapes abordées précédemment impliquent des configurations particulières au sein des dispositifs. Ces configurations permettent d'exprimer un intérêt particulier pour des aspects précis des textes à analyser. Pour reprendre l'exemple de la biocorrosion, on peut, dans ce domaine, être intéressé par les propriétés biologiques des facteurs du risque relevant du vivant ou préférer quantifier l'importance de ce risque pour préciser à un niveau inférieur quels éléments correspondent aux niveaux de risque envisagés : ces deux points de vue sur le domaine à décrire dépendent de ce qui est intéressant de voir apparaître dans les textes à analyser pour la tâche en cours (figure 29) ; nous verrons en outre l'importance de telles configurations lors des analyses automatiques. L'utilisation de l'attribut [Niveau de risque : peu dangereux

vs. dangereux *vs.* très dangereux] relève d'un jugement porté par l'auteur du dispositif sur les éléments qui le composent. Ce critère de classification relève bien d'un point de vue directement en rapport avec la tâche puisque pour tout à chacun l'eau de mer et les micro-algues ne représentent pas nécessairement de danger particulier. Nous aurons en outre l'occasion dans le chapitre 5 de revenir sur cet exemple (c.f. 5.4).

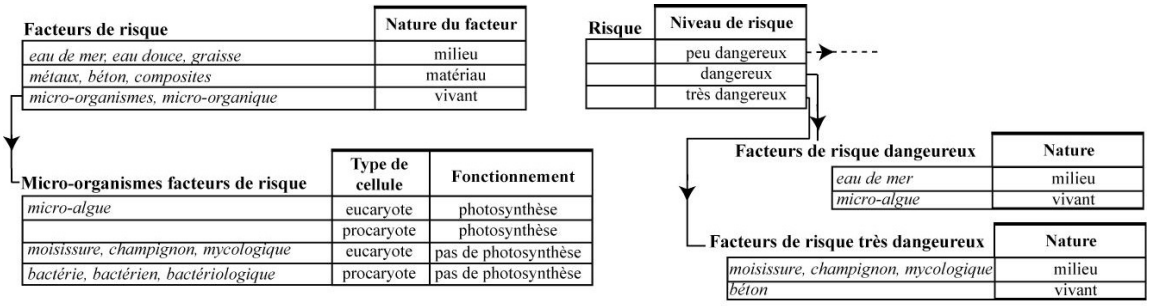

Figure 29 - Deux dispositifs LUCIA en rapport avec les micro-organismes facteurs de risque de la biocorrosion.

3.3.2 Les mots dans le discours interprété

Dans le chapitre 2, l'exemple du caviar nous a permis de montrer l'importance des objectifs d'une tâche quant aux traits à retenir pour un terme donné. Catégoriser certaines entités lexicales en fonction de leur potentialité à faire référence à des aliments *granuleux* et *salés* s'avère tout aussi utile qu'une distinction entre aliments en fonction des *valeurs socio-économiques* qu'ils évoquent. Le modèle LUCIA est souple et permet de s'adapter aux exigences des tâches pour lequel il est utilisé au niveau des propriétés et distinctions mises en lumière au sein des structures manipulées. Les deux exemples proposés pourraient être exprimés à l'aide de tables LUCIA comme dans la figure 30 : il ne s'agit pas ici d'associer des informations définitoires au concept de *caviar* mais de préciser quels éléments potentiels de signification sont intéressants pour le projet de catégorisation.

T1	Consistance	Goût
caviar	granuleux	salé
crèpe	compact	sucré
gâteau de semoule	granuleux	sucré
omelette	compact	salé

T2	Statut socio-économique
caviar, vison, diamants	luxueux
tracteur, ampoule	courant
rutabaga, tiercé	vulgaire

Figure 30 – Deux tables LUCIA **contenant *caviar*.**

La table T1 présente une catégorie avec quatre lignes différentes. La catégorie est caractérisée par la mise en exergue des propriétés physiques de consistance et de goût. Les valeurs des deux attributs binaires [Consistance : granuleux *vs.* compact] et [Goût : salé *vs.* sucré] distinguent quatre lignes à l'intérieur de la catégorie. À chacune de ces lignes, nous avons fait correspondre des entités lexicales susceptibles de correspondre aux valeurs d'attributs auxquelles elles ont été associées en tant

qu'élément de signification (dans les exemples choisis, c'est le monde culinaire qui est la référence commune). Ce type de table peut être utilisé si d'autres entités lexicales présentant l'une ou l'autre des propriétés exprimées par ces attributs peuvent se retrouver dans le dispositif de la tâche. Dans ce cas, la consistance et le goût doivent présenter un intérêt quant à la façon de parler ou la façon dont on désire voir parler de la tâche en cours[63]. La table T2 présente une catégorie avec deux lignes différentes. Une seule propriété a été choisie pour la catégorie à travers l'utilisation de l'attribut [Statut économique : luxueux *vs.* vulgaire]. Cette table pourrait être utilisée pour une tâche dans laquelle la différenciation d'appréciation de différents aliments ou objets s'effectue en fonction de leur coût ou de leur évaluation socio-économique. Elle pourrait être issue par exemple, de la lecture d'un texte de fiction. Nous citerons un extrait des *Bienheureux de la désolation* d'Hervé Bazin (1911-1996) :

Quand l'ampoule succède à la lampe à huile, le tracteur au bœuf, il s'agit d'un nouveau nécessaire, qui surclasse l'ancien, hors d'époque. Mais le vison, le diamant, le caviar seront toujours superflus.

Bazin, H., *Les bienheureux de la désolation*, Col., Points Roman, Seuil, Paris.

À partir des propos du président G.W. Bush que nous avons présentés dans le chapitre 1 (1.1 p.12), nous pouvons envisager plusieurs tables regroupant l'entité lexicale *axe maléfique* (ou *axe du mal* considérée ici comme équivalente[64]) et l'entité *axe du bien* qui lui est souvent opposée ; pour ce faire nous étudierons des textes qui reprennent ces entités. Ces tables traduisent explicitement ou non, à travers le choix des attributs et de leurs valeurs, différents points de vue sur les entités en question et laissent donc la place à différentes autres instances pour les types des catégories proposées (c'est là un des atouts de la structuration en grille). Nous nous appuierons ici sur l'étude de textes en français (nous ne faisons aucune hypothèse a priori sur la validité des transpositions culturelles entre la langue initiale – l'anglais étasunien – et celle des textes choisis). Nous nous sommes limités à l'étude de trois textes ; nous n'en reproduisons ici que des extraits (1), (2) et (3). Ces textes sont considérés en tant que partie d'un intertexte contenant le discours initial bien qu'ils expriment parfois une opinion négative sur son contenu. Les tables qui résultent de notre lecture de ces textes sont proposées en figure 31.

*Une lecture cartographique du monde démontre que l'*axe du mal *qui se déplace en fonction des cours*

[63] Ce n'est vraisemblablement pas le cas dans le quotidien des collégiens de l'enquête précédemment citée, mais cela pourrait l'être par exemple pour un physicien de la cuisine comme Hervé This : This, H., 1995, *Révélations Gastronomiques*, Col., Sciences et Gastronomie, Belin, Paris.

[64] Les deux entités sont ici considérées comme équivalentes bien qu'une analyse poussée pourrait éventuellement dégager quelques différences d'emploi. « *axe maléfique* » et « *axe du mal* » sont par exemple présentes toutes les deux (parfois dans le cadre d'une reprise anaphorique) et non distinguées dans la traduction du discours de G.W. Bush proposé par l'ambassade des Etats-Unis en Belgique (http://www.usembassy.be/fr/frpolicy/fr. bush.012902.htm) ou dans de nombreux articles de presse (« *L'axe du Mal* » d'Ignacio Ramonet – Le Monde Diplomatique de Mars 2002, « *L'Europe doit jouer un rôle important au Proche-Orient.* » de Jean-Michel FLOC'HLAY – Fenêtre sur l'Europe du 21 Février 2002 …). Notons tout de même que ces derniers articles jouent plus ou moins avec l'opposition maléfique *vs.* bénéfique ce qui constitue un sous-entendu que l'on acceptera ou que l'on regrettera pour l'interprétation des textes en question.

mondiaux dont parle Bush (…), n'est composé que de pays aux matières premières stratégiques vitales pour le fonctionnement des firmes et industries militaires (…).

(1) Boubé Bali, S., *Le Niger sur l'axe du mal de Bush***,** Les Chroniques de Madame Chang, 18/01/2004 - http://www.leschroniques-demadamechang.net/ textes/ textes_monde/textes_afrique/afrique_niger_bush_saley.htm

L'axe du Bien se conduit mal. Les États-Unis, l'Angleterre et Israël tiennent tellement à en découdre avec Saddam Hussein qu'ils mentent et trichent à qui mieux mieux et qu'ils en oublient toute décence.

(2) Laplante, L., *Un bellicisme indécent et dévastateur,* Dixit Laurent Laplante, Editions Cybérie, Québec, 12/12/2002 http://www.cyberie.qc.ca/dixit/20021212.html

Le pari américain d'une contagion démocratique au Proche Orient est en passe d'être tenu. Déjà Khadaffi se rallie au nouvel axe du bien*, et l'on sent bien la volonté israélo-palestinienne d'en finir avec le cycle de la violence et de suivre enfin la feuille de route que leur a tracé George Bush.*

(3) Climacus (pseudonyme), *Le nouvel axe du bien*, Forum « Le Monde », 20/12/2003 http://forums.lemonde.fr/perl/showthreaded.pl?Cat=&Board=guerre& Number=1132873&page=20&view=collapsed&sb=5&part=

	Evaluation
axe maléfique, axe du mal	mal
axe du bien	bien

	Régime
axe du mal	dictature
Etats-Unis, Angleterre, Israël, axe du bien	démocratie

	Matières Premières
axe du mal	statégiques
	communes

Figure 31 – 3 tables LUCIA **contenant *axe du mal*.**

Nous constatons d'emblée que ces trois tables proposent des distinctions et des regroupements qui ne pourraient pas être proposés *a priori* hors des cotextes analysés. Ce serait encore pire si on se référait aux connaissances encyclopédiques sur le monde. C'est par exemple le cas pour la table utilisant l'attribut [Régime] à l'intérieur de laquelle *Angleterre* et *axe du mal* apparaissent. Cette table traduit une prise de position conjoncturelle de l'auteur de la table (en fait ici, de notre interprétation du texte en question). L'attribut utilisé est référentiel et traduit une vision politique actualisée par les entités utilisées dans le processus de catégorisation (dictature *vs.* démocratie). Dans cette table, il n'est pas fait de distinction entre *axe du bien* et *États-Unis*, *Angleterre* et *Israël*. Ces entités n'ont pas été dissociées du point de vue de l'élément de signification choisi alors même que l'on pourrait les distinguer du simple fait de leur référence au monde : les Etats-Unis et Israël sont des nations et l'Angleterre n'est depuis 1801 qu'une partie du Royaume-Uni de Grande-Bretagne et d'Irlande du Nord. Nous pouvons noter également que les pays cités comme faisant partie de *l'axe du bien* le sont en fonction d'une interprétation particulière des propos initiaux : l'auteur du texte semble sous-entendre l'existence d'un tel groupe opposé à *l'axe du mal* et a su nommer, en fonction de ses connaissances, des pays pour s'y référer. La troisième table, utilisant l'attribut [Matières premières : stratégiques *vs.* communes], permet d'éclairer nos propos sur l'absence de *définition* des entités lexicales dans les tables. L'attribut utilisé permet de placer *axe du mal* dans une catégorie mais la valeur actualisée par l'entité ne saurait faire office de définition.

Dans cette partie, nous avons vu que les attributs mis en jeu dans les dispositifs LUCIA, permettent la mise en place de catégories pour exprimer un point de vue sur des entités lexicales observées et/ou utilisées dans des pratiques réelles.

3.3.3 Approche complémentaire entre onomasiologie et sémasiologie

Il peut paraître antinomique de défendre d'un côté la variabilité contextuelle des significations et de vouloir de l'autre proposer des descriptions de signifiés à partir de textes pour les exploiter pour l'analyse d'autres textes. Nous pouvons y voir une interférence entre les approches onomasiologiques et les approches sémasiologiques. Dans cette partie, nous expliquons en quoi finalement, notre approche est articulée autour des deux directions.

Dans [Nyckees, 1998], l'auteur oppose la recherche des clefs des significations *dans les relations qu'elles entretiennent entre elles* (approche différentielle : celle de la SI) et *dans les relations que chacune d'entre elles, considérée isolément, entretient avec le monde* (approche référentielle : celle des approches logico-grammaticales). Ces propos éclairent les différences entre l'approche sémasiologique et l'approche onomasiologique. Dans le premier cas, on tente d'associer à un signifiant donné, l'ensemble des signifiés qu'il est susceptible de véhiculer. Dans le second cas, on part de classes de signifiés pour structurer les signifiants. Pour les analyses de contenu de textes que nous proposons, nous adoptons une démarche médiane entre le sémasiologique et l'onomasiologique dont Bourion a, par exemple, souligné l'étroite complémentarité [Bourion, 2001]. Ceci nous permet, entre autres, d'être adapté à l'utilisateur en tant que producteur des représentations utiles aux analyses puisque l'approche sémasiologique reste la plus répandu et semble plus facile d'accès (voir 3.4). L'approche sémasiologique est conçue en tant que démarche partant des signes (entendus comme entités lexicales extraites d'un corpus d'observation ou proposées en fonction des connaissances convoquées pour la tâche) pour construire les représentations utilisées. C'est dans le contexte d'une pratique que l'utilisateur peut parvenir à construire des dispositifs LUCIA. L'approche sémasiologique souffre souvent de ne pas prendre en considération des contextes de rencontre des diverses significations d'un signifiant donné. Rastier [Rastier *et al.*, 1994] (voir également [Beust, 1998* : 78]) considère alors que la langue n'est vue que comme une simple nomenclature. Notre approche sémasiologique, qui tend à isoler le signe dans une représentation informatique et d'en fixer certains critères de signification observés, n'implique pas une vision uniquement référentielle de la langue. Si le point de départ de la structuration que nous proposons est bien le signe, celui-ci peut être considéré en contexte et dans le cotexte de la tâche de l'utilisateur. Il ne s'agit pas de décrire des éléments de contenu du signe de manière absolue mais de le faire en fonction d'une situation particulière. Notre approche se situe dans le sémasiologique car le signe est isolé (dans le traitement informatique et par l'utilisateur) lors de la phase de description et de catégorisation et lors des premières phases d'appariement des analyses.

Mais elle se situe aussi dans l'onomasiologique parce que les signifiés sont décrits au sein de classes en fonction de l'interprétation d'un contexte et que les analyses prennent avant tout place dans des textes. Le travail des logiciels sera justement d'évaluer la pertinence d'une telle description dans un autre cotexte que celui qui a permis de la créer. Enfin, le principe d'une *sémantique légère* que nous défendons trouvera également un intérêt à l'approche onomasiologique puisque nous verrons dans les chapitres suivants que l'organisation différentielle des signifiés nous permet de limiter les ressources du système à ce qui est utile pour la tâche en cours.

3.4 Expérience

L'évaluation de modèles centrés sur l'utilisateur est délicate. Comme nous avons pu déjà le préciser, nos travaux s'inscrivent dans un courant du TAL influencé par une branche des sciences cognitives où l'on préfère la coopération système/utilisateur à l'automatisation – on parle alors de cognition située et distribuée. En ce qui concerne nos propositions, la machine (les logiciels) n'est pas l'organe central du modèle, elle tient un rôle en rapport avec ses capacités premières : le calcul, la manipulation rapide des grandes quantités de données, l'affichage de données (diagrammes, schémas…) et l'interaction. Elle n'est pas une entité omnisciente, elle est source d'assistance et de suggestion ; c'est un compagnon personnel pour l'aide à l'interprétation et à la manipulation de documents textuels et pour la constitution des ressources nécessaires à cette assistance. Dès lors, l'évaluation de ces techniques supporte difficilement la mesure car cela impliquerait une possible mesure de qualité de l'interprétation qui est par essence ni absolue, ni universelle. La confrontation de notre modèle à des tâches déjà pourvues de techniques d'évaluation (comme la recherche documentaire ou la détection de phénomènes linguistiques singuliers avec les mesures : *rappel* et *précision*[65]) ne doit pas laisser entendre que ces techniques, par ailleurs souvent critiquées, lui sont applicables. Dans une approche centrée utilisateur, le problème de l'évaluation se trouve déplacé. Il ne s'agit pas tant d'évaluer l'application d'un modèle que l'efficacité et la faisabilité d'une interaction entre un utilisateur humain et un agent logiciel. Certaines propositions, en particulier dans le domaine du dialogue Homme/Machine préconisent déjà des modalités d'évaluation adaptées à des conditions d'interaction entre l'utilisateur et le système (*taux de compétence* et *taux d'efficacité* [Luzzati, 1996]). Cependant, LUCIA ne peut être soumis à de telles évaluations sans réflexion préalable : une adaptation est nécessaire pour pouvoir évaluer les caractéristiques des interactions proposées à l'utilisateur. Nos réalisations informatiques et les modèles

[65] Le rappel et la précision sont deux mesures utilisées dans le cadre de l'évaluation de systèmes de recherche documentaire sur des corpus préalablement traités par des experts. Ces mesures sont par exemple critiquées du fait du rôle prépondérant des experts décidant de la validité d'un document en fonction d'une requête. Le rappel et la précision pourraient être utilisés pour l'évaluation d'une tâche de détection d'emplois métaphoriques, moyennant l'accès à un corpus où les emplois métaphoriques ont déjà été repérés – ce qui, à notre connaissance, n'existe pas pour le moment. Nous reviendrons sur ces points dans le chapitre 5 (partie 5.5)

qui en sont à l'origine ne sauraient donc être évalués que par l'intermédiaire d'expériences et de discussions contradictoires sur des résultats obtenus en laboratoire ou dans des conditions réelles. L'atelier formation du CNRS *Variation, construction et instrumentation du sens*[66] que nous avons co-organisé en 2002 a été l'occasion de mettre en place une toute première évaluation. Nous souhaitions tester la capacité des participants à s'approprier les principes généraux du modèle en leur proposant de construire dans un temps imparti, un dispositif sur un sujet précis (la bourse) afin de pouvoir comparer les résultats. Cette expérience, menée conjointement avec Pierre Beust, a été présentée dans [Perlerin et Beust, 2003].

L'expérience s'est déroulée au cours de deux séances de deux heures trente chacune et avec un total de 8 participants d'horizons différents (linguistique, psychologie, ergonomie, informatique, microbiologie, sciences cognitives…) repérés par la suite par les codes suivants : SM, BA, MS, AA, JR, AP, IK et MC. Après un exposé d'environ une heure sur LUCIA et les principes d'analyse, nous avons fourni aux participants une liste de 216 entités lexicales issue du corpus *Le Monde sur CD-ROM.* Cette liste avait été obtenue à partir d'un calcul d'occurrences de termes (voir chapitre 4) sur l'ensemble des articles traitant de la bourse et de l'économie de laquelle nous avions supprimé tous les éléments non verbaux et non substantivaux[67] – les éléments de la *stop-list* pour le français fourni avec nos logiciels présentés dans le chapitre suivant (Chapitre 4 – p.111) et dont nous avions sélectionné arbitrairement 216 représentants. Les consignes données aux participants se bornaient à leur demander de construire un dispositif comme ils devraient le faire pour une tâche de veille documentaire, i.e. en organisant les entités par catégories thématiquement valables selon leur point de vue et leurs connaissances sur le domaine. Pour cette tâche, ces derniers avaient à titre d'exemple un dispositif « Météo » très proche de celui que nous présenterons dans le chapitre 5 (p.159) et qui nous a servi lors des expériences sur une étude de faits de langue. L'intérêt de leur présenter un dispositif construit pour une autre tâche que celle qui leur était soumise était de ne pas trop les influencer sur les informations à fournir.

[66] Cet atelier s'est déroulé du 10 au18 juillet 2002 sur l'île de Tatihou (50) sous la direction de Madame Anne Nicolle. Nous avons participé à son organisation. Il a donné lieu à la publication d'un ouvrage éponyme sous la direction de Madame Maryse Siksou [Siksou, 2003].

[67] Ce choix arbitraire nous a permis de limiter les catégories grammaticales des entités proposées pour simplifier le travail des participants. Rappelons que dans la partie 3.1, nous avons dit que les entités lexicales d'un dispositif LUCIA pouvaient être des noms propres et communs, des adjectifs, des verbes et des adverbes ou toutes combinaisons syntagmatiques redondantes et remarquables de ces éléments.

achat	commerciaux	front office	obligation	syndicaliste
acheteur	City de Londres	gouvernement	OPA	syndicat
action	COB	grève	opérateur	taux
actionnaire	contribuable	graphique	or	taux de change
affaires	corbeille	hausse des cours	palais Brongniart	titre
agent de change	cotation	indicateur	parité	transaction
analyste	cote	indice	participation	valeur
argent	courbe	industriel	perte	vendeur
back office	cours	inflation	petit porteur	vente
baisse des cours	Crédit Agricole	intérêt	place boursière	volume d'échange
banque	Crédit Lyonnais	investir	place financière	Wall Street
bénéfice	déflation	investisseur	portefeuille	
bénéficiaire	dévalorisation	krach	porteur	
bourse	dévaluation	libéraux	produit SICAV	
boursicoteur	dévaluer	marché	profit	
bureau	devise	MATIF	ratio	
CAC	dividende	métal	revente	
capital	échange	métal précieux	rue Vivienne	
capitaliste	économiste	middle office	salarié	
chiffre d'affaire	entreprise	mini krach	salle de marché	
chômeur	fond	monnaie	souscripteur	

Figure 32 - Liste des entités lexicales fournie lors de l'expérience.

Les participants ont presque tous travaillé sur papier car ils n'étaient pas installés à un poste de travail informatique pour créer leur dispositif. Des machines sur lesquelles se trouvait installé le logiciel LUCIA*BUILDER* (présenté dans le prochain chapitre), permettant la construction assistée de dispositifs, étaient à leur disposition dans la salle, mais il ne leur était pas explicitement demandé de s'en servir (aucune interdiction n'a pas non plus été formulée). Seul l'un des participants a utilisé son propre matériel, en l'occurrence un logiciel (*Inspiration* de Inspiration Software Inc.) originellement destiné à la réalisation de diagrammes et de schémas d'organisation.

À l'issue des deux séances d'expérience, aucun des participants des deux groupes n'a pu créer un dispositif contenant toutes les entités lexicales proposées (nous reviendrons plus loin sur les raisons de cet échec). Ils ont cependant tous pu proposer des groupes d'entités lexicales ; ils ont parfois précisé les différences qu'ils considéraient effectives au sein de ces groupes et créé des tables LUCIA avec un ou plusieurs attributs. Pour analyser ces résultats et éventuellement dégager des classes d'équivalence, des régularités ou des singularités, nous avons effectué un certain nombre de calculs (figure 33). Pour les groupes d'entités formés, qu'elles soient réunies dans des tables ou non, nous avons calculé le nombre d'entités lexicales communes 2 à 2 (noté NLC) et le pourcentage d'entités

lexicales communes entre les groupes (pourcentage d'appartenance des entités lexicales du groupe 1 dans le groupe 2 : G1/G2= NLC(G1,G2)/Card(G2)) et le taux de recouvrement T des groupes entre eux (T(G1,G2) = (G1/G2 + G2/G1)/200). Dans le tableau suivant, ces résultats sont classés par ordre décroissant de taux de recouvrement des groupes 2 à 2. (ex : AA0 représente les mots de la table n°0 du participant AA, ce groupe rassemble 89% des mots du groupes JR1).

G1	G2	NLC	G1/G2	G2/G1	c(G1)	c(G2)	T(G1,G2)
AA0	***JR1***	8	88,89%	100,00%	8	9	0,94
AA3	***MC0***	8	88,89%	80,00%	10	9	0,84
AA3	***JR0***	10	62,50%	100,00%	10	16	0,81
JR0	***MC0***	9	100,00%	56,25%	16	9	0,78
...	...	...	...	...	...	...	..

Figure 33 - Extrait du tableau relatif aux groupes de mots d'une table entière.

[AA0]	***Direction***
[AA0a] dévalorisation – dévaluation – krach – mini-krack – baisse des cours – dévaluer- déflation	*descend*
[AA0b] inflation	*monte*

Figure 34 - Table du participant AA en rapport avec les phénomènes dynamiques de la bourse.

[JR1]	***Direction***	***Connotation***
[JR1a] dévaluation – dévalorisation – dévaluer – baisse des cours - déflation	descend	-
[JR1b] krach – mini-krach	descend	*mal*
[JR1c] inflation	monte	*mal*
[JR1d] hausse des cours	monte	-

Figure 35 - Table du participant JR en rapport avec les phénomènes dynamiques de la bourse.

[MC0]	***Connotation***	***Rapport à l'action***
[MC0a] actionnaire – boursicoteur – investisseur – porteur - petit porteur	sans	rôle
[MC0b] bénéficiaire	bien	rôle
[MC0c] analyste - agent de change – économiste	sans	profession
	bien	profession

Figure 36 - Table du participant MC en rapport avec les acteurs de la bourse.

[AA3]	***Action***	***Rapport à l'activité / Résultat***
[AA3a] actionnaire - souscripteur	achat	investissement
[AA3b] bénéficiaire	reçoit	retour sur investissement
[AA3c] boursicoteur	achat/vente	investissement
[AA3d] porteur - petit porteur	-	-
[AA3e] analyste - économiste	analyse	étude/ observation
[AA3f] opérateur	-	-
[AA3g] agent de change	traitement opération	travail

Figure 37 - Table du participant AA en rapport avec les acteurs de la bourse.

Parmi les résultats obtenus des participants, 20 tables entières ont pu être soumises aux calculs exposés ci-dessus. Les tables sont comparées 2 à 2 ce qui représente 380 groupes de 2 tables à considérer dans les calculs. Au final, 13 groupes de 2 tables (6,84%) présentent un taux de recouvrement supérieur à 0,5 parmi lesquels 4 groupes de 2 tables (2,11%) présentent plus de 10 entités lexicales communes. 54 groupes de 2 tables (28,84%) présentent un taux de recouvrement non nul et 12 groupes

de 2 tables (6,32%) présentent plus de 5 entités lexicales communes. En comparant ces résultats au matériel fourni par les participants, on peut apprécier le fait que les tables ayant 2 à 2 un taux de recouvrement le plus important, concernent majoritairement les acteurs du monde boursier comme les tables MC0 et AA3 (figure 36 et figure 37) par exemple. Ce sont ces tables que l'on rencontre généralement en haut du tableau classant l'ensemble des tables formées par ordre décroissant de taux de recouvrement. Par exemple : AA3, MC0, JR0, AL1 et SG5 ont 2 à 2 des taux de recouvrement supérieur à 0,7 et présentent toutes des entités lexicales pouvant avoir trait à des personnes physiques (*boursicoteur*, *agent de change*…). Le deuxième groupe de tables qui présente des taux de recouvrement importants (au moins supérieur à 0,3) présente majoritairement des tables en rapport avec les *phénomènes* boursiers. Par exemple, AL3, JR1, AA0 et SG2 ont 2 à 2 majoritairement des taux de recouvrement supérieurs à 0,3 et sont toutes des tables relatives aux entités lexicales pouvant avoir trait à des phénomènes dynamiques (*dévaluation, déflation, baisse et hausse des cours…*- voir figure 34 et figure 35 pour AA0 et JR1). Les tables proposées se conforment principalement aux catégories ontologiques lieux, acteurs, phénomènes, tout en présentant des particularités remarquables propres à chacun des participants. On peut noter également que les deux catégories ayant fait l'objet du plus grand nombre de tables quasi-identiques (*phénomènes* et *acteur*) regroupent des entités lexicales apparaissant pour les participants comme facilement associables à ces sujets même sans recours à un corpus. Au contraire, les entités lexicales ayant généralement trait à des lieux et/ou des institutions par exemple (*entreprise*, *banque*…) semblent plus difficiles à différencier et apparaissent donc plus rarement dans des tables analogues d'un participant à l'autre. À ce propos, on peut noter que les consignes qui demandaient de construire un dispositif en rapport avec la bourse, les entités lexicales comme *entreprise* et *banque,* n'étaient pas forcément identifiées comme faisant classiquement partie de ce domaine. Il semble qu'un consensus se soit dégagé autour de certaines entités censées avoir directement trait à la bourse et que celles relevant plutôt de l'économie ou des influents de la bourse aient été plus difficiles à classer.

Des calculs similaires à ceux exposés ci-dessus ont été effectués sur les lignes des tables construites et sur les groupes d'entités lexicales non structurés en tables. Pour les lignes, les calculs ont porté sur 44 lignes comparées 2 à 2 (donc 1892 lignes à considérer dans les calculs) : 60 groupes de lignes (3,10%) présentent un taux de recouvrement supérieur à 0,5 et 54 groupes (9,51%) un taux de recouvrement non nul. On peut observer également que seuls 5 groupes de lignes (0,53%) présentent plus de 5 mots communs. Parmi les lignes à un seul mot, 9 seulement sont parfaitement identiques. Des similarités apparaissent pour des lignes à plusieurs mots (exemple : AA0 et JR1, figure 34 et figure 35 avec la même valeur d'attribut 'descend' pour [Direction]) mais ces cas restent rares au vu de l'ensemble des tables construites. Pour les groupes d'entités non structurés, une dizaine (sur 29) présente des taux de recouvrement égaux à 100% mais aucun n'est parfaitement identique d'un participant à un autre.

Au vu de tous ces résultats et après entretien avec les participants, nous avons pu constater que l'expérience présentait un certain nombre de défauts. Le premier est certainement dû au fait que le temps imparti était trop court pour la réalisation du travail demandé sans l'aide d'un support informatique dédié. Le second, plus difficile à prendre en considération pour la mise en œuvre d'une expérience au cours d'un tel atelier, est l'absence de tâche réelle comme contexte de constitution de ressources lexicales. La présence majoritaire des catégories ontologiques est très certainement due à ce manque. L'absence du corpus d'origine et donc l'impossibilité de revenir sur un texte faisant intervenir les entités lexicales proposées a également été ressentie comme un handicap par les participants : comment interpréter une entité lexicale sans cotexte ? Ces variables seront donc à redéfinir lors d'expériences ultérieures. Il est important de noter que ce retour sur corpus est envisagé dans la démarche cyclique de construction des dispositifs. Une expérience sans corpus permettait simplement de tester la faisabilité de la construction de tels matériaux, d'apprécier la capacité des participants à amorcer ce processus itératif. En l'occurrence, nous avons constaté que la méthode de construction des dispositifs s'acquiert rapidement et que les principes qui la régissent sont aisément assimilables par des non-spécialistes. Les différences et les points communs découverts au sein des résultats fournis par les participants montrent dans une certaine mesure que les utilisateurs intègrent leur propre sensibilité par rapport à un domaine - cette sensibilité pouvant relever d'une méconnaissance totale de ce domaine. À ce propos, on constate qu'une mauvaise connaissance d'un domaine n'empêche en rien les sujets d'avoir une compétence langagière sur celui-ci. On peut en déduire une différence fondamentale entre les connaissances ontologiques sur un domaine et son lexique. Nous avons pu constater différentes méthodologies de construction des dispositifs chez les participants : certains établissaient d'abord des classes de mots pour chercher ensuite à en différencier les représentants, d'autres recherchaient d'abord les différences pertinentes pour ensuite créer les tables et y faire figurer les mots du domaine. C'est ce constat qui a servi de base à nos propositions logicielles présentées dans le prochain chapitre.

Cette expérimentation a permis d'apprécier la capacité d'utilisateurs potentiels à s'approprier les principes du modèle que nous proposons. La tâche de description des significations n'est pas triviale. Cela a déjà été souligné par exemple dans [Kerbart-Orecchioni, 1998] : *« un des problèmes majeurs que pose (…) la description des structurations lexicales réside dans le fait qu'elles tiennent à la fois des systèmes diacritiques (non hiérarchiques) et des systèmes taxinomiques (hiérarchiques) ».* Cependant, nous avons pu constater que les utilisateurs sont parvenus à formuler dans un temps raisonnable des représentations lexicales reflétant un point de vue particulier sur le domaine proposé. Ceci constituait l'une des hypothèses que nous cherchions à vérifier. Durant les deux séances de cette expérimentation et, à travers les différences et les points communs entre les dispositifs et les groupes de mots fournis par les participants, reflet de leurs capacités interprétatives, nous avons pu considérer à sa juste valeur la dimension sociale et partagée du langage, latente en chaque locuteur. Cela a renforcé notre point de vue résolument centré sur l'utilisateur.

3.5 Conclusion

Dans ce chapitre, nous avons vu que le modèle LUCIA permet l'expression d'un point de vue particulier sur certaines entités lexicales d'un domaine en les organisant selon les points communs et les différences examinés de leur signification en contexte. Nous verrons dans le chapitre suivant que les entités lexicales en question peuvent provenir des connaissances des usagers ou être acquises de façon assistée depuis un corpus. Les dispositifs n'ont pas de visées ontologiques ou uniquement référentielles, et peuvent se soustraire à une certaine validité lexicologique et à un consensus large sur les propositions interprétables qu'ils englobent. À travers eux, ce ne sont pas des concepts mais des éléments de signification proposés par le lecteur/utilisateur qui sont soumis aux logiciels. Contrairement à d'autres travaux informatiques dont nous reconnaissons cependant l'influence, nous n'éludons pas le contenu référentiel des signes pour des questions *d'assainissement théorique* [Tanguy, 1997a* : 23] ou à cause d'un rejet systématique d'une sémantique dénotationnelle qui, de toute façon, reste difficilement concevable lorsqu'on prétend proposer des solutions utilisables par des non-spécialistes de la langue – nous avons pu apprécier ce fait à travers l'étude des résultats obtenus de la première expérience de réalisation de dispositifs.

Dans LUCIA, il n'y a pas de distinctions des catégories populaires et des catégories savantes [Nyckees, 1998* : 319] ; il n'y pas non plus de distinctions entre des relations entre signes et signifiés qui seraient référentielles – voire scientifiques – et d'autres que l'on pourrait penser plus naturelles ou spontanées (au moins dans une certaine synchronie - voir [Rastier, 1987*] p.161). Les relations en question ne relèvent pas du sémanticien, ce sont celles de l'interprétant dans le cadre de sa pratique : à charge pour lui d'entreprendre alors son interprétation dans le cadre des objectifs qu'il s'est fixés pour la tâche en cours ; il pourra toujours réviser ses jugements en fonction des résultats obtenus de façon automatique. Le modèle de catégorisation proposé donne un cadre computationnel efficace pour l'exploitation des attributs/valeurs, pendants informatiques des sèmes. Dans les chapitres suivants, nous nous proposons de définir les cadres d'utilisation précis du modèle et de présenter les moyens informatiques nécessaires à sa mise en œuvre concrète.