Repérage automatique de structures linguistiques en corpus : le cas des énoncés définitoires

Josette Rebeyrolle[*] & Ludovic Tanguy[**]

Cet article propose, pour le cas spécifique des énoncés définitoires, une démarche qui s'échelonne depuis une étude linguistique du phénomène jusqu'à la constitution de patrons permettant un repérage automatique des énoncés. Une attention particulière est portée aux technologies d'analyse de corpus, et l'accent est tout particulièrement mis sur les différentes pratiques à l'œuvre dans cette démarche : une pratique linguistique, une pratique des outils de repérage, et une pratique spécifique à l'étude des corpus.

This paper presents the process that leads from a linguistic study on a given type of discourse phenomena (definitions, in our case) to the practical design of morpho-syntactic patterns for their automatic retrieval in corpora. We present and assess the technological aspects of such a study, including an evaluation of the resulting patterns, and we discuss the three different kinds of skill needed in the process: generic linguistic knowledge, use of NLP tools, and corpus linguistics.

[*] ERSS (UMR 5610, CNRS / Université Toulouse II).
[**] ERSS (UMR 5610, CNRS / Université Toulouse II) et Université Toulouse II.

Cahiers de Grammaire 25 (2000), « Sémantique et Corpus », pp. 153-174

1. Introduction

Notre étude porte sur le repérage automatique dans des textes d'énoncés correspondant à des structures linguistiques précises. Au titre de cette problématique, nous nous concentrerons spécifiquement sur les énoncés définitoires, que nous jugeons aptes à exemplifier le cheminement complexe qui mène d'une étude linguistique à une recherche automatique en corpus.

L'étude des énoncés définitoires s'inscrit dans le champ des recherches qui présupposent que les actes de langage laissent des traces dans les textes et que ces traces peuvent être décrites afin de permettre le repérage automatique de ces énoncés. Reprenant à notre compte cette hypothèse, notre objectif vise l'acte de définition et consiste à montrer que les structures qui signalent des énoncés définitoires peuvent être utilisées pour identifier dans les textes les segments définitoires qui y figurent. L'étude des énoncés définitoires se justifie d'abord parce qu'il s'agit d'énoncés qui constituent un lieu privilégié où s'actualisent les relations sémantiques que les mots entretiennent entre eux et ensuite parce qu'ils ont des propriétés suffisamment stables pour que leur repérage automatique puisse être envisagé dans différents discours.

Cette question a fait l'objet de nombreuses recherches à la fois en linguistique et en informatique. Parmi les travaux linguistiques qui se donnent explicitement pour objectif la description de marqueurs lexico-syntaxiques signalant des énoncés riches en informations sémantiques, on peut citer ceux de Borillo (1996), Cartier (1998), Condamines & Rebeyrolle (à paraître), Meyer (à paraître), Pearson (1998), Pascual & Péry-Woodley (1995), Rebeyrolle & Péry-Woodley (1998). D'autres travaux abordent ce problème du point de vue de l'informatique et mettent en oeuvre des méthodes essentiellement fondées sur des bases statistiques. C'est le cas des travaux de Hearst (1998), Morin (1999) et Séguéla (1999).

Les premiers travaux (linguistiques) s'en tiennent à la description des structures qui signalent certains types de relations sémantiques (hyponymie ou méronymie, pour l'essentiel) ou certains types d'énoncés. Mais, si le repérage automatique de ces structures constitue toujours la justification de ces descriptions, son opérationnalisation n'est jamais présentée en tant que telle (à l'exception toutefois des travaux menés par l'équipe de Meyer). *A contrario*, les seconds travaux (informatiques) proposent des outils qui fournissent des accès aux segments textuels contenant des informations sur le sens des mots mais sur des bases uniquement statistiques.

L'objectif de la présente étude est double : il s'agit premièrement de voir comment se manifestent linguistiquement dans les textes les énoncés définitoires (autrement dit quelles sont les structures linguistiques qui expriment ces énoncés) (§2) et deuxièmement d'envisager le problème de l'extraction automatique des segments textuels conformes à ces structures linguistiques (§3).

2. Méthode de repérage

2.1. Exemples de structures à travers lesquelles s'exprime un énoncé définitoire dans le discours

Les formes que revêt la définition en discours ont fait l'objet d'une analyse détaillée (Rebeyrolle 2000) que nous nous contenterons de résumer ici. Nous nous contenterons d'indiquer à l'aide de quelques exemples les formes linguistiques associées aux deux classes d'énoncés définitoires qui ont été isolées : les énoncés définitoires directs et les énoncés définitoires indirects (Riegel 1990). Dans les structures énumérées dans le tableau 1, figurent les trois constituants de chaque structure canonique construite autour d'un verbe de la classe des verbes de désignation (noté Vdésigner), d'un verbe de la classe des verbes de dénomination (noté Vs'appeler) ou d'un verbe de la classe des verbes de signification (noté Vsignifier). On désigne par *SNa* le terme à définir, par le couple *SNx - X*, le syntagme qui sert à le définir. Dès lors que les éléments en relation ne sont pas des syntagmes nominaux, ils sont notés *A* et *B - X*.

Tableau 1 : *Liste des structures linguistiques pour les différentes classes d'énoncés définitoires.*

Classes d'énoncés définitoires	Structures linguistiques
Énoncés définitoires directs :	
- de désignation	[SNa Vdésigner SNx - X]
	[N_0 Vdésigner SNa SNx - X]
- de dénomination	[SNx - X Vs'appeler SNa]
- de signification	[A Vsignifier B - X]
- introduits par *c'est-à-dire*	[A, c'est-à-dire B - X]
	[B - X, c'est-à-dire A]
Énoncés définitoires indirects :	
- de classification	[SNa est un Nx - X]
- parenthétiques	[SNa (SNx - X)]
	[SNx - X (SNa)]

Les exemples fournis ici serviront à illustrer les différents types d'énoncés définitoires qui ont été distingués :

- les énoncés définitoires de désignation :

(1) *On donne le nom de boutonnière, ou bray, aux dépressions allongées, évidées dans les formations peu résistantes des séries sédimentaires ployées en ondulations anticlinales peu marquées.*

- les énoncés de dénomination :

(2) *La vase peu colonisée, recouverte plusieurs heures à chaque marée, se nomme une slikke.*

- les énoncés de signification :

(3) *Conceptualiser un terme signifie représenter chaque notion du terme par un concept dans le formalisme .*

- les énoncés introduits par *c'est-à-dire* :

(4) *Les Unités de Configuration Logiciel. Elles constituent l'élément de plus bas niveau géré en configuration, c'est-à-dire le fichier.*

- les énoncés de classification :

(5) *Un jeu de barres est un circuit triphasé auquel peuvent être raccordés tous les départs (lignes, transformateurs) à une même tension.*

- les énoncés définitoires parenthétiques :

(6) *L'expert devait se remémorer l'incident et disposait des informations contenues dans la main courante (journal de bord où sont notées toutes les interventions sur le réseau de distribution de gaz).*

2.2. Problème du repérage automatique dans des corpus des énoncés définitoires exprimés par ces structures linguistiques

Nous nous appuierons donc sur une description des propriétés linguistiques des structures qui permettent de réaliser un acte de définition en discours (Rebeyrolle 2000). Ce que nous voulons montrer dans cet article, c'est comment les outils peuvent être employés pour représenter ces structures linguistiques afin de repérer automatiquement dans un texte les énoncés définitoires qu'il contient. L'essentiel, pour nous, est donc d'évaluer la possibilité du repérage automatique des structures définitoires énumérées précédemment.

Etant donné qu'il n'est pas possible d'extraire efficacement toutes les structures à l'aide de l'outil (Yakwa, cf. § 2.3.2) dont nous disposons, celui-ci étant limité à une analyse de surface, la distinction entre les énoncés définitoires de désignation et de dénomination fondée sur la place qu'occupe la dénomination, *Na*, dans la structure thématique de ces énoncés (position initiale ou médiane pour les énoncés de désignation et position finale pour les énoncés de dénomination) a été abandonnée. Cette distinction est difficile à faire du point de vue du repérage automatique étant donné que pour établir

les patrons, on s'appuie principalement sur le pivot verbal et que les verbes qui composent les structures linguistiques de dénomination et de désignation sont généralement proches, comme le rappellent les exemples ci-dessous où *(7)* illustre la désignation et *(8)* la dénomination :

(7) *On appelle librairie un magasin où l'on vend des livres.*
(8) *Le magasin où l'on vend les livres s'appelle librairie.*

De plus, les énoncés parenthétiques, ainsi que ceux introduits par *c'est-à-dire*, caractérisés par l'absence d'une structure verbale spécifique, n'ont pas fait l'objet d'une analyse automatique[1].

Ainsi, les structures linguistiques qui ont servi à la description des patrons ont été largement réduites. Pour l'évaluation, nous avons distingué :

- pour les énoncés définitoires directs : les énoncés définitoires contenant les verbes qui marquent une relation de désignation et les verbes qui établissent une relation de dénomination sont réunis dans une même classe (classe 1) et les énoncés définitoires contenant un verbe de signification forment une deuxième classe (classe 2) ;
- pour les énoncés définitoires indirects, on s'en est tenu aux énoncés de classification (classe 3).

On va voir que toutes les propriétés des structures définitoires ne se retrouvent pas dans les patrons qui servent au repérage automatique. On soulignera, en particulier, l'ajout de contraintes qui n'ont pas été mentionnées dans les structures linguistiques définitoires et dont la présence se justifie par le changement de point de vue que nous opérons ici en passant d'une analyse linguistique descriptive à une recherche automatique basée sur des indices de surface qui doit nécessairement tenir compte de la linéarité du texte. On essaiera donc d'expliquer pourquoi la forme des patrons ne correspond pas strictement aux structures linguistiques décrites.

Dans le cadre de cet article, nous nous contenterons d'illustrer le passage de la structure linguistique à un patron utilisable automatiquement sur quelques structures définitoires : nous examinerons successivement les patrons des énoncés définitoires contenant un verbe du type *définir* (§3.1), les patrons des énoncés définitoires de signification (§3.2.1) pour finir par les patrons des énoncés définitoires de classification de forme *Na est un Nx - X* (§3.2.2).

[1] Nous verrons, entre autres, que la gestion des verbes lors du repérage automatique constitue l'une des principales difficultés.

2.2.1. Le corpus

Les énoncés définitoires étudiés ont été extraits d'un corpus d'analyse composé de quatre types de documents :

- un manuel de géomorphologie[2] : 275 000 mots.
- 34 articles[3] scientifiques appartenant au domaine de l'ingénierie des connaissances : 230 000 mots.
- deux documents[4] mis au point et utilisés au sein une entreprise française (EDF) : 205 000 mots.
- un ensemble d'articles extraits de l'Encyclopædia Universalis : 215 000 mots.

L'étude a donc été conduite sur un corpus qui réunit des textes de genres et domaines distincts, ce qui garantit une certaine généralité aux structures étudiées. Pour une étude plus précise des variations des énoncés définitoires à travers les genres textuels, cf (Péry-Woodley et Rebeyrolle 1998).

2.2.2. Liste de référence

Sur la base de notre compétence linguistique, nous avons dressé manuellement une liste exhaustive des énoncés définitoires contenus dans chacun des textes du corpus. Cette liste qui rassemble au total 1 574 énoncés définitoires est indispensable pour faire les mesures statistiques permettant d'évaluer l'efficacité[5] des structures linguistiques pour le repérage automatique des segments textuels où un énoncé définitoire est exprimé. Cette évaluation repose sur le principe suivant : il s'agit de mesurer la

2 M. Derruau (1988) Précis de géomorphologie, Masson, 7ème Edition.
Nous remercions D. Candel (responsable de LNST, FRE 2173, ex INALF, CNRS) d'avoir mis à notre disposition cet ouvrage extrait de la base SCITECH.

3 Ces articles ont fait l'objet d'une publication dans un ouvrage édité par J. Charlet, M. Zacklad, G. Kassel & D. Bourigault aux éditions Eyrolles (2000) sous le titre "Ingénierie des connaissances, évolutions récentes et nouveaux défis".
Nous remercions les membres du groupe Terminologie et Intelligence Artificielle (TIA) pour nous avoir permis d'utiliser un recueil d'articles constitué dans le cadre d'un projet mené avec la Délégation Générale à la Langue Française (DGLF) intitulé "Construction d'un thesaurus en français sur les données de l'ingénierie linguistique et de l'ingénierie des connaissances pour la recherche d'information sur la toile".

4 Il s'agit de deux guides. Le premier est consacré au développement de projets en Génie Logiciel, le second s'attache à préciser l'organisation des réseaux d'alimentation générale en énergie électrique.

5 Précisons que l'évaluation que nous nous proposons de faire ici s'applique uniquement aux structures linguistiques que nous considérons comme définitoires dans ces corpus (et qui ont fait l'objet d'une description détaillée).

pertinence des énoncés repérés automatiquement dans le corpus à partir des structures linguistiques en les confrontant aux énoncés réellement définitoires.

2.3. Aspects techniques

Nous présenterons ici le cadre technique de notre étude. Comme précisé plus haut, les patrons élaborés pour le repérage des énoncés définitoires se basent sur des informations morphosyntaxiques, et sont donc destinés à s'appliquer à des corpus enrichis ou étiquetés. L'utilisation de tels corpus nécessite donc des outils spécifiques, capables, sur le principe des concordanciers, d'effectuer des recherches en se basant sur des unités lexicales lemmatisées et des catégories morphosyntaxiques.

2.3.1. Corpus annotés

Tant dans le domaine de l'ingénierie des connaissances que de la linguistique informatique, l'utilisation de corpus annotés et d'outils d'annotation est maintenant monnaie courante. Des méthodes automatisées robustes, capables d'attribuer à chaque unité d'un texte une forme canonique (lemme) et une indication de sa catégorie morphosyntaxique (étiquette), développées initialement pour la langue anglaise, sont depuis quelques années disponibles pour le français. Il est donc logique que le développement d'applications en tout genre travaillant sur des textes prennent de telles informations comme point d'entrée de leur traitement, depuis la recherche d'information et l'extraction terminologique, jusqu'au résumé automatique. Un taux de succès dépassant les 95% sur un texte quelconque (Adda et al. 1999), combiné à la possibilité d'étiqueter des corpus de très grande taille donne d'ailleurs à ces méthodes un avantage net sur des méthodes d'analyses plus profondes, comme l'étiquetage syntaxique complet, dont le manque de robustesse nuit à son application sur de gros corpus.

Dans le cas de notre étude, les quatre textes constituant notre corpus de travail ont été automatiquement annotés à l'aide de l'outil Cordial Universités[6], sans que nous ayons apporté la moindre correction. Nous verrons par la suite que les erreurs d'étiquetage résiduelles ne nuisent pas, quantitativement, à la qualité des résultats, du moins au regard du gain que représentent les informations morphosyntaxiques.

Précisons enfin que le jeu d'étiquettes utilisé par nos patrons de recherche correspond au maximum d'information fourni par l'analyseur, c'est-à-dire celui proposé par l'action Grace pour la comparaison de différents étiqueteurs. Ainsi, nous disposons, par exemple, pour un verbe conjugué, de ses temps, mode, personne et nombre, et pour un pronom personnel, de ses personne, nombre, genre et cas. Les lemmes correspondent, comme dans la

[6] Distribué par la société Synapse Développement http://www.synapse-fr.com

totalité de ces approches, à l'infinitif pour les verbes, au singulier pour les noms, et au singulier masculin pour les adjectifs.

2.3.2. Concordanciers sur corpus annotés

Alors que les étiqueteurs sont maintenant très répandus, il peut être paradoxal de constater que des outils génériques, capables d'interroger de tels corpus annotés et d'en extraire des segments, sont eux beaucoup plus rares. Chaque application construite pour manipuler et exploiter de telles informations dispose de ses propres méthodes pour parcourir le texte, mais les fonctionnalités varient alors suivant les objectifs suivis. Le genre de travail que nous effectuons dans cette étude se base sur un outil générique d'interrogation de corpus étiquetés, qui reprend le principe des concordanciers travaillant sur des textes nus. Il s'agit de l'outil Yakwa[7], une plate-forme d'interrogation de textes sur la base de patrons morphosyntaxiques, fonctionnant sur le même principe que l'interface XKWIC du projet CQS (Christ 1994).

Alors qu'un concordancier « simple » permet à son utilisateur de définir ses critères de recherche dans un texte sur la seule base des formes de surface (en utilisant toutefois des troncatures du type `défini$` pour représenter tout mot commençant par « défini »), un concordancier sur corpus annoté comme Yakwa permet d'utiliser les informations liées aux catégories morphosyntaxiques et aux formes canoniques des unités lexicales du texte. Il est également possible de définir des patrons, c'est-à-dire des séquences de marqueurs individuels, chacun de ces marqueurs correspondant à un mot du texte. Ainsi, une séquence comme :

```
« prendre » Article t$
```

correspond à toute séquence composée d'une forme conjuguée ou non du verbe *prendre*, suivie immédiatement d'un article quelconque, suivi de tout mot commençant par la lettre *t*.

Il est, de plus, possible d'utiliser des contraintes plus lâches, en autorisant un certain nombre d'unités à s'insérer entre deux marqueurs individuels, voire à autoriser des unités facultatives d'un certain type entre deux marqueurs

```
« prendre » Article {Adjectif}2 t$
```

Ce patron reprend le précédent en autorisant un maximum de deux adjectifs après l'article.

[7] Pour plus de détails sur Yakwa, qui était, à l'origine, le module de profilage de textes de la plate-forme DiET (Tanguy et al. 1999), nous invitons le lecteur à se référer au site WWW suivant : http://www.univ-tlse2.fr/erss/membres/tanguy/yakwa.html.

Le tableau 2 fournit un résumé des notations utilisées ici pour décrire les patrons morphosyntaxiques :

Tableau 2 : Liste des éléments composant les patrons morphosyntaxiques.

Catégories morphosyntaxiques		Expressions régulières	
Adv	Adverbe	**NON X**	Exclusion de X
Det	Déterminant	**(X|Y)**	Disjonction
Num	Adjectif numéral	**{X}n**	De 0 à n mots vérifiant X
Prép	Préposition	**{X}***	0 ou plus mots vérifiant X
Pro	Pronom	**n**	de 0 à n mots quelconques
ProPers	Pronom Personnel	*****	0 ou plus mots quelconques
Vbe	Verbe	**« X»**	Forme lemmatisée de X

D'autres outils du même type restent disponibles, avec des nuances : le logiciel SATO (Daoust 1996), fonctionne sur le même principe, mais sans que les unités du texte soient désambiguïsées. Ainsi, la référence à une catégorie grammaticale dans le patron de recherche, comme Verbe, correspondra dans le texte à toute forme dont la graphie peut correspondre à un verbe (les formes *porte* ou *ferme* seront extraites même lorsqu'elles sont employées dans le texte comme nom ou adjectif). Le moteur d'interrogation de la partie catégorisée de la base Frantext[8] permet la définition de grammaires d'interrogation atteignant un pouvoir d'expression similaire, mais reste bien entendu limité au corpus de Frantext. Enfin, le système Intex (Silberztein 1993) permet également la définition de patrons de recherche sous la forme de grammaires locales représentées par des automates à état fini.

La puissance d'expression disponible pour la définition de patrons correspond *a priori* au maximum de ce qu'il est possible d'exiger en se limitant à des informations morphosyntaxiques. Nous verrons, dans le détail de l'étude que nous avons menée à l'aide de cet outil, à la fois l'intérêt de cette technologie, tout comme le problème posé par le passage d'un schéma linguistique à un tel patron morphosyntaxique.

3. Résultats

Au vu de la multiplicité des structures établies lors de la première partie, nous avons sélectionné ici une structure linguistique de la première classe (classe 1 : énoncés définitoires directs de désignation et de dénomination), à savoir la structure qui rassemble les constructions du verbe *définir*. C'est sur ce schéma que nous expliciterons plus en détail les différentes étapes qui conduisent à l'élaboration d'un patron complexe dont les performances en corpus sont satisfaisantes.

8 http://zeus.inalf.fr/

Par la suite, nous résumerons plus succinctement d'autres types de structures, en précisant les scores atteints et les patrons employés. Nous pourrons ensuite discuter, au vu de l'ensemble de ces structures, de la disparité des résultats et des gains relatifs des différentes étapes pour chaque patron.

3.1. Traitement des structures linguistiques contenant le verbe *définir*

Les trois constructions du verbe *définir* doivent être accessibles par le patron de repérage automatique. Il s'agit de :

- la construction agentive : `[N0 définit Na comme Nx – X]`

(9) *Nous définissons le système d'information comme l'ensemble des moyens de traduction et d'utilisation de connaissances.*

- la construction passive : `[Na est défini comme Nx – X]`

(10) *La période d'actualisation est définie comme la période de temps sur laquelle s'échelonnent les dépenses dont on veut connaître le total actualisé.*

- la construction pronominale passive : `[Na se définit comme Nx – X]`

(11) *Une relation se définit comme un sous-ensemble du produit cartésien des extensions des concepts composant sa signature.*

Les énoncés correspondant à ces trois constructions sont réunis dans une même sous-liste de référence et doivent donc être accessibles par un même patron. C'est la construction de ce patron que nous allons décrire à présent. Nous fournirons le détail de chacune des étapes de cette construction, en partant d'un patron correspondant à un outil basique d'exploration de texte non annoté (réduit à la forme verbale), jusqu'à un patron mettant en jeu les fonctionnalités avancées de Yakwa permettant de se rapprocher des constructions du verbe *définir*. A chaque étape, nous indiquerons les valeurs de *rappel* et de *précision*[9] et formulerons des commentaires sur la nature du patron.

[9] Le *rappel* est le pourcentage d'énoncés définitoires qui sont retrouvés par le patron. La *précision* est le pourcentage d'énoncés retrouvés par le patron qui sont des énoncés définitoires. Pour plus de détails sur ces mesures, nous renvoyons à Baeza-Yates & Ribiero-Neto (1999).

3.1.1. Description détaillée des patrons

Les énoncés définitoires qui comportent le verbe *définir* couvrent un ensemble assez cohérent de contextes aisément repérables puisqu'ils restent centrés autour d'une même unité lexicale.

Patron 1 : `défini$`

Rappel : 100% - *Précision* : 5,15%

Commençons par observer les résultats de ce patron extrêmement basique, susceptible d'être utilisé par n'importe quel outil de recherche en plein texte (qu'un simple traitement de texte est susceptible de fournir). Le premier problème de ce patron est que la simple troncature ne suffit pas à distinguer les formes verbales de "définir" du substantif (*définition*) et des adjectifs ou adverbes (*définitif*, *définitivement*), ce qui explique partiellement le bruit produit. En effet, un taux de précision de 5% indique que seul un énoncé repéré sur vingt est un énoncé définitoire.

Patron 2 : « `définir` »

Rappel : 100% - *Précision* : 7,24%

Pour résoudre cet inconvénient, nous utilisons ici la lemmatisation du corpus pour ne retenir que les formes verbales. Nous noterons un tel marqueur élémentaire « `définir` », les guillemets indiquant que l'on déclare la forme canonique d'une unité recherchée. Une alternative à l'utilisation d'un corpus lemmatisé consiste à dresser un inventaire exhaustif des formes verbales envisagées, qui sont, dans ce cas, en nombre fini. Ainsi, ce patron est équivalent à une disjonction de ces formes fléchies (*définit* OU *définis* OU *définissent* OU *définissons*, etc.). Les intérêts de la lemmatisation sont donc ici avant tout la simplicité et la systématicité.

Patron 3 : « `définir` » `* comme`

Rappel : 100% - *Précision* : 73,33%

L'étape suivante correspond à une première extension syntagmatique du patron. Il s'agit ainsi de prendre en compte la seconde unité lexicale du schéma général, *comme*, à l'aide d'outils très généraux, travaillant ici encore sur texte nu. Un patron de ce type peut, sur la base du précédent, exprimer la présence d'une occurrence du verbe *définir* suivie, dans la même phrase, ou dans un contexte droit délimité, d'une occurrence de *comme*. Le patron 3 revient, dans un premier temps, à rechercher une occurrence de *comme* à droite de *définir*, dans la seule limite de la phrase, comme l'indique le signe « * » séparant les deux marqueurs.

Patron 4 : « définir » 6 comme

Rappel : 90,91% - *Précision* : 92,59%

Le patron 4 se veut une première tentative de réduction du bruit résiduel, en bornant la distance, en nombre d'occurrences, entre les deux items lexicaux. Nous avons proposé la valeur 6, valeur moyenne correspondant au meilleur score entre 2 et 10. La valeur 6 est en effet considérée comme suffisante pour couvrir les cas d'insertion d'une apposition ou d'une locution adverbiale qui forment, généralement, des séquences d'au plus 6 unités. Cette valeur n'est toutefois pas suffisante pour des termes complexes, très courants en langue spécialisée, comme *facteur d'atténuation entre deux valeurs non consécutive*s. Cette restriction brute laisse également sous silence des compléments circonstanciels comme *dans ce type de discours*, dont la taille vient se rajouter à celle du terme défini.

Il est aisé de voir que le taux de *rappel* chute de façon inacceptable, même si la *précision* gagne quant à elle 20%. La méthode utilisée au stade suivant consiste alors en l'utilisation d'un schéma récursif, mettant en oeuvre une fonctionnalité avancée de l'outil Yakwa. Il s'agit de ne pas borner quantitativement la distance entre les deux marqueurs, mais au contraire d'interdire certaines catégories morphosyntaxiques, les verbes dans notre cas.

Patron 5 : « définir » (Non Vbe)* comme

Rappel : 100% - *Précision* : 91,67%

Ainsi, le patron 5 se lit de la manière suivante : on recherche une occurrence du schéma « définir » * comme, pour lequel aucune forme verbale n'est présente entre *définir* et *comme*. Cette méthode permet, de façon souple, de circonscrire les syntagmes nominaux, ainsi que la plupart des compléments circonstanciels.

Patron 6 :
« définir » (Non Vbe)* comme (Non Adv,ProPers,Prép)

Rappel : 98,18% - *Précision* : 96,53%

La dernière étape consiste, afin d'améliorer encore les résultats, à filtrer les dernières phrases non pertinentes en interdisant, pour l'unité située immédiatement à droite de *comme*, les adverbes, pronoms personnels et prépositions. Cette contrainte permet d'évacuer de l'ensemble des énoncés correspondant à ces patrons les utilisations de syntagmes adverbiaux introduits par *comme* (*comme dans la plupart des cas*, *comme précédemment*) ou les subordonnées incises (*comme on l'a vu à la section 2*).

Les scores de *rappel* et de *précision* pour chacun des patrons précédemment décrits sont résumés dans la figure 1. Nous y avons ajouté une troisième mesure, la *F-mesure*, définie par la formule suivante (Baez-Yates & Ribiero-Neto 1999) :

```
F-mesure=2 x rappel x précision / (rappel + précision)
```

Cette dernière mesure permet de résumer la performance d'un patron sous un seul score. Elle permet notamment d'expliquer que la légère perte de *rappel* du patron 6 est compensée par le gain en *précision*.

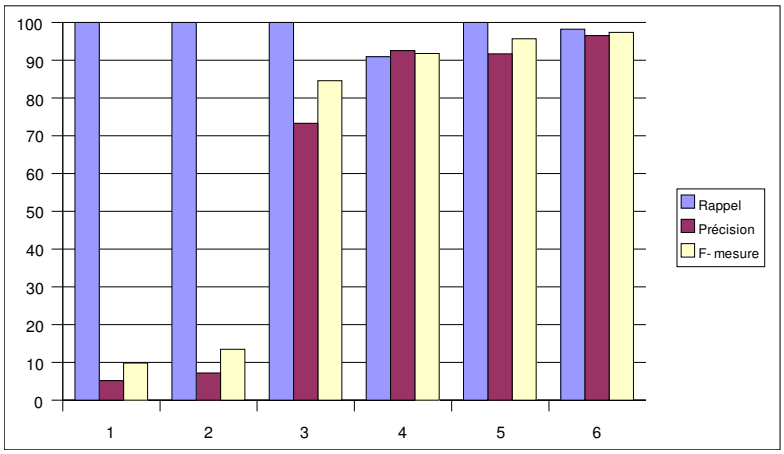

***Figure 1** : Scores des patrons des structures contenant le verbe définir*

3.1.2. Discussion

Pour synthétiser la progression motivant ces six patrons, nous devons prendre en compte plusieurs considérations.

La technologie : Dans chaque cas, sauf pour le passage de 5 à 6, une nouvelle fonctionnalité est mise en jeu. On peut voir que le principal gain est celui de pouvoir rechercher des schémas extensibles, entre deux unités lexicales (patron 3), mais aussi le schéma récursif du patron 5, qui est d'autant plus précieux qu'il permet une progression dans la fameuse zone des « plus de 90% », qui, suivant la loi de Zipf, est celle où les progrès sont les plus difficiles.

La motivation par la structure linguistique : Chaque passage qui consiste en un enrichissement du patron correspond bien sûr à une élaboration en vue de rejoindre les structures linguistiques énumérées en 3.1, pour peu que la technologie utilisée le permette. Toutefois, cette impulsion ne s'étend qu'au patron 3 dans notre progression. Si nous envisageons de développer un patron introduisant les syntagmes nominaux *SNa* et/ou *SNx* comme les deux patrons suivants, on observe une chute dramatique du taux de rappel.

Patron 7 : `Nom (Non Vbe)* « se | être » « défini`

```
(Non Vbe)* comme (Non Adv,Pro,Prép)
```

Patron 7 bis :
```
Nom|Pro (Non Vbe)* « définir »
(Non Vbe)*Nom (Non Vbe)* comme
(Non Adv,Pro,Prép)
```

Ces deux patrons sont ce que l'on peut envisager comme étape suivante afin de se rapprocher au plus près des structures linguistiques. Le premier cherche à capter les constructions passives et pronominales passives, et le second les constructions agentives du verbe *définir*. Les résultats pour ces deux patrons (i.e. on prend comme résultat l'union des énoncés produits par chacun des patrons) sont les suivants :

Rappel : 40% - *Précision :* 97%

On obtient, en d'autres termes, le meilleur taux de *précision*, mais le taux de *rappel* est inacceptable. La raison de cette chute est la complexité des réalisations à prendre en compte. Que ce soit la présence d'auxiliaires modaux (*Na peut se définir comme*), les appositions (*Na, défini comme*) les constructions infinitives (*il est utile de définir Na comme*), la multiplicité des énoncés « hors-norme » rendent la tâche d'énumération des patrons de surface très fastidieuse, voire inaccessible.

La raison de cet état de fait s'explique de la manière suivante. Nous n'utilisons pour appliquer ces patrons que des informations morphosyntaxiques de surface, et non une véritable analyse syntaxique. Dès lors, les perturbations observées sur l'axe syntagmatique sont très importantes, trop pour être captées par des schémas pointus comme ceux tentés par les patrons 7.

Le principe à appliquer dans ce cas semble être celui d'une description minimale, quitte à rester bien souvent très éloigné de la structure linguistique. Cette façon de procéder est à mettre en parallèle avec la méthode d'extraction de termes proposée par D. Bourigault pour l'outil LEXTER (Bourigault 1996). En effet, dans le cas de Lexter, les patrons morphosyntaxiques utilisés pour le repérage de syntagmes nominaux complexes sont définis *négativement*, *i.e.* en indiquant quelles sont les unités qui en forment les frontières (comme les verbes, les pronoms, etc.), et non par l'explicitation exhaustive des schémas des syntagmes eux-mêmes.

Cette remarque nous conduit ainsi à identifier un troisième axe pour décrire l'évolution des patrons :

La pratique du corpus : Dans le cas des dernières progressions (patrons 5 et 6), il est envisageable, mais plus difficile, d'attribuer aux contraintes supplémentaires un statut de résultat d'une réflexion linguistique traditionnelle. Que ce soit l'interdiction des verbes entre *définir* et *comme*, ou l'interdiction des adverbes et prépositions à droite de *comme*, ces contraintes sont plus facilement justifiables par une pratique directe des patrons, et notamment par l'observation des énoncés non pertinents, que par une analyse *a priori* hors corpus (ce qui n'est pas le cas du schéma central *définir comme*,

bien entendu).

Concrètement, cette phase d'affinage des patrons se fait par observation directe des énoncés non pertinents produits par les patrons précédents et conduit donc naturellement à l'interdiction de certains éléments.

Cette approche empirique de la définition des patrons, en prenant en compte progressivement les contraintes et leurs résultats nous a conduit à dégager un certain nombre de principes opératoires dans l'élaboration d'un patron pour un schéma linguistique quelconque :

- préférer un sous-schéma récursif à une limite fixe entre deux pivots du patron
- rester indépendant vis-à-vis des constructions actives/passives, i.e. ne pas préjuger des formes verbales précises
- se contenter, pour l'ajustement fin du patron, d'observer (et d'interdire) les catégories des unités situées à proximité immédiate de la séquence repérée.

3.2. Traitement de deux autres structures linguistiques définitoires

Nous présenterons ici les résultats synthétiques concernant deux gammes de patrons correspondant à des structures de la classe 2 (énoncés définitoires de signification) et de la classe 3 (énoncés définitoires de classification).

3.2.1. Description des patrons pour la structure [A Vsignifier B - X]

Les énoncés définitoires de signification recouvrent des structures construites autour des verbes *signifier*, *vouloir dire* et *entendre par*. En voici des exemples :

(12) *Floculer signifie former des flocons, des grumeaux, qui peuvent aller jusqu'à coaguler ensemble.*
(13) *Le terme de régolite, qui veut dire étymologiquement roche fragmentée (du grec regnumi, je fragmente) devrait être synonyme d'altérite.*
(14) *Nous entendons par primitives conceptuelles une expression du domaine allant du mot à un groupe de mots désignant une notion du domaine que nous devons représenter dans la taxonomie.*
(15) *Par sol, on entend non seulement le sol pédologique, mais les dépôts meubles superficiels, héritages altérés sur lesquels le sol pédologique se développe et dont l'ablation est préjudiciable.*

Étant donné la multiplicité des structures, notre façon de procéder pour ces énoncés est de proposer un ensemble de patrons, et non plus un patron unique. A chacune des étapes décrites ci-dessous, c'est l'ensemble des énoncés captés par l'un des sous-patrons proposés qui est évalué. Ces patrons travaillent ainsi de concert.

Patron 1 :

```
« signifier »
« vouloir » « dire »
« entendre »
```

Rappel : 100% - *Précision* 41,94%

Ce premier patron se contente d'utiliser la lemmatisation du corpus. Le faible nombre de contraintes a pour effet de produire un taux de *rappel* maximal et en contrepartie un pourcentage de *précision* faible.

Patron 2 :

```
« signifier »
« vouloir » « dire »
« entendre » * par
par * « entendre »
```

Rappel : 100% - *Précision* 48,60%

La prise en compte, dans ce deuxième patron, de la préposition qu'exige le verbe *entendre* dans l'acception retenue ici ne produit pas une significative progression du pourcentage de *précision*. C'est avec le patron 3 que l'on obtient un résultat significatif.

Patron 3 :

```
(Non Pro (sauf ProRel)) « signifier » (Non « que »)
« vouloir » « dire » (Non « que »)
« entendre » (Non Vbe)* par
par (Non Vbe (sauf modaux))* « entendre »
```

Rappel : 98,08% - *Précision :* 70,83%

Ce patron interdit, à gauche du verbe *signifier*, la présence d'un pronom autre qu'un pronom relatif et, à droite de ce verbe, comme de *vouloir dire*, la conjonction *que*. Les cas où le complément du verbe *signifier* est une conjonctive sont exclus car, pour marquer la définition d'un mot, le sujet des verbes *signifier* ou *vouloir dire* doit être un signe et non un événement. C'est ce que souligne C. Wimmer (1987, p. 63), s'agissant de *signifier*, quand elle affirme que : "*que* introduit à des fins de construction une phrase, c'est-à-dire l'évocation d'une réalité événementielle". Pour le verbe *entendre*, on exclut la présence d'un verbe entre *entendre* et la préposition *par*, en n'excluant toutefois pas les énoncés utilisant un auxiliaire modal du type : *par A, il faut entendre...*

Ce dernier patron allie donc les propriétés linguistiques des constructions verbales des verbes *signifier* et *vouloir dire* à des informations observées en corpus. Ces diverses contraintes complètent la description du patron afin de restreindre le nombre d'énoncés extraits (et donc de limiter le *bruit*).

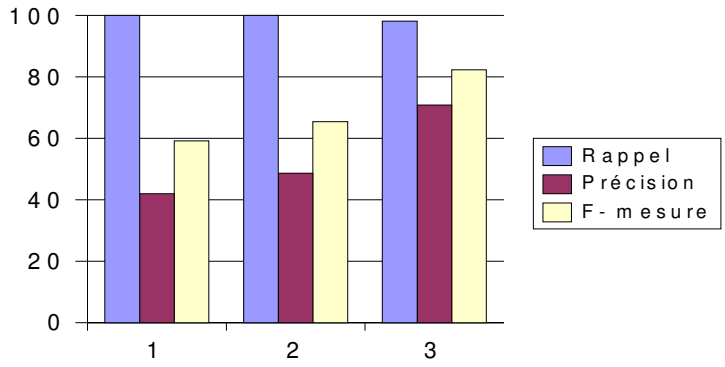

Figure 2 : *Scores des patrons pour les verbes de signification*

3.2.2. Description des patrons pour la structure [SNa est un Nx - X]

Le repérage automatique des énoncés définitoires de classification de structure [SNa est un Nx - X] peut être conduit au moyen des patrons suivants :

Patron 1 : `est|sont un|une|le|la|les|l'|des`

Rappel : 93,53 % - *Précision* : 16,83%

Ce premier patron, qui suit au plus près la réalisation discursive de la structure, ne fournit pas de résultats satisfaisants en ce qui concerne le rappel.

Patron 2 : `«être» * Det|Num`

Rappel : 99,73% - *Précision* : 3,94%

La généralisation que permet le recours à la lemmatisation et à la catégorisation dans ce second patron rend possible la prise en compte de toutes les occurrences considérées comme pertinentes (le *rappel* atteignant presque les 100%). Mais elle conduit à une augmentation très importante du bruit. Il est donc crucial de préciser ce patron en ajoutant des contraintes supplémentaires, comme nous proposons de le faire dans le patron 3.

Patron 3 :

`«être» (Adv)1 Det|Num 2 (Nom sauf L)`

Rappel : 94,90% - *Précision* : 21,70

Résultats :

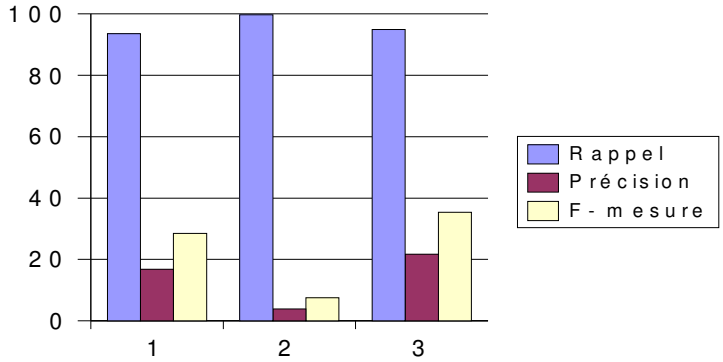

Figure 3 *: Scores des patrons pour les énoncés de classification*

Dans ce troisième patron, il faut noter en particulier que certains items lexicaux ont été interdits à droite du déterminant de *Nx* comme l'indique le marqueur (Nom sauf L), où L est une liste de substantifs interdits dans cette position. Cette contrainte permet d'ignorer certaines formes ayant un statut de déterminant, comme *le cas de, un exemple de*, etc. Toutefois nous devons nous poser le problème de l'établissement de cette liste, qui nécessiterait à la fois une étude systématique et une adaptation au corpus.

Si le pourcentage de *rappel* (94,90%) montre que la quasi totalité des énoncés (contenus dans la sous-liste de référence des énoncés de classification) sont automatiquement repérés, le pourcentage de *précision* (21,70%) fait, en revanche, apparaître la difficulté d'un repérage automatique des énoncés définitoires de cette classe. Les patrons construits sur le schéma *être un* produisent en effet un très grand nombre d'énoncés qui ne peuvent recevoir une interprétation définitoire.

Le repérage automatique des énoncés de structure [SNa est un Nx - X] pose au moins deux types de difficultés :

- cette structure n'est pas, on le sait, exclusivement réservée à l'assertion d'une relation d'inclusion hiérarchique - nombreux sont ceux à avoir signalé que la relation d'hypo/hyperonymie permet d'expliciter toute sorte de classification.
- à cela s'ajoute surtout la difficulté strictement technique de prendre en considération toutes les variations qui affectent la *mise en texte* de cette structure générale. La description du patron oscille, en effet, entre trop de rigueur et trop de laxisme. Si le patron est trop précis, le *silence* produit est très important, si, au contraire, le patron est trop lâche, autrement dit s'il prévoit trop de distance entre les différents éléments qui le composent, le *bruit* croît alors très vite puisque la structure recherchée se confond donc avec des structures linguistiques sans aucun rapport.

La description proposée dans le patron 3 se situe à mi chemin entre ces deux extrêmes mais ne parvient pas à fournir un résultat significatif.

Pour conclure, il faut souligner les difficultés qui se posent lorsque l'on veut utiliser le contexte discursif dans une recherche automatique du type de celle que nous venons de présenter. Nous avons en effet montré (Rebeyrolle 2000) combien les situations dans lesquelles le terme à définir, *Na*, fait l'objet d'une première mention, généralement immédiate, dans le contexte gauche sont favorables pour marquer le passage du discours référentiel au discours définitoire :

(16) *Une section se reconnaît par le fait qu'elle est composée de blocs de texte (les paragraphes) situés sous un titre. Le titre est un bloc court isolé et typographié dans une fonte plus grasse, éventuellement souligné et numéroté.*

Il faut cependant se poser la question du statut que l'on doit donner à cette contrainte discursive. Dans la mesure où elle n'est pas systématique, cette contrainte n'intervient pas au même niveau que celles que nous avons utilisées jusqu'ici. Elle se situe, selon nous, à un second niveau, et constitue un indice supplémentaire qui peut être employé pour ordonner les énoncés extraits par les patrons. Son utilisation expérimentale, en faisant intervenir d'autres fonctionnalités de Yakwa comme la mémorisation en cours de recherche, donne un taux de précision dépassant les 40% pour un rappel inférieur à 30%. On peut utiliser cette contrainte pour attribuer aux énoncés qui s'y conforment un "indice de confiance" plus élevé que celui que l'on donne à ceux dans lesquels *Na* ne fait pas l'objet d'une reprise.

Les situations de reprise du terme à définir, si elles ne sont pas absolument systématiques, apportent donc un poids supplémentaire aux patrons définitoires de classification et nous invitent à envisager la possibilité d'une pondération différentielle des énoncés extraits par ces patrons en attribuant un indice de confiance plus élevé à ceux qui remplissent cette contrainte.

4. Conclusion

Nous avons présenté ici une expérience décrivant une démarche complète de repérage automatique en corpus d'un type d'énoncés dont le fonctionnement linguistique a été précisément décrit. Ce passage de la description linguistique à une représentation opérationnelle résulte d'un compromis entre :

a) la description des propriétés linguistiques des structures étudiées,
b) la technologie influençant le mode d'expression des structures,
c) la pratique empirique de la recherche en corpus.

Si ces trois facteurs interagissent tout au long de la démarche, il n'en est pas moins possible d'attribuer à chacun une période privilégiée. Ainsi, le point de départ reste bien entendu la description linguistique des structures.

Les premières étapes concernent ensuite l'utilisation des possibilités techniques. Et c'est particulièrement lors des phases finales d'affinage des patrons que le corpus prévaut.

Si la description linguistique permet de mettre au jour un faisceau de contraintes et de facteurs qui conditionnent l'insertion textuelle des énoncés définitoires, elle ne peut cependant pas rendre compte des variations qui affectent leur environnement syntagmatique. Or, une recherche automatique basée sur des indices de surface est tributaire de la linéarité du texte. Ainsi distinguons-nous deux types de contraintes :

- Les contraintes que l'analyse linguistique a mises en lumière sont des contraintes que l'on pourrait qualifier de *positives*, étant donné qu'elles se présentent sous la forme d'une description des éléments morpho-syntaxiques qui composent les structures étudiées ;
- Les contraintes de surface que l'observation des occurrences en corpus révèle se présentent quant à elles sous la forme d'un ensemble d'éléments qu'il s'agit d'exclure, aussi parlerons-nous de contraintes *négatives*.

Nous avons pu dégager à cette occasion les limites d'une approche qui consisterait à envisager au préalable ces variations, et qu'une plus grande efficacité est atteinte en raisonnant négativement, i.e. en excluant certaines occurrences plutôt qu'en énumérant celles qui sont acceptables.

Les fonctionnalités avancées d'un outil comme Yakwa ont pu être évaluées quantitativement. La progression de nos patrons est, nous l'avons vu, rythmée par la possibilité d'exprimer des contraintes de plus en plus complexes. Notamment, l'utilisation de corpus annotés constitue un avantage évident par rapport aux textes bruts.

Enfin, nous avons également démontré les limites d'une approche fondée sur des critères de surface, pour le repérage de structures ayant une forte composante sémantique, comme c'est le cas des énoncés définitoires de classification. Même si pour ce type de phénomènes une analyse syntaxique complète serait de peu de secours, il est probable qu'un enrichissement de l'analyse de surface améliorerait notablement les résultats dans le cas général. La mise à disposition de méthodes robustes d'analyse syntaxique partielle, capables d'aborder de larges corpus (voir Bourigault et Fabre dans ce volume), est donc une piste à suivre.

Références bibliographiques

Adda, G., Mariani, J., Paroubek, P., Rajman, M. & Lecomte, J. (1999), "L'action GRACE d'évaluation de l'assignation des parties du discours pour le français", in *Langues* 2-2, pp. 119-129.

Baeza-Yates, R. & Ribeiro-Neto, B. (1999), *Modern Information Retrieval*, ACM Press Books, Massachusetts, Addison-Wesley.

Borillo, A. (1996), "Exploration automatisée de textes de spécialité : repérage et identification de la relation lexicale d'hyperonymie", in *LINX* 34-35, pp. 113-124.

Bourigault, D., Gonzalez-Mullier, I. & Gros, C. (1996), "LEXTER, a Natural Language Processing tool for terminology extraction", in Actes *Seventh International Congress on Lexicography (EURALEX'96)*, pp. 771-779, Göteborg University, Department of Swedish. Göteborg, Sweden.

Cartier, E. (1998), "Analyse automatique des textes : l'exemple des informations définitoires", In Actes *Rencontre Internationale sur l'Extraction, le Filtrage et le Résumé Automatiques*, pp. 6-18, Sfax, Tunisie, novembre 1998.

Christ, O. (1994), "A modular and flexible architecture for an integrated corpus query system", in Actes *Third Conference on Computational Lexicography and Text Research*, pp. 23-32, Budapest, Hongrie, 7-10 juillet 1994.

Condamines, A. & Rebeyrolle, J. (à paraître), "Searching for and identifying conceptual relationships via a Corpus-based approach to a Terminological Knowledge Base (CTKB): Method and Results", In L'homme et alii (à paraître).

Daoust, F. (1996). *SATO (Système d'Analyse de Textes par Ordinateur)*. Manuel de référence, Service d'analyse de textes par ordinateur (ATO), Version 4.0, Université du Québec à Montréal, Montréal.

Hearst, M.A. (1998), "Automated discovery of Wordnet relations", in C. Fellbaum (ed), *Wordnet : an electronic lexical database*, *Language Speech and Communication* 5, The MIT Press, pp. 131-151.

L'homme M.-C., Jacquemin C. & Bourigault D. (éds) (à paraître), *Recent Advances in Computational Terminology*, Amsterdam/Philadelphia : John Benjamins Publishing Company.

Meyer, I. (à paraître), "Extracting knowledge-rich contexts for terminography : a conceptual and methodological framework", in L'homme et alii (à paraître).

Morin, E. (1999), "Acquisition de patrons lexico-syntaxiques caractéristiques d'une relation sémantique", *Traitement Automatique des Langues* 40-1, pp. 143-166.

Pascual, E. & Péry-Woodley, M.-P. (1995), "La définition dans le texte", in J.-L. Nespoulous & J. Virbel (éds), *Textes de type consigne - Perception, action, cognition*, pp. 65-88, Toulouse : PRESCOT.

Pearson, J. (1998), *Terms in Context*. Amsterdam/Philadelphia : John Benjamins Publishing Company.

Péry-Woodley, M.-P. & Rebeyrolle, J. (1998), "Domain and genre in sublanguage text: definitional microtexts in three corpora", in A. Rubio, N. Gallardo, R. Castro & A. Tejada (eds), Actes *First International Conference on Language Resources and Evaluation*, pp. 987-992. Grenade, Espagne, 28-30 mai 1998.

Péry-Woodley, M.-P. (2000), *Une pragmatique à fleur de texte: approche en corpus de l'organisation textuelle*. Habilitation à Diriger des Recherches, Université de Toulouse-Le Mirail, Toulouse.

Rebeyrolle, J. & Péry-Woodley, M.-P. (1998), "Repérage d'objets textuels fonctionnels pour le filtrage d'information : le cas de la définition", in Actes *Rencontre Internationale sur l'Extraction, le Filtrage et le Résumé Automatiques*, pp. 19-30, Sfax, Tunisie, novembre 1998.

Rebeyrolle, J. (2000), "Utilisation de contextes définitoires pour l'acquisition de connaissances à partir de textes", in Actes *Journées Francophones d'Ingénierie de la Connaissance (IC'2000)*, Toulouse, IRIT, pp. 105-114.

Rebeyrolle, J. (2000), *Forme et fonction de la définition en discours*, Thèse de Doctorat Nouveau Régime.

Riegel, M. (1990), "La définition, acte du langage ordinaire - De la forme aux interprétations"., in J. Chaurand & F. Mazière (éds), *La définition*, pp. 97-110, Paris : Larousse.

Séguéla, P. (1999), "Adaptation semi-automatique d'une base de marqueurs de relations sémantiques sur des corpus spécialisés", in Actes *Troisièmes rencontres Terminologie et Intelligence Artificielle (TIA'99)*, pp. 31-42. Nantes, France.

Silberztein, M. (1993), *Dictionnaires électroniques et analyse automatique de textes : le système INTEX*, Paris, Masson.

Tanguy, L., Armstrong, S., Bouillon, P. & Lehmann, S. (1999), "DIET : Diagnostic et évaluation des systèmes de traitement de la langue naturelle", in *Langues*, 2-2, pp. 140-150.