Construction d'une base de connaissances lexicographiques :
Les marqueurs superficiels dans les définitions spécialisées du *Trésor de la Langue Française*

Nabil Hathout
Institut National de la Langue Française – CNRS
Château du Montet. Rue du Doyen Roubault
F–54500 Vandœuvre-lès-Nancy
e-mail : hathout@ciril.fr

12 février 1996

Introduction

La diffusion dans un large public de la micro-informatique, et plus précisément des lecteurs de CD-Rom, a rendu viable économiquement l'édition de dictionnaires électroniques. Ces dictionnaires constituent une source importante de connaissances lexicographiques. Leur existence dans un format électronique rend possible leur manipulation par des outils informatiques. Il n'est cependant pas possible de les utiliser comme des lexiques informatiques car les informations qu'ils contiennent ne sont pas représentées de façon formelle et homogène. Certains dictionnaires, comme le *Robert Électronique* par exemple, ne distinguent même pas les définitions des exemples. Les articles de ces dictionnaires ne peuvent, par conséquent, être utilisés que comme des corpus de textes lexicographiques.

L'INaLF s'est engagé, depuis 1992, dans un projet d'informatisation du *Trésor de la Langue Française* (TLF). Dans ce cadre, Jacques Dendien (1994) a réalisé une première maquette qui permet la consultation informatique du volume XIV du TLF. Ce système distingue tous les objets textuels qui composent les articles, et permet de les sélectionner séparément si l'utilisateur le souhaite.

Les recherches que nous présentons se situent en aval de ce projet. Elles ont pour objectifs l'**extraction de connaissances lexicographiques** à partir du TLF et l'organisation de ces connaissances en une **base de connaissances lexicologiques** (BCL).

La première section de cet article décrit dans ses grandes lignes la **méthode d'analyse** que nous proposons. La seconde section aborde deux points importants de cette méthode : la **caractérisation sémantique** des segments qui composent les définitions et la structuration des informations qu'ils contiennent sous forme de **fiches lexicographiques**.

1 Construire un lexique formel

La constitution d'une BCL à partir des articles du TLF peut être décomposée en plusieurs étapes :

- la constitution d'un **lexique formel** qui regroupe les représentations formelles des *définitions* ;
- la désambiguïsation du sens des mots qui apparaissent dans les représentations qui composent le lexique formel (Martin, 1994) ;

- la correction des omissions, incohérences, imprécisions... contenues dans les définitions ;
- la prise en compte des *restrictions ou spécifications d'emploi*, des *conditions d'emploi entre-crochets*, des *synonymes* et des *antonymes* ;
- l'enrichissement de la BCL à l'aide des informations contenues dans les *exemples* (ex. détermination de la structure argumentale des prédicats) ;

Nous abordons ici la première de ces étapes.

Les *définitions* constituent le centre des articles d'un dictionnaire et véhiculent l'essentiel de l'information définitoire. Par ailleurs, elles font l'objet de nombreuses études en lexicologie, en terminologie et en linguistique (Chaurand & Mazière, 1990). Pour autant, leur analyse automatique en vue d'en extraire des informations lexico-sémantiques est une tâche non triviale qui pose de nombreux problèmes (Bougarev & Briscoe, 1989; Artola Zubillaga, 1993; Atkins & Zampolli, 1994).

À la suite de (Martin, 1992), nous considérons qu'une définition est une relation entre un *défini* (la vedette de l'article) et un *définissant* (le texte qui paraphrase ou décrit le défini). Les définissants ont le plus souvent une structure syntaxique et sémantique simple. Leur forme canonique comprend deux parties :

- une partie *genus*, à savoir un hyperonyme du défini ;
- une partie *differentia* qui décrit les caractères discriminants qui distinguent l'ensemble dénoté par le défini dans celui dénoté par le *genus*.

Pour construire un lexique formel à partir des définitions du TLF, nous avons adopté une méthode d'analyse partielle, **dirigée par la sémantique**. Dans un premier temps, nous nous limitons aux seules définitions spécialisées (i.e. qui relèvent d'un domaine de spécialité). Ainsi, l'analyse de chaque définition peut être guidée par le **canevas des fiches lexicographiques** de son domaine, c'est-à-dire par un ensemble de relations sémantiques habituellement utilisées pour définir les entités de ce domaine (ex. `est-composition-de` ; `a-pour-origine` ; `a-pour-fonction`). Ces fiches, inspirées à la fois de celles des actions MUC (Chinchor et al., 1993), des *réseaux définitoires* proposés dans (Martin, 1993) et des entrées du *Lexique génératif* (*Generative Lexicon*, abrégé GL (Pustejovsky, 1995)), constituent, une fois remplies, les entrées du lexique formel.

Segmentation. Un définissant est une phrase dans laquelle sont articulées un certain nombre d'informations sur le défini. Il est en général possible d'identifier, pour chacune de ces informations, un segment dans lequel elle est exprimée. De façon duale, chaque définissant peut être divisé en un certain nombre de segments dont chacun exprime une information particulière sur le défini. Par exemple, on peut découper le définissant du substantif « *ruthénium* », de la manière suivante :

(1) [Élément] [de numéro atomique 44] [, de symbole *Ru*] [, métal] [blanc argent] [, brillant] [, dur] [, cassant] [, utilisé en particulier en alliage dans la fabrication de contacts électriques ainsi que pour certaines catalyses] .

On peut, de ce fait, remplacer l'analyse globale des définissants par une **analyse locale** de chacun de leurs segments. On améliore ainsi la **robustesse** et l'**efficacité** de l'analyseur, puisque les segments ont des structures syntaxiques plus simples que les phrases et que l'échec d'un traitement sur un segment ne compromet pas l'analyse des autres segments.

Pour segmenter un définissant, il faut en faire une analyse **superficielle** qui comprend :

1. L'étiquetage morpho-syntaxique des mots de la phrase (Brill, 1992; Lecomte & Paroubek, 1994).
2. La décomposition de cette phrase en suites de mots élémentaires. On peut s'inspirer pour cela des techniques proposées dans (Alshawi, 1989; Constant, 1991).
3. Le rattachement des groupes prépositionnels, en se basant sur des tables (Boons, Guillet, & Leclère, 1976; Guillet & Leclère, 1992) et sur l'apprentissage endogène (Bourigault, 1994).

4. La recomposition des segments par découpage/regroupement des suites de mots.

L'exemple (1) permet de voir qu'il existe des **segments élémentaires** composés d'un seul « élément sémantique » (ex. [dur], [brillant]), et des **segments complexes** composés de plusieurs éléments, (ex. un prédicat et un de ses arguments). Dans la suite de l'article, nous nous intéressons surtout aux segments complexes car les segments élémentaires présentent des structures syntaxiques triviales et n'ont donc pas besoin d'être analysées. Leur représentation formelle est également élémentaire puisqu'elle est constituée par un lien vers (i.e. un pointeur sur) l'entrée, dans le lexique formel, du lemme de la forme qu'ils contiennent.

Caractérisation des segments. L'étape suivante de l'analyse d'un définissant consiste à identifier le type sémantique de l'information contenue dans chaque segment. Par exemple, on peut associer aux segments de (1) les étiquettes suivantes :

(2) [Élément]$_{\text{classification}}$ [de numéro atomique 44]$_{\text{numéro atomique}}$ [, de symbole *Ru*]$_{\text{symbole chimique}}$ [, métal]$_{\text{classification}}$ [blanc argent]$_{\text{couleur}}$ [, brillant]$_{\text{luminance}}$ [, dur]$_{\text{résistance à la pression}}$ [, cassant]$_{\text{élasticité}}$ [, utilisé en particulier en alliage dans la fabrication de contacts électriques ainsi que pour certaines catalyses]$_{\text{utilisation}}$.

Cette caractérisation permet d'utiliser, pour chaque segment, une procédure d'analyse adaptée aux spécificités de son type.

Le calcul du type des segments peut être réalisé de différentes manières. Par exemple, la caractérisation d'un segment élémentaire se fait par héritage de la classe sémantique de l'entrée associée à sa représentation formelle. Pour les segments complexes, elle peut être déterminée par une recherche d'indices superficiels, comme des marqueurs métalinguistiques ou des configurations syntaxiques.

Analyse partielle des segments. Une fois les segments catégorisés, on peut en faire une analyse partielle, c'est-à-dire que l'analyseur construit une représentation syntaxico-sémantique qui ne prend pas nécessairement en compte tous les mots du segment. De plus, elle peut être constituée de plusieurs parties lorsque le rattachement de certains constituants à l'arbre principal n'est pas décidable. Par exemple, on pourrait avoir pour le dernier segment de l'exemple (1), les représentations données en figure 1. (Pour des raisons de mise en page, les deux principaux sous-arbres de la première représentation sont dessinés séparément.)

Dans cette représentation, la locution adverbiale « *en particulier* » et la conjonction de coordination « *ainsi que* » ont été ignorées. D'autre part, le groupe prépositionnel « *en alliage* » n'a pas pu être rattaché à l'arbre principal. Une analyse sémantique fine, ultérieure, permettrait de le rattacher comme modificateur du prédicat `fabrication`.

L'analyse partielle des segments est guidée par les champs qu'il faut remplir dans la partie de la fiche lexicographique correspondant au type sémantique du segment. Par exemple, dans le cas d'un segment `utilisation`, ces champs sont, outre le prédicat lui-même, son agent, son thème, le domaine...

Construction d'une représentation formelle du définissant. Cette étape correspond au remplissage proprement dit des fiches lexicographiques. Elle consiste à traduire les représentations syntaxico-sémantiques des définissants en éléments de fiches. Par exemple, à partir de l'exemple (1), il est possible de produire la représentation proposée en figure 2, inspirée du formalisme GL.

Discussion. L'objectif de la décomposition du processus d'analyse est de ne considérer, pour les deux derniers traitements (i.e. l'analyse syntaxico-sémantique et la construction de la représentation formelle), que les éléments qui interviennent dans le calcul de la représentation des informations précisées dans le canevas des fiches lexicographiques du domaine de la définition. On limite ainsi la sur-génération de représentations possibles par ces deux modules, cette tendance étant due à la complexité des traitements qu'ils réalisent.

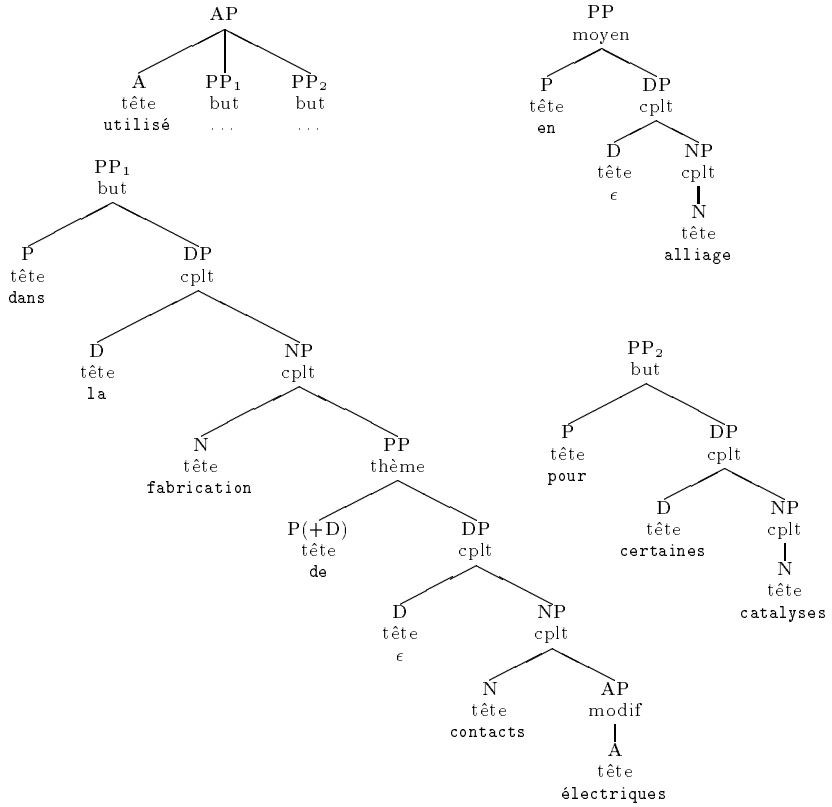

FIG. 1 - *Représentations syntaxiques du dernier segment de l'exemple (1)*

Parmi les quatre étapes présentées dans cette section, la segmentation joue un rôle essentiel car les trois suivantes en dépendent fortement. Il est donc indispensable qu'elle ait une bonne précision. D'autre part, c'est la plus difficile car la plus « superficielle ».

2 Classes sémantiques et marqueurs superficiels

Cette section aborde de manière plus approfondie deux des points de la méthode que nous venons de présenter : les fiches lexicographiques en 2.2 et la caractérisation du contenu sémantique des segments en 2.3. Ce travail s'appuie sur une étude linguistique dont les principaux résultats sont données en 2.1.

2.1 Étude linguistique

L'étude présentée ici porte sur les définitions de noms concrets des domaines de la **botanique** et de la **chimie**. Cette étude a pour objectifs d'**identifier les informations sémantiques** qui apparaissent dans ces définitions et de **collecter les marqueurs superficiels** de ces informations.

2.1.1 Botanique

Les définitions de noms concrets de ce domaine sont composées d'informations qui se répartissent dans cinq classes sémantiques. Les marqueurs superficiels de ces informations sont présentés dans le tableau 1.

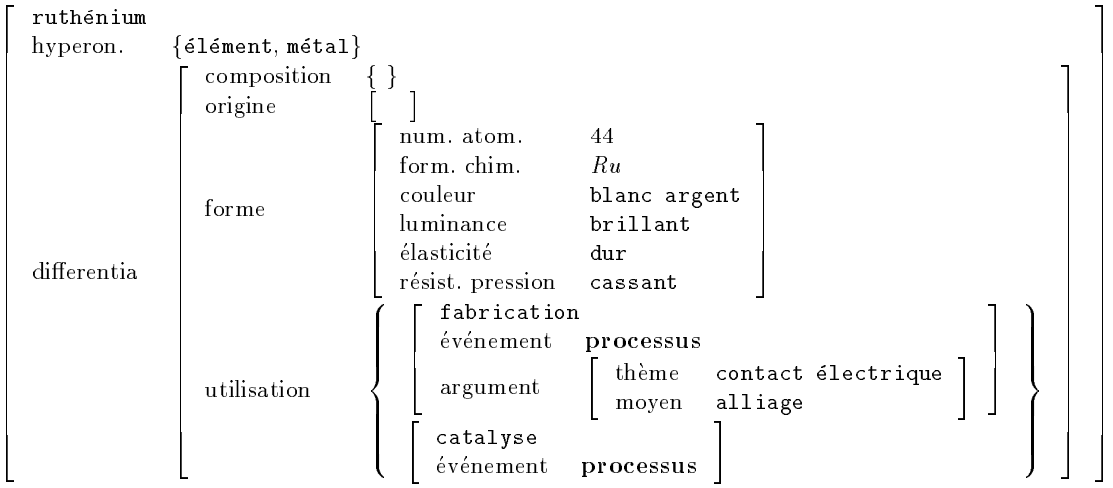

FIG. 2 - *Représentation formelle pour l'exemple (1)*

Classification. On peut distinguer trois formes de classification. La première, lexicographique, correspond à la partie *genus* du définissant (ex. `arbrisseau`). La deuxième, terminologique (ex. `cornouiller du Canada`), se confond généralement avec la précedente. La dernière, « botanique » (ex. `de la famille des Papilionacées`), est relative au système taxonomique de ce domaine, les niveaux de ce système étant l'espèce, le genre, la famille, l'ordre, la classe... Dans les définitions, la classification botanique se distingue des deux premières par le fait qu'elle est systématiquement signalée par un marqueur explicite, à savoir le nom du niveau taxonomique.

Morphologie. Cette classe regroupe les informations relatives à la taille, la couleur, l'odeur, le goût, la disposition, les parties et la ressemblance.

Durée de vie. Elle est exprimée par trois adjectifs de la langue de spécialité botanique : `vivace` ; `annuelle` ; `bisannuelle`. Ces informations ne sont donc jamais marquées.

Localisation. Elle est essentiellement géographique, et est souvent marquée par des verbes qui décrivent le mode de développement des plantes (ex. `croître dans` ; `pousser sur` ; `se développer près de`).

Utilisation. Dans les définitions de botanique, les informations de finalité, de fonction, etc. sont toutes relatives à l'utilisation des plantes. Ces informations concernent presque toujours une partie de la plante et sont souvent signalées par des marqueurs.

Marqueurs. Le tableau 1 présente les marqueurs superficiels d'informations sémantiques qui ont été identifiés par l'étude linguistique. Il donne également le « type lexical » de ces marqueurs ; les abréviations utilisées sont : LE = lexicalement explicite ; $EVAS$ = expression à verbe ou adjectif support ; LNE = lexicalement non explicite.

2.1.2 Chimie

On retrouve pour la chimie un certain nombre des classes sémantiques identifiées pour la botanique. On a ainsi, pour les noms concrets, les classes suivantes :

Classe	Type	Marqueurs
Classification	LE	variété ; espèce ; genre ; tribu ; sous-famille ; famille ; ordre ; sous-classe ; classe ; sous-embranchement ; embranchement ; division
	LE	appartenir ; comprendre ; renfermer ; regrouper ; grouper ; faire partie de
	LE	représenté par ; type
Odeur	LE	odeur ; odorant ; odoriférant
Goût	LE	goût ; saveur
Disposition	LE	disposé en ; se présenter en ; porté ; porter
	LNE	présenter
Méréonomie	LE	muni ; contenir
	LNE	caractérisé par ; remarquable par
Approximation	LE	ressembler à ; semblable à ; proche de ; avoir l'aspect de ; rappeler
Localisation	EVAS	croître dans ; croître en ; croître sur ; cultivé dans ; cultivé sous ; cultivé en ; se cultiver en ; pousser à ; pousser dans ; pousser sur ; se développer dans ; se développer près de ; vivre au bord de ; vivre dans ; vivre sur
	EVAS	abonder dans ; présent/ADJ jusqu'en ; venir sur ; visible sur
	LE	enfoui/ADJ dans ; d'origine ; originaire de ; répandu dans ; répandu en
	LE	région ; mer ; pays ; montagne ; zone ; hémisphère
Utilisation	LE	utilisé comme ; utilisé dans ; utilisé en ; utilisé pour ; employé dans ; employé en ; employé pour ; avoir un usage ; fournir ; servir à ; servir de ; donner
	EVAS	cultivé pour ; cultivé comme ; cultivé à des fins
	LNE	propriété

Tab. 1 - *Botanique : les classes sémantiques et leurs marqueurs.*

Classification. Comme en botanique, il y a, en chimie, trois formes de classification, mais contrairement à la classification botanique, celle qui est relative à la nomenclature de la chimie n'est pas signalée par des marqueurs explicites.

Composition. Les informations de composition jouent, en chimie, un rôle similaire aux informations méréonomiques en botanique : les produits chimiques sont souvent des mélanges de produits plus simples qui les caractérisent ; les plantes sont, pour leur part, caractérisées par leurs parties naturelles.

Origine. Les informations d'origine correspondent à celles de localisation (qui n'est autre que l'origine spatiale) dans les définitions de botanique.

Utilisation. Les informations d'utilisation et de fonction sont, en chimie comme en botanique, bien adaptées à une analyse à base de marqueurs superficiels car elles sont presque toujours signalées. On constate, par ailleurs, que plusieurs marqueurs d'utilisation sont communs aux deux domaines.

Propriétés perceptibles. Cette classe regroupe d'une part, les propriétés naturellement perceptibles : la couleur, l'odeur, le goût ; la densité ; l'élasticité ; la luminance ; la résistance aux chocs ; la volatilité... et d'autre part, des propriétés ne pouvant être mises en évidence qu'à l'aide

d'instruments ou par des expériences, comme: l'explosibilité; le magnétisme; la radioactivité; la résistance à la corrosion; la solubilité dans l'eau; la solubilité dans les solvants; la toxicité... La plupart des segments qui réalisent ces informations sont élémentaires (ex. `cassant`; `dur`; `explosif`; `siccatif`), et ne contiennent donc pas de marqueur superficiels.

Propriétés chimiques théoriques. Dans les définitions de chimie, les propriétés « théoriques » (formule chimique; symbole; poids atomique; numéro atomique) se distinguent des propriétés perceptibles par le fait qu'elles sont systématiquement signalées par des marqueurs explicites. Elles sont, en ce sens, similaires aux informations de classification taxonomique en botanique.

Marqueurs. Le tableau 2 regroupe les marqueurs d'informations sémantiques dans les définitions de chimie.

Classe	Type	Marqueurs
Classification	LE	`de la famille de` ; `du groupe de`
Composition	LE	`combinaison de` ; `composé de` ; `contenir` ; `mélange de`
	LE	`composé de` ; `constitué par` ; `muni de`
	LE	`constituant`/SBC `de` ; `entrer dans la composition de`
Origine	LE	`dériver de` ; `dérivé`/ADJ `de` ; `extraire de` ; `extrait`/ADJ `de` ; `formé au cours de` ; `formé par` ; `se former dans` ; `obtenu au cours de` ; `obtenu de` ; `obtenu en` ; `obtenu par` ; `obtenu à partir de` ; `produit`/ADJ `par` ; `produit`/SBC `de` ; `résulter de` ; `synthétisé à partir de` ; `tiré de`
	EVAS	`contenu`/ADJ `dans` ; `présent`/ADJ `dans` ; `répandu dans`
	LNE	`trouver`
Utilisation	LE	`donner` ; `destiné à` ; `employé comme` ; `employé contre` ; `permettre de` ; `servir à` ; `source de` ; `utilisé comme` ; `utilisé dans` ; `utilisé en` ; `utilisé pour`
	LNE	`propriété`
Odeur	LE	`odeur`
Formule chimique	LE	`formule`
Poids atomique	LE	`poids atomique` ; `masse atomique`
Numéro atomique	LE	`numéro atomique`

Tab. 2 - *Chimie : les classes sémantiques et leurs marqueurs.*

2.2 Fiches lexicographiques

Les classes d'informations sémantiques qui composent habituellement les définitions d'un domaine technique jouent un rôle central dans l'analyse de ces dernières. Du fait du nombre et de la diversité des domaines (le tome XIV en contient près de 800), on ne peut établir, pour chacun d'eux, la liste de ces classes en réalisant une étude linguistique approfondie d'un sous-ensemble de ses définitions. Par ailleurs, les classes sémantiques issues de telles études présentent certains inconvénients. Elles ne prennent pas en compte l'existence de marqueurs permettant de les identifier, comme le montrent les différences qui existent entre les classes de 2.1.1 et le tableau 1, et celle de 2.1.2 et le tableau 2. Étant constituées indépendamment des autres domaines, elles laissent une place importante aux particularités de leur domaine (ex. la taxonomie botanique par rapport à la classification), et sont donc plus difficilement généralisables. Ceci a pour conséquence une absence de cohérence entre les classes de domaines différents, d'où une impossibilité de les utiliser telles

quelles comme canevas de fiches lexicographiques : une harmonisation préalable de ces classes est nécessaire si on veut réunir au sein d'un même lexique les représentations formelles de définitions de domaines différents.

Canevas générique. Une solution à ces problèmes consiste à définir une liste de classes génériques que l'on pourra instancier pour chaque domaine. Cette liste comprend deux parties, l'une **commune** à tous les domaines et l'autre **spécifique** à chaque domaine particulier. En nous appuyant sur l'étude linguistique dont les résultats sont présentés en 2.1, sur les travaux de Martin (1993) et de Pustejovsky (1995), nous proposons que la partie commune contienne les informations de classification, de composition, d'origine et d'utilisation. Toutes les autres informations sont regroupées dans la partie spécifique.

Les informations spécifiques servent à distinguer le défini des autres entités du domaine en se basant sur les caractères du défini qui sont pertinents *pour ce domaine*. Elles constituent le noyau de la partie *differentia* du définissant, et correspondent au *rôle formel* de la *structure qualia* de GL. Ceci est illustré par le tableau 3.

Canevas générique		GL	
Partie	Classe	Structure	Rôle
commune	classification	héritage lexical	
	composition	qualia	constitutif
	origine		agentif
	utilisation		télique
spécifique			formel

Tab. 3 - *Correspondance avec GL*

La liste générique induit directement un **canevas générique** qui définit la structure des entrées du lexique formel et qui peut être représenté comme suit :

(3)
$$\begin{bmatrix} \texttt{vedette} & \\ \text{commune} & \begin{bmatrix} \text{classification} & \ldots \\ \text{composition} & \ldots \\ \text{origine} & \ldots \\ \text{utilisation} & \ldots \end{bmatrix} \\ \text{spécifique} & \ldots \end{bmatrix}$$

Ce canevas apporte une réponse aux inconvénients précédemment cités : la généralité des canevas des fiches lexicographiques de chaque domaine particulier est garantie par le fait que tous dérivent de la même structure générique ; les entrées de domaines différents sont ainsi compatibles, en particulier au niveau du contenu des classes et de leur dénomination (à condition de fixer au préalable une nomenclature de propriétés spécifiques).

Le tableau 4 met en correspondance les classes de 2.1.1 et 2.1.2 avec le canevas générique qui vient d'être proposé. On voit dans ce tableau que le canevas générique induit sur ces classes des modifications importantes. Ainsi, les informations de classification relatives aux taxonomies botanique et chimique deviennent spécifiques car tous les domaines ne disposent pas de taxonomies ou de nomenclatures similaires. La durée de vie des plantes perd l'importance qu'elle avait tandis que les informations méréonomiques sont promues en étant rattachées au champ **composition** des fiches lexicographiques.

Instanciation du canevas générique. Le canevas générique doit être instancié à deux niveaux différents pour pouvoir être utiliser pour encoder les définitions d'un domaine technique. La

Canevas générique		Botanique	Chimie
commune	classification	classification lexicographique classification terminologique	classification lexicographique classification terminologique
	composition	méréonomie disposition	composition
	origine	localisation	origine
	utilisation	utilisation	utilisation
spécifique		classification botanique morphologie (autres) durée de vie	classification chimique propriétés perceptibles propriétés chimiques théoriques

Tab. 4 - *Correspondances des classes sémantiques*

première instanciation doit déterminer quelles classes d'informations apparaissent dans la partie spécifique (ex. couleur, taille, toxicité...).

La seconde concerne la réalisation des classes d'informations listées dans cette structure. Par exemple, le fait que les définitions de tous les domaines contiennent des informations d'origine ne signifie pas qu'elles soient réalisées (et en particulier marquées) de manière identique. En botanique, ces informations concernent la localisation des plantes car ce sont des objets concrets naturels et inanimés. En chimie, elles peuvent également portent sur les processus d'extraction ou de fabrication des objets artificiels. La nature des objets du domaine constitue ainsi l'un des paramètres de cette seconde instanciation.

2.3 Marqueurs superficiels

L'étude des définitions de botanique et de chimie nous a permis de distinguer trois types lexicaux de marqueurs : les marqueurs lexicalement explicites, les expressions à verbe ou adjectif support et les marqueurs lexicalement non explicites. Les marqueurs peuvent également être caractérisés par le type des informations sémantiques qu'ils signalent. Ces deux classifications définissent une typologie qui permet d'envisager l'utilisation, pour chacun de ses types, de techniques de collectes semi-automatiques spécifiques.

Marqueurs lexicalement explicites. Les marqueurs lexicalement explicites d'informations appartenant à la partie commune du canevas (3) sont relativement indépendants des domaines particuliers. On peut donc envisager d'en établir une liste générale à partir d'un corpus composé de la totalité des définitions spécialisées. Les marqueurs d'origine et d'utilisation peuvent être collectés à partir de configurations comme `<part. passé> + <prép.>` ou `<adj.> + <prép.>` extraites de manière automatique. Dans ces configurations, `<prép.>` doit être spatiale ou temporelle pour l'origine ; elle doit pouvoir introduire un domaine (ex. `en`, `dans`) ou une finalité (ex. `pour`, `comme`) pour l'utilisation. Les résultats de cette collecte peuvent ensuite être triés automatiquement à l'aide de techniques similaires à celles que propose Bourigault (1994) : par exemple, on peut constituer une liste de **candidats marqueurs** d'utilisation en sélectionnant les verbes et les adjectifs qui apparaissent avec plusieurs prépositions d'utilisation et qui n'apparaissent pas avec des prépositions d'autres types. Les listes ainsi obtenues doivent ensuite être validées manuellement pour supprimer les candidats qui ne sont pas des marqueurs.

Les informations spécifiques sont ou bien fortement dépendantes des domaines techniques (ex. la classification botanique) ou bien relativement indépendantes de ces derniers (ex. l'odeur, le goût). Elles ne peuvent, dans les deux cas, être mises en évidence que par des études linguistiques fines. On ne peut, de ce fait, envisager de collecter leurs marqueurs de manière systématique. Par ailleurs, l'importance de ces informations est assez variable d'un domaine à l'autre.

Expressions à verbe ou adjectif support. Les expressions à verbe ou adjectif support qui signalent des informations de la partie commune de (3) dépendent plus du domaine que les marqueurs lexicalement explicites. On ne peut donc pas envisager, pour eux, une collecte générale sur l'ensemble du corpus des définitions spécialisées. On peut en revanche réaliser des collectes séparées dans les définitions de chaque domaine en s'appuyant sur les prépositions associées aux différentes classes sémantiques. Le filtrage automatique doit alors conserver les prédicats (ex. les verbes) qui apparaissent plusieurs fois avec des prépositions de types différents.

On peut noter que les marqueurs d'informations de la partie spécifique du canevas, identifiés dans les définitions de botanique et de chimie, sont tous lexicalement explicites. En particulier, aucun n'est une expression à verbe ou adjectif support.

Marqueurs lexicalement non explicites. Les marqueurs lexicalement non explicites constituent un phénomène marginal que nous ne prendrons pas en compte dans la suite de nos recherches. L'étude d'un petit nombre de marqueurs lexicalement non explicites potentiels permet d'avancer quelques hypothèses : ils sont très peu nombreux (cf. tableaux 1 et 2) ; les informations qu'ils marquent varient en fonction du type du défini (ex. objet concret, entité abstraite, propriété, processus...) ; certains d'entre eux sont aussi des opérateurs lexicographiques (ex. `propriété`).

Conclusion

Les recherches présentées dans cet article se placent dans le cadre de la constitution d'une BCL à partir du TLF. Nous nous sommes en particulier intéressé à la caractérisation, à l'aide de marqueurs superficiels, du contenu sémantique des segments qui composent les définitions. Ce travail nous a, d'autre part, permis de proposer un schéma d'organisation des informations sémantiques contenues dans les définitions spécialisées. Nous avons également proposé une typologie lexicale des marqueurs, qui, combinée avec leur classification en fonction des informations sémantiques qu'ils signalent, permet d'adapter les méthodes de collecte aux caractéristiques de chaque type.

L'évaluation de la méthode de caractérisation du contenu sémantique des segments est l'une des prochaines étapes de cette recherche. On doit en particulier déterminer la proportion de segments marqués par rapport au nombre total de segments qui pourraient l'être (i.e. les segments complexes). Par ailleurs, ce travail doit être complété par l'étude des définitions spécialisées des autres catégories de définis (noms abstraits, verbes et adjectifs).

Remerciements

Je remercie Christiane Jadelot pour l'aide qu'elle m'a apportée dans les premières phases de cette recherche. Je tiens également à remercier Andrée Borillo pour les discussions que nous avons eues concernant les aspects linguistiques de ce travail, de même que Robert Martin, Pascal Amsili, Josette Lecomte et Patrick Paroubek pour leurs commentaires sur des versions antérieures de cet article. Je remercie les participants au séminaire de DEA des Sciences du langage de l'Université Toulouse Le Mirail pour leurs réactions et leurs remarques. Cet article doit également beaucoup aux discussions qui ont eu lieu lors des réunions du projet ICE « Traitement automatique des textes, terminologie et bases de connaissances » (*Action cognition, communication et ingénierie de la langue*).

Références

Alshawi, H. (1989). Analysing the dictionary definitions. In B. Bougarev & T. Briscoe (Réds.), *Computational Lexicography for Natural Language Processing*, chap. 7, pp. 153–169. London, UK: Longman.

Artola Zubillaga, X. (1993). *Conception et construction d'un système intelligent d'aide dictionnairial (SIAD)*. Thèse de doctorat, Université du Pays Basque, Donostia.

Atkins, B. T. S., & Zampolli, A. (Réds.). (1994). *Computational Approches to the Lexicon.* Oxford, UK: Oxford University Press.

Boons, J.-P., Guillet, A., & Leclère, C. (1976). *La structure des phrases simples : Constructions intransitives*, Vol. 8 of *Langue et Culture.* Genève: Librairie Droz.

Bougarev, B., & Briscoe, T. (Réds.). (1989). *Computational Lexicography for Natural Language Processing.* London, UK: Longman.

Bourigault, D. (1994). *LEXTER, un Logiciel d'EXtraction de TERminologie. Application à l'acquisition de connaissances à partir de textes.* Thèse de doctorat, École des Hautes Études en Sciences Sociales, Paris, France.

Brill, E. (1992). A Simple Rule-Based Part of Speech Tagger. In *Proceedings of the Third Conference on Applied Natural Language Processing*, Trento, Italie: ACL.

Chaurand, J., & Mazière, F. (Réds.). (1990). *La définition.* Paris, France: Larousse.

Chinchor, N., Hirschman, L., & Lewis, D. D. (1993). Evaluating Message Understanding Systems: An Analysis of the Third Message Understanding Conference (MUC-3). *Computational Linguistics*, *19*(3), 409–450.

Constant, P. (1991). *Analyse syntaxique par couches.* Thèse de doctorat, École Nationale Supérieure des Télécommunications, Paris, France.

Dendien, J. (1994). Le projet d'informatisation du TLF. In É. Martin (Réd.), *Les textes et l'informatique*, chap. 3, pp. 31–63. Paris, France: Didier Érudition.

Guillet, A., & Leclère, C. (1992). *La structure des phrases simples : Constructions transitives locatives*, Vol. 26 of *Langue et Culture.* Genève: Librairie Droz.

Lecomte, J., & Paroubek, P. (1994). Premiers essais de l'assignateur de catégories grammaticales d'E. Brill sur des textes français. Rapp. tech., INaLF, Nancy, France.

Martin, R. (1992). *Pour une logique du sens* (Seconde édition). Linguistique nouvelle. Paris, France: Presses Universitaires de France.

Martin, R. (1993). Inférences et définitions lexicographiques. In F. Henry (Réd.), *Pour l'informatisation du* Trésor de la Langue Française *(Rapport Préalable).* Nancy, France: INaLF.

Martin, R. (1994). Dictionnaire informatisé et traitement automatique de la polysémie. In É. Martin (Réd.), *Les textes et l'informatique*, chap. 5, pp. 77–114. Paris, France: Didier Érudition.

Pustejovsky, J. (1995). *The Generative Lexicon.* Cambridge, Mass.: MIT Press.