Variation syntaxique et contextuelle dans la mise au point de patrons de relations sémantiques

Nathalie Aussenac-Gilles, Anne Condamines

Chapitre 4

Variation syntaxique et contextuelle dans la mise au point de patrons de relations sémantiques

1. Introduction

Depuis une quinzaine d'années, des marqueurs linguistiques sont employés comme un moyen potentiel de repérer des relations conceptuelles en corpus. Il s'agit d'éléments formels (typographiques, lexicaux, syntaxiques) dont on fait l'hypothèse qu'ils peuvent être utilisés de manière plus ou moins systématique pour accéder, dans des textes, à une relation lexicale ou mieux conceptuelle, déterminée *a priori* (le plus souvent, hyperonymie, méronymie, cause). Ce processus c'est ni immédiat ni complètement automatique. Il porte d'abord sur la sélection d'un certain nombre de phrases dont une partie est conforme au schéma défini par le patron. Il requiert ensuite une interprétation, éventuellement guidée par des suggestions faites sur la base des hypothèses de la signification *a priori* de ces patrons, pour identifier des termes en relation et la sémantique de cette relation. Enfin, la dernière étape consiste à s'éloigner un peu plus du texte pour décider d'une représentation conceptuelle de la relation. En ce sens, c'est une forme de filtrage sémantique au sens défini dans cet ouvrage (cf. Chapitre 1). Il s'agit bien d'un processus visant à sélectionner des éléments textuels en fonction de leur sens, et à « forcer » leur interprétation en termes de relation conceptuelle en fonction de critères linguistiques.

La notion de marqueur est mise en oeuvre par des communautés différentes, les unes venant plutôt de la linguistique de corpus et de la terminologie, les autres de l'informatique - Traitement Automatique des Langues (TAL), recherche d'information ou ingénierie des connaissances (IC) -. Cette apparente similarité d'utilisation ne rend pas compte des divergences de points de vue qui peuvent exister : un marqueur décrit dans toute sa finesse (caractérisation précise du contexte

---

Chapitre rédigé par Nathalie AUSSENAC-GILLES et Anne CONDAMINES

syntaxique dans lequel il joue ce rôle) par un linguiste peut n'avoir aucune pertinence pour un informaticien s'il ne peut le mettre en œuvre avec les outils dont il dispose. Inversement, les potentiels des analyses automatiques conduisent à définir des marqueurs plus efficaces mais difficiles à mettre au point. Et surtout, les exigences de l'automatisation conduisent à vouloir apprendre automatiquement ces marqueurs à partir de corpus, conduisant à définir des marqueurs dont la portée linguistique n'est pas évaluée. Enfin, la disponibilité d'un volume croissant de documents sous format électronique, et en particulier sur le web, offre un champ d'étude aux dimensions nouvelles.

Avec le recul qu'apportent les expériences d'analyse et d'utilisation de marqueurs de relation, il semble que l'on puisse à présent redéfinir cette notion de manière plus fine et plus pertinente. En particulier, depuis quelques temps, est apparue l'idée de la variation du fonctionnement d'un marqueur non seulement en fonction du contexte distributionnel mais aussi de la situation de production des textes, c'est-à-dire non seulement en fonction de la syntaxe mais aussi en fonction du genre textuel (cf. 3.4). Examiner le fonctionnement des marqueurs permet ainsi d'évaluer sous un autre jour la complémentarité entre caractéristiques linguistiques et caractéristiques extra-linguistiques dans le fonctionnement sémantique.

Ce chapitre se propose d'interroger la variation du fonctionnement des marqueurs en croisant les regards du TAL, de l'IC et de la linguistique de corpus. La section 2 situe la notion de marqueur dans l'historique de la terminologie et de la linguistique de corpus. Nous y montrons comment les hypothèses de stabilité et de régularité du sens ont été progressivement remises en question, bousculant ainsi les hypothèses sur le caractère fixe des interprétations possibles des marqueurs. Nous précisons ce que couvre cette notion en linguistique ainsi qu'en informatique pour le TAL et l'IC, dans le cas particulier de l'utilisation de marqueurs pour repérer des relations sémantiques dans des textes. L'analyse linguistique présentée dans la section 3 détaille les différents types de variation identifiés dans l'interprétation des patrons ou des phrases qu'ils permettent de retrouver. La section 4 fait le point sur les implémentations informatiques des patrons, et leur capacité à prendre plus ou moins en compte ces variations. Nous concluons en proposant de renouveler le point de vue actuel sur les marqueurs et la manière dont ils permettent d'extraire de l'information dans les textes.

2. Marqueurs et relations

2.1. Marqueurs et terminologie

Sous l'influence de l'approche du Cercle de Vienne pendant longtemps, la terminologie (inspirée par les travaux de Wüster) est apparue comme un moyen stable et supposé sûr d'accéder à la connaissance ; mais il fallait pour cela qu'elle soit fixée et contrôlable. La question qui se posait et qui se pose encore pour beaucoup de terminologues est celle du rapport de nomination avec la réalité :

« Nous pouvons également formuler [les hypothèses philosophiques et cognitives] à travers une liste de questions qui depuis des siècles reçoivent toujours de nouvelles réponses : - Qu'est-ce que la réalité et en quoi consiste-t-elle ? Que sont les objets dans la réalité ? - Comment pouvons-nous connaître une telle réalité ? Comment pouvons-nous connaître et saisir des objets ? - Que pouvons-nous dire de la réalité ? Comment pouvons-nous dénommer des objets ? » [BUD 07] : 13.

« Les fondements de la théorie de la connaissance du type ontologie réaliste sont une base solide et un point de départ pour les ontologies actuelles, auxquelles ont recours l'informatique et le traitement cognitif du langage et du savoir. Elles sont aussi à la base de la science de la terminologie » (ibid., 19).

Comme l'a montré M. Slodzian [SLO 95], la doctrine wustérienne se fonde sur le postulat d'une langue universelle permettant l'accès à la connaissance. Avec les langues « spécialisées », associées à des domaines parfaitement maîtrisés par des experts, ce type de doctrine a cru trouver confirmation de son hypothèse de la possibilité d'une langue qui, à défaut d'être pure, pourrait être purifiée (normalisée). Un tel point de vue a pour objectif de tenir éloignés les termes des discours qui les utilisent et surtout de viser une normalisation dont on pense qu'elle garantira des échanges transparents entre différentes communautés dans le même domaine, comme le souhaite Wüster :

«[…] jusqu'à une date récente, la linguistique n'a fait valoir que l'évolution libre, non dirigée, de la langue. C'est l'usage effectif de cette dernière qui, dans la langue commune, sert de norme. On peut appeler cette norme la norme descriptive. En revanche, en terminologie, fertile en notions et en termes, cette évolution libre de la langue mène à une confusion inacceptable… » ([WUS 81] : 65).

Depuis une quinzaine d'année, ce point de vue a évolué sous l'influence d'un double constat. D'une part, la terminologie ainsi fixée ne correspondait pas à celle utilisée dans les documents d'entreprises, il semblait donc impossible d'imposer une norme qu'on n'avait aucun moyen de faire appliquer. D'autre part, l'utilisation de

ces normes préfixées et génériques pour certaines applications de traitement automatique était tout à fait inefficace si on changeait de point de vue applicatif [CHA 96]. Plusieurs travaux ont fait état de la diversité des modèles existant dans un même domaine, des différences d'analyse et de structuration qu'ils reflètent, enfin de leur inadéquation fréquente à de nouvelles utilisations [SMA 08].

Plutôt qu'une approche censée aller des objets vers leur nomination, une nouvelle approche s'est mise en place qui va des discours réels vers les ressources terminologiques ou ontologiques (désormais ressources termino-ontologiques) pour mieux pouvoir utiliser ces ressources afin d'accéder automatiquement à certaines informations dans les textes [AUS 07]. C'est donc sous l'influence des besoins des entreprises (améliorer leur gestion de la documentation) mais aussi par le rapprochement des questions de terminologie avec des questions de linguistique qu'une évolution s'est faite vers la prise en compte des discours comme source de la terminologie. En réalité, c'est à un changement de paradigme complet qu'on a assisté. Les discours en effet ne sont pas stables, contrôlables et définissables comme le sont les objets référentiels et travailler à partir de textes, c'est être confronté à la variation et à la nécessité d'interpréter :

« Au lieu de partir d'une ontologie préfixée, dont le texte ne serait qu'une manifestation toujours partielle et imparfaite, [la conception rhétorico/ herméneutique] cherche à faire émerger corrélativement des régularités et des singularités, et à leur faire correspondre, par construction interprétative, des fonds et des formes sémantiques. » ([RAS 01] : 90).

Le même type d'évolution s'est fait dans la constitution d'ontologies. Si beaucoup de chercheurs visent encore à construire des ontologies générales, beaucoup s'orientent maintenant vers la constitution d'ontologies régionales, liées à un domaine, voire, à une application (cf. Chapitre 6). Les textes jouent un rôle majeur puisqu'ils sont censés contenir les manières de dire et de penser de communautés qui ont en commun une activité, le plus souvent professionnelle. Les ressources termino-ontologiques sont représentées sous forme d'un réseau composé de nœuds (les termes) reliés par des arcs (des relations plus ou moins liées au domaine). Leur essor met donc également en avant les relations sémantiques et leur repérage dans la langue écrite. Les relations jouent un rôle clé à plusieurs titres : 1) elles permettent de définir des concepts avec précision ; 2) la construction de ces modèles devant être rapide voire automatisée, une des solutions est de les construire à partir de textes, et les relations forment alors des indices identifiables dans les textes ; 3) ces modèles sont utilisés entre autres pour l'annotation de textes et pour l'extraction d'information à partir de textes, ce qui suppose encore de repérer des concepts, des instances de concepts et des relations sémantiques dans les textes. Dans ce contexte, l'automatisation du repérage des relations est affichée comme l'objectif de travaux qui font appel soit aux statistiques pour exploiter les régularités

de l'usage de la langue, soit à la linguistique pour mettre en œuvre des approches par patrons, soit encore à l'apprentissage automatique pour « apprendre » certains de ces patrons [BUI 05] [BAR 04].

Les logiciels qui exploitent informatiquement des marqueurs de relation pour identifier des relations font une ou plusieurs des hypothèses suivantes : celle de la stabilité de ces marqueurs d'un corpus à l'autre ou d'un domaine à l'autre, celle de la stabilité d'interprétation de la relation ainsi identifiée, enfin celle de la capacité à reconnaître les éléments mis en relation.

Avant même la question des marqueurs, se pose celle des relations que des éléments sont censés « marquer ».

2.2. La question des relations

2.2.1 Représentation des relations

On ne peut pas étudier le fonctionnement des marqueurs sans s'interroger sur la question des relations. Dans les modes de représentation qui sont utilisés pour représenter les ressources termino-ontologiques, les relations jouent un rôle aussi important que les concepts :

« If concepts are seen as the basic building blocks of conceptual structure, then relationships are the mortar that holds it together » ([GRE 02] , viii).

Il est même assez fréquent que pour décider de conserver ou non un candidat-terme dans une ressource termino-ontologique, le fait qu'il soit en relation, dans un corpus, avec un autre terme ou candidat-terme soit déterminant [BAR 04] [GRA 04].

Un constat semble évident bien que peu souvent noté : dans la grande majorité des termino-ontologies existantes, les relations sont moins spécialisées que les termes qu'elles réunissent. Ainsi, dans UMLS (Unified Medical Language System), un thesaurus médical réalisé par la NLM (National Library of Medicine), aucune des 54 relations retenues n'est incompréhensible pour un non-expert[1]. Certaines sont assez nettement typiques du domaine médical comme *diagnoses* ou *developmental-form-of* mais elles restent accessibles à des non-experts, qui fréquentent aussi des médecins. La plupart sont très générales, indépendantes du domaine *: ingredient-of, location-of, produces* … La spécificité du domaine se trouve ainsi pratiquement tout entière dans les termes reliés et, éventuellement, les qualificatifs.

[1] http://www.nlm.nih.gov/research/umls/META3.HTML

2.2.2 Des relations aux marqueurs

La pertinence des réseaux relationnels pour représenter la connaissance peut être remise en cause (voir par exemple [RAS 95]). Mais tel n'est pas notre propos dans ce chapitre. Nous partons de l'hypothèse que nous acceptons qu'il y a une pertinence et un intérêt à vouloir représenter la connaissance (ou un point de vue sur cette connaissance) qu'on peut trouver dans un corpus sous la forme d'un réseau de termes. Il est clair que si les relations sont accessibles à des non-spécialistes, les marqueurs (tout au moins la plupart des marqueurs) qui leur sont associés le sont probablement aussi. Ce qui ne veut pas dire qu'on peut retrouver ces marqueurs par simple réflexion introspective, c'est-à-dire associer spontanément à une relation l'ensemble des marqueurs qui peuvent lui être associés. Nous verrons que, dans beaucoup de cas, les marqueurs ne peuvent pas être spontanément décrits. Il n'en reste pas moins que, dans les faits, on fonctionne avec une liste plus ou moins consciente de relations auxquelles on cherche à associer des marqueurs. De fait, les relations qui ont été les plus étudiées sont la relation d'hyperonymie [CRU 02] [HEA 92] [BAR 04], la méronymie [GUA 96] [WIN 87] et la relation de cause [GAR 98] [KHO 02] [GIR 02].

Sans doute peut-on considérer que ces relations sont les plus fréquentes, quel que soit le domaine, même si elles ne sont pas nécessairement les plus fréquentes dans un corpus donné. Pour autant, elles ne sont sans doute pas aussi stables et consensuelles qu'on pourrait le penser. D'une part, il existe plusieurs types de relations méronymiques, causatives et même hyperonymique. D'autre part, selon les domaines et le genre textuel, ce ne sont pas les mêmes relations méronymiques, hyperonymiques ou causatives qui vont apparaître. Enfin, selon le point de vue que l'on a sur la représentation que l'on veut construire, les relations (les modes de représentation) peuvent changer, comme nous allons le montrer.

2.3. Marqueur de relation conceptuelle en linguistique et en informatique : diversité et points communs

Nous venons de présenter comment la notion de marqueur s'est développée dans la problématique de la construction de ressources termino-ontologiques, lorsqu'il est apparu que cette élaboration pouvait être faite à partir de textes. Cette notion de marqueur existe depuis longtemps en linguistique ; et bien avant l'élaboration de ressources termino-ontologiques, des sémanticiens se sont intéressés à certaines des réalisations linguistiques qui pouvaient manifester certaines relations comme l'hyperonymie (générique-spécifique) ou la méronymie (partie à tout). De nombreux travaux se sont ainsi développés en terminologie textuelle, aussi bien en informatique qu'en termino-linguistique. Mais dans les premiers temps, ces travaux se situaient dans la perspective d'une vision stable du fonctionnement des termes et

des relations en discours. Ce n'est que récemment que l'idée d'une variation en fonction de différents paramètres est apparue.

Dans cette partie, nous situons la notion de marqueur, et des notions proches comme celle de patron, dans les travaux en linguistique et en informatique. Nous soulignons l'hypothèse commune de stabilité et introduisons les éléments relatifs à la variation dont l'étude sera détaillée dans la partie suivante.

2.3.1 Origine linguistique de la notion de marqueur

Bien que ne venant pas originellement de la linguistique, le terme de marqueur est assez fréquemment utilisé dans cette discipline. Il dénomme un élément langagier auquel on peut attribuer une interprétation de manière consensuelle. Il relève ainsi d'une connaissance du fonctionnement métalinguistique. Par exemple, le suffixe *–ette* marque la notion de petitesse dans *maisonnette* ou *voiturette*.

D'un point de vue linguistique, [MEY 01] décrit ainsi la nature de trois types de marqueurs de relation : *lexical patterns* ("involving one or more specific lexical items"), *grammatical patterns* and *paralinguistic patterns* ("which include punctuation, as well as various elements of the general structure of a text"). L'auteur précise aussi que "complex in their nature, and in the way they can be realized in text: they are sometimes unpredictable, polysemic, and/or domain-dependent. ([MEY 01] : 290). Cette description fait déjà apparaître nettement la variation possible du fonctionnement des marqueurs et, dans un certain nombre de cas, leur imprévisibilité, variation dont nous reparlerons longuement.

2.3.2 Marqueurs et patrons en informatique

L'informatique, à travers les travaux sur l'analyse automatique des langues ou la linguistique informatique, a emprunté cette notion pour la redéfinir [GRA 04]. Les objectifs de définition, d'utilisation et d'implémentation des marqueurs divergent entre informatique et linguistique. Pour l'informatique, les marqueurs sont une des techniques possibles pour accéder au contenu informationnel de documents numériques contenant du texte. Ils sont donc utilisés pour rechercher des informations précises, dans le cadre du traitement automatique des langues pour comprendre, résumer, analyser des textes, en modélisation de connaissances (construction et peuplement d'ontologies par exemple) ou encore en extraction d'information. En extraction d'information, l'objectif est de repérer dans des textes des phrases concernant des « individus », des « instances » pour ensuite figer les régularités observées dans une représentation, sous un format exploitable automatiquement. L'informatique se focalise donc sur les aspects techniques, la forme et l'efficacité de traitement des textes à l'aide des marqueurs pour atteindre un objectif donné, plus que sur leur fonctionnement détaillé dans les textes.

Par exemple, en recherche d'information, un marqueur est défini ainsi : « forme linguistique faisant partie de catégories prédéfinies (grammaticales, lexicales, syntaxiques ou sémantiques) dont l'interprétation définit régulièrement le même rapport de sens entre les termes». ([HAD 02] : 37). Cette définition réduit le marqueur à un seul élément linguistique, se focalise sur les éléments constitutifs du marqueur plus que sur son rôle. Elle suppose que le patron, lui, renvoie à une forme plus complexe, composée de plusieurs marqueurs organisés de manière particulière. Mais le plus souvent, marqueur et patron sont pris comme synonymes, ou alors le patron est considéré comme l'implémentation d'une caractérisation au niveau linguistique qui serait le marqueur.

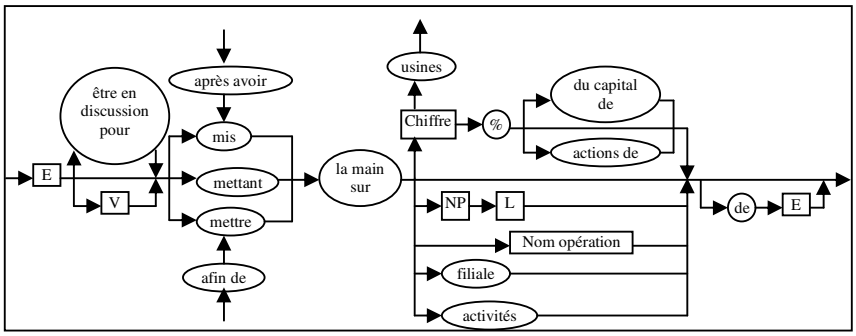


Fig. 4.1 : exemple d'automate d'extraction d'information [AMA 01]

L'implémentation des marqueurs correspond généralement à des automates à états finis, à des expressions régulières balayées par un programme qui cherche à retrouver en corpus la séquence d'éléments ainsi définis. Par exemple, en figure 4.1, le schéma représente le patron « mettre la main sur » permettant de retrouver une phrase qui indique une relation entre deux entreprises (notées E). Les éléments composant l'expression varient suivant les approches : mots, lemmes, catégories sémantiques, catégories grammaticales, rôles syntaxiques, symboles propres au logiciel pour marquer les répétitions par exemple, contraintes de mise en forme etc. Dès que le patron caractérise les informations cherchées autrement que comme des chaînes de caractères, une analyse préalable du texte, plus ou moins fouillée, doit être réalisée : analyse syntaxique, sémantique, repérage de la mise en forme, repérage de classes sémantiques, etc. Le repérage de relations correspond donc à une tâche complexe, faisant appel à une chaîne de traitements automatiques ou manuels des textes avant de pouvoir projeter des patrons.

2.3.3 Marqueurs de relation conceptuelle

De manière générale, aussi bien en linguistique qu'en informatique, un marqueur de relation conceptuelle permet de repérer de manière consensuelle la présence d'une relation entre deux éléments. Par consensuelle, nous entendons une interprétation acceptable au moins par un groupe de locuteurs partageant un point de vue.

Remarquons que la notion de marqueur de relation se situe plutôt dans la perspective d'interprètes qui ne sont pas les interlocuteurs initialement visés (et c'est le cas lors de la construction de ressources termino-ontologiques à partir de textes). Le marqueur de relation est ainsi associé à une volonté définitoire du lecteur interprète, c'est-à-dire, dans le cas qui nous intéresse, à la possibilité de représenter une portion de texte sous la forme d'un triplet terme/relation/terme. On peut ainsi parler de genre interprétatif comme nous l'avons fait dans [CON 07], c'est-à-dire de l'existence d'un ensemble de locuteurs qui peuvent s'accorder sur l'interprétation de passages de textes. Fish parle de communautés interprétatives qu'il définit de la sorte (il parle de convention plutôt que de consensus) :

« Les significations que [l'ego] confère au texte ne sont pas les siennes mais trouvent leur source dans la (ou les) communauté(s) interprétative(s) sur laquelle (ou lesquelles) il repose. On pourrait dire aussi que [ces significations] sont à la fois subjectives et objectives : elles sont subjectives parce qu'inhérentes à un point de vue particulier et donc non universelles ; et elles sont objectives parce que le point de vue qui les délivre est public et conventionnel plutôt qu'individuel ou singulier. » ([FIS 07] : 74).

```
divers_X_comme_Y
^divers$ {ADV|PRP.*|NOM|DET.*|ADJ|KON}* ">^comme$"

X-dénomination-de-Y
>^(porter|appliquer|employer|réserver|recevoir|prendre|utilise
r|donner|proposer|mériter)$ 6 ">^[cl]e$" ">^(nom|terme|mot|exp
ression|vocable|appellation|désignation|dénomination)$"
```

Fig. 4. 2 : exemples de patrons sous forme d'expressions régulières

L'informatisation de la recherche de relations suppose de définir des patrons incluant ces marqueurs. Les composants d'un patron peuvent être des éléments sémantiques (comme la liste des verbes (porter|appliquer| …) sur la figure 4.2), des catégories grammaticales, des mots recherchés dans leur forme exacte ou à partir de leur lemme, ou encore des éléments de phrase non caractérisés (sur la figure 4.2, 6 signifie que 6 mots au plus se trouvent entre le premier verbe reconnu dans la liste et « ce » ou « le », * signifie un nombre indifférencié de mots). Suivant les objectifs

poursuivis, les éléments constitutifs des patrons se contentent de caractériser les seuls indices de la relation, ou alors ils ciblent à la fois la relation et les entités qu'elle relie, ou un seul de ces éléments (sur la figure 4.2, le patron « divers_X_comme_Y » caractérise seulement le premier élément en relation (X), par les catégories situées entre « divers » et « comme »). Plus on souhaite automatiser l'interprétation des fragments de discours ainsi retrouvés (et donc ne pas faire intervenir l'interprétation humaine), plus il faut caractériser précisément les éléments explicitant la relation et les termes mis en relation. Les applications les plus exigeantes en la matière sont l'extraction d'information, l'objectif étant de constituer des tables ou des bases de données enregistrant des informations spécifiques mises en relation (des « instances » de concepts comme des noms de maladies, de gènes ou d'entreprises, des valeurs de paramètres, etc.).

2.3.4 Stabilité *vs* variations

Tant dans les études linguistiques « classiques » que dans leur mise en œuvre informatique, les hypothèses fortes derrière la notion de patron sont la régularité et la stabilité de la sémantique du patron dans le corpus : on s'attend à pouvoir réutiliser des marqueurs identiques d'un corpus et d'un domaine à un autre, on s'attend à retrouver, à l'aide de ces mêmes marqueurs, des éléments linguistiques révélant une relation ayant un sens particulier, et éventuellement aussi, à une place déterminée par le patron, des termes ainsi mis en relation. Ces termes seront considérés comme des indicateurs de classes sémantiques ou d'instances spécifiques de ces classes selon les applications. Or, ce qui est stable avant tout, c'est le patron lui-même. L'informatique fournit de bons instruments de repérage de régularités de forme et à ce titre, les patrons jouent bien leur rôle. Malgré tout, leur mise au point est loin d'être triviale, comme l'ont montré Rebeyrolle et Tanguy [REB 00]. Plusieurs ajustements sont nécessaires pour parvenir à couvrir les différentes formes que l'on souhaite caractériser, et seulement celles-là. Ces ajustements découlent de deux variations au moins :

- la nature des informations recherchées, et l'on pourrait rajouter « ce qui va en être fait, le traitement informatique envisagé de ces informations » ;

- la nature du corpus : les corpus varient tant par leurs spécificités sémantiques et syntaxiques, que par leurs caractéristiques situationnelles.

De fait, ce qui est beaucoup moins stable, c'est l'interprétation des fragments de discours (le plus souvent des phrases) retournés par un patron. Or dans la visée informatique, cette variation n'est pas du tout prise en compte, la plupart des définitions faisant l'hypothèse d'un rapport relation/marqueur stable avant même la mise en discours. Cette variation dérange, et elle est ignorée au profit d'une

exploitation la plus automatique possible. On traite alors uniformément des données diverses, ce qui risque d'augmenter le silence et le bruit de l'analyse.

Nous allons revenir sur la diversité des points de vue que linguistes, terminologues et informaticiens portent sur la notion de marqueur, en nous interrogeant sur la manière dont chaque matérialisation ou implémentation favorise ou non la prise en compte de la variation. Nous commencerons (partie 3) par une étude linguistique visant à caractériser la nature et la signification des variations qui apparaissent dans les textes, d'une part, dans le triplet terme/marqueur/terme et, d'autre part, dans le passage du discours à la représentation sous la forme terme/relation/terme. Nous poursuivrons (partie 4) par un panorama et une étude critique des différentes implémentations informatiques actuelles des marqueurs, en dessinant des pistes pour faire évoluer ces approches vers une meilleure prise en compte des éléments de variation mis en avant par l'analyse linguistique.

3. Fonctionnement des marqueurs : variations linguistiques

La vision d'un fonctionnement idéal des marqueurs serait celle où, à un triplet terme/marqueur/terme, on pourrait associer un triplet terme/relation/terme de manière systématique, c'est-à-dire permanente et univoque. L'instabilité de cette équivalence peut venir de différents phénomènes que l'on peut résumer sous trois rubriques : ceux qui concernent le triplet contenant le marqueur (donc le discours) ; ceux qui concernent le triplet contenant la relation ; enfin, ceux qui concernent l'interprétation du triplet avec marqueur pour le transformer en triplet avec relation.

3.1. Variation des triplets

Les deux triplets qui nous intéressent sont, d'une part, le triplet terme/marqueur/ terme qui concerne le fonctionnement dans le texte et, d'autre part, le triplet terme/relation/terme associé à des choix de représentation.

3.1.1 Instabilité du triplet terme/marqueur/terme

La structure terme/marqueur/terme concerne nécessairement le discours et sa complexité. De fait, il est assez rare que ce triplet apparaisse de manière évidente. La plupart du temps, leur mise au jour demande un travail d'interprétation guidé par l'intention de construire un triplet terme/relation/terme. Ainsi, il peut manquer un des termes, les trois éléments peuvent apparaître dans un ordre différent, ou encore les deux termes peuvent ne pas avoir la même catégorie grammaticale.

3.1.1.1 Un des termes est absent

Cela peut être le cas par exemple lors de la reprise par une anaphore pronominale. Seul alors un des termes est explicitement présent dans la phrase. Il faut alors résoudre l'anaphore, ce qu'un humain fait assez facilement mais un outil beaucoup plus difficilement. Cela peut aussi être le cas dans la reprise par une anaphore infidèle. On sait que cette structure se construit souvent sur une relation d'hyperonymie. Ainsi dans :

Un chat est entré, cet animal au pelage clair…

Il y a une relation d'hyperonymie entre *animal* et *chat*.

De fait, la présence de cette relation est moins fréquente qu'on pourrait le penser [CON 05]. Par ailleurs, dans certains cas, il est impossible d'identifier un terme antécédent, c'est tout un paragraphe qui joue ce rôle. Ainsi dans l'exemple :

A la mi-1988, le portefeuille d'investissements indirects [du Japon] dépassait les 100 milliards de dollars. Sauf krach -hypothèse plausible - tout indique que ce mouvement va continuer. (Le Monde Diplomatique).

Mouvement pourrait avoir une fonction d'hyperonyme mais il n'y a pas d'hyponyme identifiable dans le contexte qui précède.

Autre difficulté, cette relation peut s'instaurer entre deux phrases, ce qui ne simplifie pas le traitement automatique qui, généralement, s'intéresse plutôt à la recherche d'information à l'intérieur d'une seule phrase.

3.1.1.2 Les positions des éléments ne sont pas stables

L'ordre terme1/marqueur/terme2 n'est pas toujours respecté. On peut trouver des exemples avec marqueur/terme1/terme2 :

Avec 4 roues motrices, la voiture X est la plus performante.

(le marqueur est *avec*, l'holonyme *voiture* et le méronyme *4 roues motrices*)
ou encore avec terme2/marqueur/terme1 (avec un même marqueur et une même interprétation mais un changement d'orientation de la relation). Ainsi dans

La rose comme fleur de décoration est particulièrement appréciée

Il y a une hyperonymie entre *fleur de décoration* et *rose*. Dans :

Une fleur de décoration comme la rose se vend bien.

On retrouve la même hyperonymie mais avec une inversion de position entre T1 et T2 (et un changement de déterminant).

3.1.1.3 Nature grammaticale différente pour T1 et T2

> *Le contrôle a posteriori consiste, pour un ouvrage donné, à examiner la manière dont ont été appliquées les doctrines techniques.*

Dans cette phrase, le terme 1 est un groupe nominal (*contrôle a posteriori*), alors que le terme 2 est un groupe verbal qui pourrait, le cas échéant être transformé en groupe nominal (*examen de la manière d'appliquer les doctrines techniques*).

3.1.2 Instabilité du triplet terme/relation/terme

Cette instabilité concerne les choix de représentation d'un même triplet terme/marqueur/terme. Deux cas peuvent se présenter. D'une part, la décision de répartir les éléments entre les termes et la relation peut varier. D'autre part, une même relation peut être déclinée différemment.

3.1.3 Variation dans la répartition des éléments

Le choix de l'organisation des éléments est guidé par un principe d'économie représentationnelle. Dans l'exemple suivant :

> *La phase d'intégration du composant peut commencer lorsque l'ensemble des éléments logiciels ont été codés.*

une relation « conditionne » a été identifiée. Un examen plus attentif des exemples nous a amenés à réorganiser cette relation en « conditionne le début de » et « conditionne la fin de ». L'alternative aurait été d'intégrer *début* et *fin* dans les termes eux-mêmes (par exemple, *début de la phase d'intégration du composant* est un syntagme qui, dans ce contexte, semble pouvoir être considéré comme un terme). C'est pour des raisons d'économie (réduire le nombre de termes) que nous avons préféré intégrer *début* et *fin* dans la relation elle-même.

3.1.3.1 Déclinaison de la relation

Les travaux qui portent sur l'analyse des relations proposent dans la plupart des cas de décliner une même relation en sous-relations qui semblent pouvoir être rattachées à une relation principale, associable à un marqueur.

Ainsi, plusieurs auteurs ont proposé d'organiser la relation de méronymie, sur les bases d'une analyse de type linguistique et/ou cognitif [KEE 08]. [VIE 07], [WIN 87] par exemple proposent la classification suivante : composant/objet, membre/collection, portion/masse, constituant/objet, activité/phase, zone/lieu. Certains marqueurs orientent directement vers un certain type de méronymie (comme être situé dans pour la relation zone/lieu) mais pour certains autres, liés à la méronymie en général (comme être composé de), il peut être plus difficile

d'identifier quelle relation de partie à tout est concernée, surtout si on n'est pas compétent dans le domaine.

Le flan est composé de lait, d'œufs et de sucre (constituant/objet)

La quiche est composée de flan et de jambon (composant/objet).

3.2. Instabilité du passage du triplet terme/marqueur/terme au triplet terme/relation/marqueur

On est là au cœur du problème : il s'agit de décontextualiser les éléments discursifs pour en faire des éléments d'une représentation dont on fait l'hypothèse que, d'une part, elle pourra rendre compte d'autres parties de texte « exprimant » la même relation sous une autre forme et que, d'autre part, elle sera consensuelle (acceptable par un groupe de locuteurs).

3.2.1 « Polysémie » des marqueurs

Nous avons vu des exemples où un même marqueur peut renvoyer à plusieurs déclinaisons d'une même relation (comme *être composé de* pour plusieurs relations méronymiques). Mais beaucoup de marqueurs peuvent être associés à plusieurs relations comme le signalait Meyer [MEY 01]. Citons par exemple *faire partie de* qui peut renvoyer soit à la méronymie, soit à l'hyperonymie [MAR 07] :

Les chiens font partie des mammifères

est ainsi plutôt à interpréter comme rendant compte d'une hyperonymie alors que

Les feuilles font partie des composants des arbres

est à interpréter comme évoquant une méronymie.

3.2.2 Effet rhétorique

Les marqueurs de relation n'échappent pas à un fonctionnement discursif visant à un effet rhétorique particulier. Un exemple de ce phénomène consiste à vouloir atténuer l'effet injonctif en utilisant des tournures considérées comme plus neutres. Dans les exemples suivants, extraits d'un manuel rédigé à EDF par le service qualité, un marqueur qui, hors contexte, serait considéré comme un marqueur de succession (conjonction de subordination temporelle + passif) doit être compris comme une obligation d'effectuer une tâche avant de pouvoir en commencer ou en terminer une autre. Ainsi,

La phase d'intégration du composant peut commencer lorsque l'ensemble des éléments logiciels ont été codés.

Doit être comprise comme la phase d'intégration du composant ne peut commencer que lorsque l'ensemble des logiciels ont été codés *et donc la phrase comme exprimant une relation de condition.*

3.2.3 Interprétation directe vs indirecte

Dans certains cas, le marqueur lui-même semble donner directement l'interprétation sous la forme d'une relation. Dans d'autres cas, cette interprétation n'est pas aussi clairement posée et ne peut être déduite que par la prise en compte d'un certain nombre d'éléments. Ainsi, une phrase comportant *chez*, peut être interprétée comme comportant une méronymie ; par exemple, dans :

Chez les colobinés, le nez fait saillie sur la lèvre supérieure

il y a une relation de méronymie entre *nez* et *colobinés.*

En réalité, nous le verrons ci-dessous, il est inexact de dire que *chez* marque la méronymie. *Chez* permet de mettre un élément en position thématique (en l'occurrence un animal ou une plante) le reste de la phrase donnant des informations sur cet élément. Le plus souvent, cette information est de type méronymique (anatomique en fait) mais seulement si le texte relève des sciences naturelles et a une visée didactique. D'un point de vue linguistique, il est difficile de dire que *chez* « marque » une méronymie. En revanche, d'un point de vue informatique, il est avéré que dans certains textes ayant certaines caractéristiques, dans plus de 50% des occurrences, on aura une relation de méronymie entre N1 et N2 [CON 00].

3.3. Rôle du co-texte dans l'interprétation : prise en compte de la syntaxe

Les éléments appelés marqueurs correspondent souvent à des éléments isolés : verbes (*provoquer* pour la cause), noms (*partie* pour la méronymie), prépositions (*pour* pour le but). Or, il est possible de contraindre les contextes dans lesquels apparaissent ces éléments afin d'améliorer leur possibilité de jouer le rôle de marqueurs. Ainsi, *provoquer* seul peut renvoyer à bien d'autres choses que l'expression de la cause. Mais il y a beaucoup plus de chances de trouver cette relation dans les textes si on limite les cas examinés à ceux qui sont précédés ou/et suivis d'une nominalisation déverbale [GAR 98] :

Une compression trop franche provoque une défense de la paroi abdominale (cause) Vs L'enfant provoque le chien (pas de cause).

Même constat pour *partie* qui renvoie plus sûrement à une méronymie si le contexte est : *avoir pour partie* ou *être une partie de* que par exemple *gagner la partie*. Dans certains cas cependant, la structure syntaxique ne suffit pas à sélectionner les seuls cas où l'élément peut être associé à une relation. Prenons

l'exemple de *chez* dont nous avons parlé. L'étude de cette préposition a montré qu'elle apparaît dans trois types de contextes syntaxiques [CON 01]:

1- *Chez* (det1) N1, structure présentative det2 N2
   *Chez les lémurs, il existe des zones glandulaires circumgénitales (*N1 *: Lémurs,* N2 *: zones glandulaires circumgénitales).*
2- *Chez* (det1) N1, det2 N2 prédicat
   Les callosités ischiatiques sont séparées chez les mâles comme chez les femelles (N1 *: mâles, femelles,* N2 *: callosités ischiatiques).*
3- *Chez* (det1) N1, (det3 N3) prédicat det2 N2
   L'arrivée du printemps crée une sorte de fièvre chez les observateurs d'oiseaux (N1 *observateurs d'oiseaux*, N2 *fièvre,* N3 *arrivée du printemps*).

Aucune de ces structures ne favorise l'interprétation méronymique. Pour chacune d'elles, il existe des exemples où l'interprétation méronymique est présente (ici en 1 et 2) et d'autres où cette même interprétation est impossible (comme en 3). Il faut donc essayer de trouver un autre élément qui pourrait jouer un rôle dans l'établissement d'une interprétation régulière. C'est ce que nous proposons avec la notion de genre textuel dans le prochain paragraphe.

3.4. Rôle du genre textuel dans l'interprétation

3.4.1 La notion de genre textuel

La notion de genre textuel renvoie à l'idée d'une variation dans l'apparition et le fonctionnement d'éléments langagiers, en lien avec des éléments situationnels. Cette notion est tout à fait pertinente pour les marqueurs de relation, c'est-à-dire ce qui concerne leur présence (le fait qu'ils apparaissent ou non) et leur interprétation (le fait que tel élément langagier ait une forte probabilité d'être interprété comme un marqueur relationnel dans tel ou tel type de texte).

La prise en compte de la situation permet tout d'abord de repérer des cas où bien que le locuteur utilise les marqueurs définitoires, ces passages ne seront pas retenus dans une termino-ontologie car la définition n'est pas consensuelle mais propre au locuteur. C'est le cas par exemple des « définitions » dans des contextes littéraires. Ainsi l'extrait de Germinal suivant

Le brigand est le vrai héros, le vengeur populaire, le révolutionnaire en acte.

a les caractéristiques d'un contexte définitoire mais le genre textuel dans lequel il apparaît fait qu'on ne le retiendra pas comme une définition de *brigand.*

La prise en compte d'une variation en lien avec le genre textuel n'est apparue que récemment en terminologie [ROG 00] ou en TAL. Pourtant, elle pourrait

permettre non seulement d'améliorer les résultats d'analyse mais aussi, si l'on peut expliquer le lien entre certaines régularités langagières et des caractéristiques extra-linguistiques, d'anticiper les fonctionnements. Des travaux prenant en compte le genre textuel existent désormais [REB 00], [CON 09], [AUS 08c], [MAR 08] … Mais ils devraient encore se développer du point de vue linguistique. C'est en effet un champ de recherche qui n'a pas encore été complètement exploité.

Dans la constitution de ressources termino-ontologiques, et en particulier dans la description des marqueurs de relation, les éléments extra-linguistiques qui peuvent jouer un rôle dans la caractérisation d'un genre sont, d'une part, le domaine, spécialisé ou pas, étant entendu qu'on associe souvent un fonctionnement terminologique à un domaine spécialisé et, d'autre part, la visée, l'intention de communication. En effet, tout texte s'accompagne d'une intention de communication qui peut avoir une influence sur les manières de communiquer (dans l'exemple ci-dessus, le souhait de modérer l'objectif injonctif a amené à utiliser un marqueur de succession qu'il faut interpréter comme renvoyant à une relation de condition). Ces deux éléments - domaine dont relève le texte et visée de communication - peuvent intervenir de manières diverses dans le fonctionnement des marqueurs. Examinons ainsi deux prépositions, *avec, chez*, qui peuvent être associées à la méronymie et leur fonctionnement en fonction des caractéristiques des textes.

3.4.2 Importance de la visée : le cas de *avec*.

	GERMINAL	GEO	Toy catalogue	Small-adds	Itineraries
Nombre mots	206,700	230,000	93,000	22,600	48,000
Avec	667	432	236	185	114
Avec méronymique	43 (3%)	55(12.7%)	161(68.2 %)	141(76.2 %)	75(64.6%)

Tableau 4.1 : comportement de *avec* dans différents genres de corpus[2].

Dans certains énoncés, *avec* peut être considéré comme un marqueur de méronymie (*une robe avec des dentelles*). L'étude d'un grand nombre d'énoncés comportant cette préposition a montré que cette possibilité de marquage méronymique est loin d'être équiprobable dans tous les textes [CON 06]. Sans plus détailler le fonctionnement de cette préposition, nous pouvons dire que les textes ayant une probabilité élevée de présence de *avec méronymiques* ont la même visée. Il s'agit de valoriser un objet en mettant en évidence une ou plusieurs de ses parties « saillante(s) ». La saillance peut être de nature perceptive ou « commerciale » :

Vous repèrerez la place grâce à une église avec une coupole (itinéraires)

[2] Germinal, roman de Zola ; GEO, manuel de géomorphologie ; Toy catalogue, catalogue de jouets ; Small-adds, petites annonces immobilières ; Itineraries, descriptions d'itinéraires

Maison avec piscine (petites annonces immobilières).

Il semble bien que seule cette visée joue un rôle dans le fonctionnement de *avec méronymique*. En effet, dans le manuel GEO (tableau 4.1), qui relève bien d'un domaine spécialisé (géomorphologie en l'occurrence), *avec* n'a pas souvent un fonctionnement méronymique alors que la relation partie à tout y est fréquente. Ce sont alors d'autres marqueurs de méronymie qui sont utilisés.

3.4.3 Importance de la visée et du domaine : chez

Tout comme *avec*, *chez* peut être associé à un sens méronymique dans certains énoncés (tableau 4.2). Mais cette interprétation n'est pas équiprobable selon les textes. Les textes les plus déclencheurs de l'interprétation méronymique sont ceux qui concernent les sciences naturelles et qui ont une visée didactique. En effet, lorsque le domaine relève bien des sciences naturelles mais que la visée n'est pas didactique, *chez* est peu associé à des énoncés méronymiques (25,2 % dans le tableau 4.2). De même, si la visée est didactique mais le domaine n'est pas celui des sciences naturelles, on obtient de mauvais résultats (0).

	EUbio	Manuelsbio	LMbio	EUculture
Nombre de mots	24,770	82,000	49,440	79,700
Chez	155 (0.63%)	91 (0.11%)	107 (0.22 %)	103 (0.13%)
Chez méronymiques	83 (53.6%)	48 (53 %)	27 (25.2 %)	0

Tableau 4.2 : Comportement de chez dans différents genres de corpus[3].

3.4.4 Domaine et marqueurs de relation

Il est évident que le domaine joue un rôle sur au moins deux points :

- Présence vs absence d'une relation (*est le symptôme* de, spécifique du domaine médical)

- Présence vs absence de marqueurs (*chez* n'apparaît pas dans le domaine minéral (c'est *dans* qui est choisi, *dans le diamant* …)

Il semble moins certain en revanche que le domaine seul ait une influence sur l'interprétation des marqueurs. Ainsi, Marshman et L'Homme [MAR 06] font une étude comparée des verbes marqueurs de cause issus de deux corpus de deux domaines différents : informatique et médical. Elles notent des similitudes de

[3] EUbio : Encyclopaedia Universalis, sciences naturelles, Manuelsbio : Manuels de biologie, LMbio : Journal Le Monde, domaine sciences naturelles, EUculture : Encyclopaedia Universalis domaines culturels.

fonctionnement comme la fréquence de certains verbes et de certaines interprétations. Mais il existe aussi de nombreuses divergences et les auteurs suggèrent que ces divergences soient dues au moins en partie au fait que les genres étudiés sont différents selon les domaines (articles de recherche pour le corpus médical, textes didactiques et de semi-vulgarisation dans le corpus informatique). Ces éléments, qui, pour nous relèvent plutôt de la visée, pourraient expliquer les variations d'un corpus à l'autre. Sous-jacente à cette question, se pose la nécessité de clarifier des notions comme genre, registres, visée, domaine, spécialité, types de textes ainsi que se propose de le faire Lee [Lee 01]. Les travaux sur le fonctionnement des marqueurs de relations pourraient contribuer à cette clarification.

3.5. Co-texte et/ou contexte

Le co-texte (la syntaxe) et le contexte (la situation) interviennent probablement dans la compréhension et particulièrement dans l'interprétation d'éléments langagiers comme marqueurs de relation. Dans la perspective d'un traitement automatique, il serait nécessaire de voir quelle peut être la meilleure façon de faire intervenir ces deux éléments. La prise en compte du contexte peut se faire relativement facilement par la caractérisation des textes. La difficulté tient au problème qu'il peut y avoir à définir ces caractéristiques. Il n'est pas certain qu'une liste figée de méta-données de caractérisation puisse être actuellement envisageable. Par ailleurs, l'étude du fonctionnement des marqueurs en lien avec cette caractérisation est loin d'avoir été faite pour tous les marqueurs, du point de vue linguistique. La prise en compte du co-texte est plus classiquement faite en TAL. Mais sa mise en œuvre peut être couteuse et il peut s'avérer plus efficace de laisser l'utilisateur trier les exemples proposés (pour éliminer le bruit), surtout dans les cas où la prise en compte de la situation (contexte) a permis de prédire que le nombre de cas pertinents sera très élevé (cas de *avec* par exemple).

4. Point de vue informatique : prise en compte de la variation dans les outils informatiques de recherche de relation par patrons

La recherche de relations conceptuelles dans des textes a été longtemps considérée en informatique sous l'angle de l'extraction d'information : il s'agit alors de rechercher des types d'information très précis, dont on peut fournir une caractérisation lexicale, syntaxique ou sémantique au système informatique. Cette approche fonctionne bien dans des collections de textes dont la rédaction présente des régularités et, à défaut, pour des collections volumineuses. Les patrons de fouille sont alors sophistiqués, l'objectif étant d'identifier le plus automatiquement et précisément possible de nombreuses occurrences des éléments recherchés. La

recherche de relations est aujourd'hui une problématique également au cœur du web sémantique et de l'ingénierie des ontologies [STA 01]. Elle intervient soit en phase de construction d'ontologies à partir de textes [MAE 00], soit en phase d'utilisation d'ontologie pour trouver des relations entre occurrences de concepts dans des textes, dans un objectif d'annotation [RUI 08]. Dans ce contexte, la formulation en discours des informations recherchées ne sont pas ou mal connues, et dont difficilement caractérisables dans un patron. Cependant, certains invariants (ou supposés tels), comme la présence systématique de relations hiérarchiques dans l'ontologie, autorise une recherche par patron sur la base de réutilisation / adaptation, ou bien d'apprentissage des patrons à partir de phrases validées manuellement. Plus récemment, la caractérisation de structures récurrentes dans les ontologies, également appelées patrons d'ontologies (*ontology patterns*), a conduit à focaliser les informations recherchées sur les équivalents linguistiques de ces *patterns* [AGU 08] ou encore sur des éléments plus complexes que les seules relations binaires.

Nous nous plaçons donc du point de vue de l'ingénierie des ontologies pour étudier les logiciels et approches existantes et juger de leur capacité à prendre en compte les phénomènes de variations identifiés au niveau linguistique et décrit en partie 3. Nous rappelons d'abord les étapes classiques de la recherche de relation par patron, les types d'outils disponibles aujourd'hui pour la réaliser et les caractéristiques que nous avons retenues pour en discuter les capacités à gérer des variations. Nous détaillons ensuite chacune de ces étapes pour revenir sur les liens entre les techniques possibles dans les systèmes d'aide à la recherche de relations et leur capacité à prendre en compte la variation.

4.1. Logiciels supports à la recherche de relations par patron

4.1.1 Etapes principales de la recherche de relation par patrons

La plupart des systèmes d'aide à l'extraction de relations supposent un processus organisé selon les étapes suivantes :

- *traitement automatique préliminaire du corpus* : il s'agit de prévoir les traitements de bas niveaux classiques (repérage de frontières de mots et de phrases, de syntagmes, analyse des catégories grammaticales, recherche de mots ou d'entités spécifiques, lemmatisation ou radicalisation etc.) nécessaires aux traitements qui suivront ;
- *mise au point des patrons* : il s'agit de définir des patrons adaptés à la nature des relations recherchées, au corpus et au format requis par le logiciel de projection ;
- *projection des patrons ou recherche d'instances* : le système recherche en corpus toutes les occurrences du patron et retourne les phrases les contenant ;

- *interprétation des résultats au niveau terminologique* : cette interprétation porte sur chacune des occurrences du patron, et consiste à la valider ou non, à identifier la nature de la relation conceptuelle et les termes mis en relation ; cette phase est parfois sautée pour aller directement vers la suivante ;
- *enrichissement d'un modèle de connaissances* (« ontologisation » de la relation) : il s'agit de passer de l'interprétation terminologique à une interprétation conceptuelle : décider de représenter la relation terminologique sous forme d'une relation ou d'une instance de relation, fixer le nom (le type) de cette relation, choisir les concepts (son domaine et son co-domaine) ou les instances de concepts qu'elle relie.

4.1.2 Différents outils d'aide à la recherche de relations

Notre étude s'appuie sur trois états de l'art récents [AUG 08] [HAL 08] (tous deux tirés du numéro spécial de *Terminology* sur les approches par patron pour l'extraction de relations) et [STA 01] [PAN 08], plus orientés vers l'utilisation de techniques d'apprentissage pour automatiser l'identification des relations. Les premiers auteurs mentionnent une douzaine de travaux de ce type qui se sont intéressés à l'évaluation de leurs résultats. Les seconds comparent 11 outils de ce type en évaluant leur portabilité, leur capacité à apprendre de nouveaux patrons et à traiter des relations autres que les relations hiérarchiques. Près de dix d'approches se basant sur l'apprentissage sont répertoriées dans [PAN 08], pour des applications allant de systèmes question-réponse à l'annotation sémantique de documents. Nous distinguons quatre types d'outils pour assister la recherche de relations :

- *des plates-formes générales d'analyse de texte* : Gate[4], LinguaStream[5], Alvis[6] (cf. Chapitre 7). Conçues initialement pour faciliter la définition de chaînes de traitements automatique du langage à différents niveaux dans des textes, ces plates-formes bénéficient d'un regain d'intérêt dans le cadre de la construction d'ontologies à partir de textes. Dans des systèmes comme ANNIC défini à l'aide de Gate, les traitements aident à identifier des concepts et des relations [MAY 08].
- *des outils indépendants d'extraction de relations :* Prométhée [MOR 99], Caméléon [SEG 01], RelExt [SCU 05], ou d'extraction de termes et de relations : le système d'extraction de taxonomies de [BAR 04], Terminoweb [BAR 06], Expresso [PAN 06], Snowball [AGI 00] …
- *des environnements généraux de construction d'ontologies* à partir de textes (souvent associé à une méthode) qui intègrent une aide à la recherche

[4] http://gate.ac.uk
[5] http://www.linguastream.org/
[6] http://www-lipn.univ-paris13.fr/~hamon/PlateformeTAL-ALVIS/index.html

de relations, comme Text-To-Onto [MAE 00] [STA 01], OntoLearn [VEL 06] ou la plate-forme Terminae [AUS 08b] avec son module Linguae.

- *des logiciels d'extraction de relations très spécifiques* à certains domaines de connaissances et aux corpus relatifs à ces domaines. Un des domaines les plus explorés est la bioinformatique, avec des systèmes comme PASTA [GAI 03], RelationAnnotator [MUK 06] ou les travaux de Khélif [KEH 07], de Nédellec [NED 04], de Sheth [RAM 08].

Selon leur nature, ces logiciels ne couvrent pas toutes les phases du cycle que nous avons identifié, et assistent ou automatisent ces tâches à des degrés divers. La phase systématiquement traitée est la projection des patrons. Pour chaque étape, nous ferons un inventaire des aides apportées.

Le degré d'automatisation de chaque étape est d'autant plus élevé que le système est spécialisé pour trouver un type particulier de relations ou pour traiter des corpus de forme régulière dans un domaine bien cerné. Le fait d'automatiser des phases de la recherche de relation ne préjuge en rien de la prise en compte de la variation. Par exemple, certains systèmes favorisent l'apprentissage de manière à s'adapter au mieux aux spécificités d'un domaine ou d'un corpus, et en cela, ils sont pertinents par rapport à la variation de forme et de pertinence des patrons d'un domaine à l'autre. En revanche, cette automatisation fait souvent l'hypothèse qu'un patron appris aura ensuite une interprétation stable sur tout le corpus, c'est-à-dire que toutes les occurrences retournées rendront compte du même type de relation.

4.1.3 Eléments de comparaison des logiciels à chaque étape du processus

Nous retenons cinq caractéristiques qui vont nous permettre de discuter de la prise en compte des variations identifiées par l'étude linguistique par les logiciels de recherche de relation à chaque étape :

- *stabilité* : hypothèse de stabilité implicite ou explicite justifiant le fonctionnement du logiciel à cette étape ;
- *variation* : éléments de variation potentiels, bloqués, favorisés et/ou pris en compte ;
- *interprétation* : place laissée (ou non) à l'interprétation des résultats de l'étape par un humain, nécessité ou non de valider les résultats obtenus ;
- *compétence* : compétence requise pour utiliser logiciel à cette étape, et en interpréter / valider les résultats
- *aide / rôle du logiciel* : aide apportée à l'utilisateur / part de la tâche prise en charge par le logiciel.

Nous récapitulons sous forme de tableau l'évaluation de ces caractéristiques pour les logiciels que nous avons étudiés (tableau 4.3).

Logiciel	Stabilité	Variation	Base de patrons réutilisables	Aide à la mise en forme de patrons	Evaluation de la qualité des phrases	Interprétation Recherche des termes reliés	Interprétation Proposition de type relation	Interprétation Proposition de concepts
Caméléon [SEG 01] [JAC 06]	Association patron / type de relation	Adaptation des patrons aux corpus Vérification de la sémantique de chaque relation	75 patrons ; une 15 aine de type de relations	Proposition de patrons à adapter Interface simplifiant la saisie des parties	Manuelle	Suggestions tirées d'un extracteur de termes	Type associé au patron Peut être remis en question	Concepts associés aux termes Concepts gérés dans ontologie à composante lexicale
Expresso [PAN 06]	Patrons trouvés sur internet et associés à des relations fixes	Certains patrons sont propres à chaque corpus	Patrons généraux avec un fort rappel et une faible précision Relations IS-A et autres	Patrons abstraits à partir des occurrences des termes en relation + évaluation de la fiabilité	Classées par degré de fiabilité (analogue et liée à celui des patrons)	Catégories associées aux termes Domaine biologie	5 types de relations étudiés	s'arrête aux relations entre termes (assimilables à des concepts)
OntoLearn [VEL 06]	Type de relation et leurs différentes formulations linguistiques	Checkers (patrons) ad hoc pour le corpus considéré.	10 checkers pour 10 types de relation	Pas de possibilité de définir de nouveau patron		Exploite les termes d'un thésaurus. Compare les relations à celles du thésaurus	Type associé au patron. Automatique et fixe.	incorpore des catégories sémantiques WordNet (signature de relation connue)
Prométhée [MOR 99]	Régularité des formes des relations au sein d'un corpus	D'un domaine (corpus) à l'autre : nouveaux patrons	Patrons pour la relation EST-UN	Apprentissage		oui	Fixé pour le patron au moment de l'apprentissage	Reste au niveau des termes
RelExt [SCU 05]	Types de relation et leur forme par domaine	Patrons par type de relation	Aucun	Apprentissage à partir de termes spécifiques au corpus et de patrons verbe-objet appris	Ne s'applique pas	Termes sélectionnés par critères statistiques sur corpus étiquetés syntaxiquement	Fonction du verbe reconnu	Termes regroupés en classes selon critères statistiques
Relation-Annotator [MUK 06]	Termes désignant les concepts recherchés Types des relations UMLS Liée au domaine	Différents patrons par type de relation	non	Mise au point manuelle de patrons Apprentissage à partir de couples de termes en relation	Qualité des instances de relation trouvées Précision	Termes associés aux concepts recherchés	Un des types dans UMLS, en fonction du patron retrouvé	Fixés a priori
SnowBall [AGI 00]	Termes désignant les concepts recherchés (listes d'entités nommées)	Multiples formes pour le type de relation recherché	non	Apprentissage à partir de couples de termes en relation, évalue la qualité du patron (confiance)	Qualité des instances de relation trouvées	Entités nommées associées aux catégories sémantiques recherchées	fixe	Fixés a priori (localisation – organization)

Termino-Web [BAR 06]	Association patron / type de relation	Possibilité de définir de nouveaux patrons et de nouveaux types de relation adaptés à un corpus et une modélisation	9 types de relations et près de 250 patrons pour l'anglais	Très réduite et très simple : éditeur 1 ligne, patron = forme exacte recherchée	Phrases = context snippets, validation manuelle	Extraction de termes indépendante Repérage de « knowledge rich contexts »	TypeRel fixé au moment de définer le patron	Aucune proposition de relation
Text-ToOnto [MAE 00] [STA 01]	Forme d'expression des relations	Type de relations non fixes liés au corpus	non	Règles d'association apprises à partir du corpus	Ne s'applique pas	Termes reliés aux concepts	Type relation non interprété (à ajouter manuellement)	Propose les « meilleurs » concepts dans hiérarchie
WWW2rel [HAL 08]	Domaine et corpus Association patron / type de relation liée au domaine et au corpus	Patrons adaptés de corpus tirés du web, types de relation propres au corpus Dépendance patrons - domaine	Test de portabilité de patrons dans d'autres domaines (synonymie, est-un, …)	Apprentissage des patrons	Précision F-score par rapport à corpus annotés à la main. Classe les instances trouvées par "knowledge pattern range".	Recherche de relations entre instances	Type de relation associé au patron	Ne s'applique pas

Tableau 4.3 : place de l'interprétation dans plusieurs systèmes de projection de patrons

4.2. Mise au point des patrons

4.2.1 Spécificités de la mise au point des patrons

Cette étape suppose une *stabilité* des patrons lorsqu'il est fait usage de patrons dits « génériques », caractérisant toujours les mêmes relations, et ce au moins dans le corpus à l'étude, au plus quels que soient le domaine et les corpus. Elle requiert de l'utilisateur une *compétence* minimale en syntaxe et sémantique, et si possible en traitement automatique des langues et informatique pour anticiper comment le système va les utiliser. La plupart des recherches pour outiller cette étape constatent rapidement que l'hypothèse de *stabilité* est simplificatrice : les mêmes patrons sont plus ou moins fiables, plus ou moins productifs, et renvoient à des types de relations qui peuvent varier. Au moment de la mise au point, prendre en compte la variabilité suppose d'évaluer, adapter ou rejeter des patrons. *L'interprétation* humaine intervient donc à plusieurs reprises : pour abstraire, à partir de phrases du corpus, une caractérisation qui sera le patron, pour identifier le type de relation en présence à partir des phrases retournées par les patrons, enfin pour décider de retenir, adapter ou rejeter un patron. Pour simplifier cette tâche, le système peut fournir une base de patrons « génériques » ou « adaptables », des indices de fiabilité sur ces patrons ou des traces de leur comportement observé sur différents corpus (fréquence d'apparition, rappel et précision) ; il peut proposer une interface dédiée guidant la mise en forme des patrons, suggérer les termes en relation pour guider l'interprétation des projections ; le système peut aussi automatiser tout ou partie de la mise au point en procédant par apprentissage de nouveaux patrons à partir d'exemples annotés manuellement.

4.2.2 Forme des patrons et lien avec leur mise au point

Il est important de repérer la diversité de forme qui se cache derrière l'appellation de patron. A chaque forme correspondent des hypothèses différentes de variabilité et de stabilité, une mise au point plus ou moins simple et un mode de projection particulier sur les corpus. Nous listons ici quelques unes de ces formes :

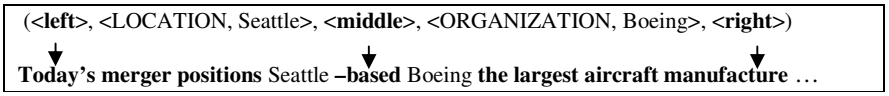
***mot1* [+ *mot2*]** : Une ou plusieurs formes lexicales complètes ou tronquées pour retrouver des variantes de forme : ex : *chez* ou *chez* 'en début de phrase' [SIE 08] [FEL 04]. Ce format suppose de multiplier l'inventaire des mots pouvant se trouver dans un patron sans véritable caractérisation. La mise au point est accessible à des non spécialistes de la langue, mais leur efficacité (rappel et précision) peut être difficile à optimiser.

A term1 B term2 C : Dans cette forme utilisée dans Caméléon [SEG 01], l'approche de Marshman [MAR 07] ou TerminoWeb [BAR 06], term1 et term2 sont les termes recherchés en relation (ce qui suppose de produire une caractérisation syntaxique ou sémantique des termes recherchés), et A, B et C caractérisent le contexte autour des termes. Les éléments composant le patron peuvent être des catégories grammaticales ou sémantiques, des formes lexicales exactes ou lemmatisées, des éléments « indéterminés », ces éléments pouvant être optionnels, obligatoires ou répétés, etc. Beaucoup de patrons simplifient ce modèle en éliminant A et C, ou même en ne caractérisant ni term1 ni term2, comme les patrons du système Expresso [PAN 06] ou ceux appris par Snowball qui ont la forme suivante : **ENTRY**/NN is/VBZ a/DT type/NN of/IN **TARGET** où ENTRY renvoie à term1 et TARGET à term2, NN à un syntagme nominal, DT à un déterminant, VBZ à un verbe et IN à une préposition. Enfin, comme [BYR 99], on peut simplifier la recherche en fixant une catégorie sémantique connue pour term1 et/ou term2, en général des entités nommées faciles à identifier.

Dans ces deux premières formes, tous les éléments du patron sont supposés d'égale importance. Ils sont recherchés dans l'ordre où ils apparaissent. La mise au point peut être fastidieuse car tous les éléments sont cruciaux pour retrouver les bonnes phrases et toutes les phrases à reconnaître. Ainsi, pour anticiper tous les cas de figure dus aux instabilités des triplets terme/marqueur/terme mentionnés aux 3.1, il faut anticiper toutes les variantes possibles a priori et définir autant de patrons.

Or d'autres travaux considèrent qu'un patron comporte des éléments clés « le cœur du patron » et des éléments plus secondaires. L'hypothèse de stabilité est plus forte pour le cœur du patron (caractéristique d'un type de relation) tandis que la mise au point concerne les éléments secondaires, plus variables, qui influencent l'efficacité opératoire, mais ne changent pas la nature du patron ni la sémantique de la relation. Les trois exemples qui suivent illustrent ce choix.

LEFT term1 MIDDLE term2 RIGHT : Dans SnowBall [AGI 00], les patrons ont une forme analogue aux précédents, mais l'élément séparant les termes en relation, appelé *middle,* est jugé plus important que les éléments droit et gauche.



Exemple 4.1. Un patron Snowball et une phrase qu'il permet de reconnaître

De plus, le système procède en deux temps : il commence par la reconnaissance de catégories sémantiques (LOCATION et ORGANIZATION dans l'exemple 4.1) à l'aide de jeux de règles propres à ces catégories, puis les patrons sont appris / appliqués en s'appuyant sur l'élément central (middle).

LEFT VERBE_DEF NEXUS RIGHT : Pour étudier en espagnol des patrons de définition construits à partir de verbes, Sierra distingue les parties à gauche (LEFT) et à droite (RIGHT) du syntagme composé du verbe (VERBE-DEF) et éventuellement d'une préposition le modifiant (NEXUS) [SIE 08]. Ici encore, les mots situés *entre* les termes, et particulièrement un verbe, sont supposés plus importants que les autres éléments de caractérisation du contexte.

SI Indicateur reconnu ALORS (SI Contextes reconnus ALORS conclure RELATION ou RELATION(Arg1,Arg2)) : Dans l'approche initiée par J.P. Desclés [DES 91] et mise en œuvre entre autres dans Coatis [GAR 98], les travaux de Jackiewicz [JAC 96] et ContextO [MIN 02], un patron correspond à une règle (cf. exemple 4.2) qui distingue la recherche d'un *indice déclencheur* d'abord, et d'un *contexte* confirmant cet indice dans un 2e temps. Ce découpage facilite la mise au point du patron et rend plus efficace sa projection. A priori, l'indice déclencheur est relativement stable, il s'agit souvent d'une catégorie sémantique de verbes par exemple, alors que la caractérisation du contexte peut varier sensiblement. Le contexte peut être affiné dans un 2e temps pour réduire le bruit, l'essentiel étant de bien repérer tous les indices déclencheurs. De plus, la position de cet indice par rapport aux termes est variable (pas d'a priori sur un rôle plus important du « milieu »). Cela autorise des patrons souples, et permet de gérer l'instabilité du triplet terme/marqueur/terme mentionnée en 3.1.

Indicateur : VerbePrésentatif
si Contexte 10 mots autour du verbe contient un indice *partie document3*
si Contexte 10 mots autour du verbe contient un indice de *partie document1*
alors reconnaître un énoncé de thématique

avec partie document3 : « dans les lignes » « dans le texte » « dans la partie » …

avec partie document1 : « qui suivent » « ci-dessous » « précédent » …

phrase ainsi reconnue : <Dans les lignes> <qui suivent>, … nous présentons …

Exemple 4.2. Adaptation d'une règle d'exploration contextuelle [MIN 02]

4.2.3 Disponibilité de banques ou bases de patrons

Les patrons les plus mentionnés dans la littérature sont les patrons de Hearst [HEA 92]. Plusieurs approches font l'hypothèse qu'ils sont universels et renvoient systématiquement à une relation hiérarchique entre classe et sous-classe. Il est clair que la mise au point de nouveaux patrons est coûteuse et que la réutilisation de patrons existants est souhaitée par tous ceux qui construisent des modèles à partir de texte. Cependant, une réutilisation immédiate semble utopique. Comme cela est prévu depuis les premières versions de Caméléon [SEG 01] et comme l'a confirmé l'expérience rapportée dans [JAC 06] [AUS 08c], il est plus judicieux de prévoir d'adapter des patrons disponibles, de les considérer comme des points de départ et

non comme des patrons génériques. Dans ce cas, toute information sur les résultats obtenus sur d'autres corpus est une aide au choix et à l'interprétation des patrons. Plusieurs tentatives de capitalisation ont vu le jour depuis les dix dernières années. On trouve des listes de patrons plus ou moins formalisés dans des articles et thèses, par exemple, pour l'anglais, le français ou l'espagnol. Il est plus rare que les patrons soient implémentés dans des logiciels comme le proposent Caméléon [SEG 01] [AUS 08c] pour le français, Terminoweb [BAR 06] pour l'anglais ou Skeleton [FEL 04] pour le catalan.

4.2.4 Fiabilité et qualité des patrons

Pour [PAN 08], un patron fiable est un patron attesté sur un gros volume de documents, ayant éventuellement un rappel faible mais surtout une précision élevée. Leur mesure de fiabilité d'un patron s'appuie sur la « qualité » des paires d'instances retrouvées dans ce corpus, calculée selon le potentiel d'information mutuelle existant entre les termes reliés. Cette fiabilité tend à privilégier des patrons très précis, et conduit à définir de nombreuses variantes pour une même relation. Les auteurs s'en servent pour juger de la pertinence de patrons génériques à adapter au corpus. D'autres mesures de ce type sont proposées dans [MAE 00] et [MAE 02].

[HAL 08] s'appuie sur des mesures comme le Kappa pour évaluer les relations trouvées, en vérifiant que les termes identifiées ont une corrélation plus forte que s'ils apparaissaient par hasard ensemble. Ces métriques permettent aussi de juger de la qualité de patrons appris à partir du corpus, au moyen des phrases qu'ils retournent. C'est dans le même esprit que sont évalués les patrons de Caméléon. Pour [JAC 06], la qualité et la fiabilité d'un patron ne sont pas intrinsèques, mais bien liées à un corpus. Leur expérience souligne les fortes variations de rappel et de précision d'un même patron d'un corpus ou d'un genre à un autre. Les patrons sont documentés par leur rappel et leur précision sur différents corpus ce qui constitue des indices de leur pertinence pour fouiller un nouveau corpus.

4.2.5 Apprentissage de patrons par induction

L'article [PAN 08] illustre bien les ambitions d'automatisation du processus d'identification des relations autant que des patrons, et présente un éventail de techniques qui permettent d'assurer ces processus. L'étape clé à automatiser est alors appelée « induce recurrent patterns ». Un des premiers travaux en la matière est le système ASIUM [FAU 98] qui classe automatiquement les contextes d'apparition de verbes pour trouver des classes de relations. A la base, on trouve presque systématiquement l'algorithme DIPRE de Bri décrit dans [HAL 08], réutilisé et adapté dans Prométhée [MOR 99], dans WWW2rel [HAL 08], par [RUI 07] ou [BLO 07]. Ce procédé suppose de disposer de ressources minimales pour amorcer le cycle d'apprentissage : connaître des couples de concepts en relation et

les termes désignant ces concepts. Le processus se décompose alors en (1) extraction de patrons : formulation de patrons à partir des contextes où apparaissent les termes en relation, (2) généralisation des patrons jugés proches par abstraction des éléments qui les composent, (3) évaluation des patrons ainsi obtenus. A partir de phrases contenant des termes désignant des concepts en relation, Expresso infère un patron formé d'éléments grammaticaux de surface qui généralisent le plus grand nombre d'instances [PAN 06]. Une approche similaire est à la base de [RUI 07] ou de Snowball [AGI 00]. Une évolution récente d'Expresso [PAN 08] propose d'aller plus loin en utilisant ensuite un algorithme de « clustering » pour abstraire des classes en relation, à partir des termes en relation dans les phrases.

Les avantages de cet apprentissage sont un gain de temps et une capacité à traiter de gros volumes de documents. Les limites sont avant tout la qualité et la faible généricité des patrons appris. Ils restent proches des formes exactes trouvées dans les textes. Les résultats sont de meilleure qualité en extraction d'information, dans des domaines très spécialisés, avec de gros volumes de textes écrits selon un style très réguliers et disponibles sur support numérique. Un compromis consiste à envisager une aide à l'abstraction des composants de patron pas à pas, sous la forme de propositions à valider à la demande. C'est la solution envisagée pour faire évoluer Caméléon et simplifier la définition de nouveaux patrons [AUS 08a] à partir de phrases contenant des termes en relation.

4.3. Projection des patrons

4.3.1 Spécificités de la projection des patrons

Cette étape est généralement automatique, l'utilisateur n'a pas de maîtrise dans la plupart des systèmes (donc pas d'interprétation). La manière dont les patrons sont projetés est pourtant déterminante sur la qualité de leurs résultats et le temps nécessaire à cette tâche. Comprendre l'algorithme de projection facilite également une mise au point adaptée ou optimale des patrons. Or dans la plupart des systèmes, l'algorithme utilisé est peu décrit, et c'est par la pratique que l'utilisateur comprend les impacts de la structure ou du contenu des patrons sur leur efficience.

4.3.2 Modalités de la projection des patrons et enjeux pour la variation

La projection des patrons peut répondre à trois types d'objectifs : extraction d'information, construction d'ontologie, peuplement d'ontologie. Elle va se réaliser différemment dans chacun des contextes. Plus on considère les patrons comme stables, plus la projection peut être systématique et linéaire. La projection varie dans la manière et la nature des éléments recherchés. Dans les trois contextes, on trouve les types d'algorithmes suivants :

Projection linéaire : Le cas le plus simple considère les éléments constituant le patron de manière équivalente et dans l'ordre où ils apparaissent : le texte est exploré de manière linéaire pour rechercher une occurrence du premier élément du patron, et, s'il est reconnu, des éléments suivants selon le même principe. Cette approche, retenue par exemple dans Caméléon et Prométhée, permet à l'utilisateur de comprendre aisément le fonctionnement des patrons. Elle est efficace pour rechercher des lemmes ou des patrons privilégiant les verbes comme ceux de [SIE 08] ou [SOL 08]. En revanche, elle s'avère peu performante dès que l'on autorise en début de patron des catégories grammaticales aussi usuelles qu'un déterminant. Le patron est « activé » (au sens où l'on va recherche l'élément suivant qui le compose) dans presque toutes les phrases, sans pour autant être reconnu.

Restriction de la recherche sur des phrases ciblées : Une première manière de focaliser la recherche est de ne s'intéresser qu'à des portions de textes dont on sait, par d'autres indices, qu'elles ont plus de chances de contenir une relation. Par exemple, le système RelExt recherche les relations dans des phrases contenant des couples de termes qui co-occurrent fréquemment en en collocation [SCH 05]. Lorsqu'une ressource (ontologie ou terminologie) fournit des termes déjà en relation, une manière de les repérer dans les textes est de projeter d'abord les couples de termes, et de ne chercher les patrons de relation que dans les contextes contenant ces couples. C'est la suggestion retenue par Séguéla [SEG 01] ou par Chagnoux [AUS 08a] pour retrouver de nouvelles occurrences de relations.

Recherche en priorité de certaines parties du patron : Le fait de distinguer, dans un patron, différentes parties en fonction de leur rôle par rapport à l'identification d'une relation et des termes reliés, permet de définir des stratégies de recherche plus optimales. Lorsqu'il est présent, c'est souvent le verbe qui est alors considéré comme élément privilégié.

Avec des patrons de la forme LEFT VERBE-DEF NEXUS RIGHT [SIE 08], la projection correspond à une recherche focalisée en premier lieu sur le verbe VERBE-DEF, puis sur l'élimination des contextes jugés non valides selon des règles portant sur le NEXUS en priorité, puis la structure des parties LEFT et RIGHT. L'identification des éléments mis en relation (LEFT et RIGHT) se fait à l'aide d'expressions régulières caractéristiques de termes. Selon un principe analogue, et pour l'anglais, la recherche d'instances des patrons définis par Agichtein [AGI 00] porte en priorité sur l'élément appelé *middle,* et situé entre les termes en relation.

Dans le cas de l'exploration contextuelle, la projection consiste à projeter les règles d'exploration, à tester d'abord les conditions portant sur l'indice déclencheur (focus) avant de tester les éléments du contexte et d'identifier les termes en relation. Il existe différentes opérationnalisations de cette approche, dont certaines sont récapitulées dans [MIN 02]. Cette exploration accélère le balayage du corpus à la

recherche d'une instance, et autorise des variations dans le contexte syntaxique des termes mis en relation.

4.3.3 Limites de l'informatique à bien gérer la projection des patrons

La manière dont les logiciels d'extraction de relations gèrent cette étape et cherchent à optimiser la mise au point des patrons traduit une certaine ignorance des phénomènes linguistiques et en particulier de la variation.

En particulier, l'efficacité des patrons se mesure en termes de rappel (capacité à trouver toutes les occurrences du patron en corpus) et précision (capacité à trouver des instances valides, contenant bien une relation du type attendu). Or la recherche d'une efficacité de rappel et précision conduit à multiplier les patrons pour une même « forme canonique linguistique » pour parvenir à un optimum opérationnel. Ceci conduit à un paradoxe, qui est de définir des patrons plus précis et moins génériques alors qu'un patron se veut une abstraction.

Cette évaluation purement quantitative escamote des éléments sémantiques plus fins : un « bon patron » est souvent retenu parce qu'il est un patron productif et parce qu'il rend compte de la relation avec fiabilité. On trouve par exemple comme critère pour retenir des patrons appris qu'ils produisent au moins 3 occurrences sur le corpus étudié ou que leur rappel soit au dessus d'un seuil fixé. Ensuite, une approche quantitative n'invite pas à capitaliser l'impact sur ses performances de la nature des éléments d'un patron modifiés lors de sa mise au point. Cette mauvaise maîtrise conduit à tâtonner en fonction du but recherché (trouver des instances du patron) et non par l'analyse de la langue.

4.4. Interprétation des résultats au niveau terminologique

L'interprétation des phrases retournées par un patron prend un statut différent durant la mise au point du patron et en phase d'utilisation du patron pour un objectif de modélisation conceptuelle. Nous nous plaçons ici dans le second cas, après que l'on ait fait l'hypothèse, durant sa mise au point, que le patron rendait compte d'un certain type de relation. Après la projection, on s'attend alors à trouver dans chaque phrase retournée les termes mis en relation soit là où le patron les spécifie, soit autour du patron, et a priori dans la même phrase.

Le fait de fixer la structure et la position des termes en relation rend le patron « rigide » et réduit les possibilités d'anticiper des formes de phrase non prévues. Une première difficulté est donc de trouver un compromis entre l'aide à l'interprétation (en fixant la caractérisation des termes) et la capacité du patron à anticiper des variations mineures dans les phrases. De plus, la variabilité du sens de

la relation révélée par le patron au sein d'un même corpus oblige à vérifier chacune des occurrences du patron. Pour simplifier cette tâche, certains systèmes filtrent les phrases retournées en fonction de leur qualité ou de leur fiabilité, calculée selon des critères numériques [HAL 08] [BAR 06] [MAE 02] ; ils peuvent suggérer un type de relation pour cette occurrence et parfois mettre en évidence les termes identifiés en relation. TextToOnto exploite la hiérarchie des concepts pour proposer une relation au « meilleur » niveau d'abstraction selon les relations lexicales trouvées [MAE 02]. D'autres systèmes projettent les résultats d'un extracteur de termes, comme Caméléon (qui utilisait les résultats de Lexter), OntoLearn ou des termes extraits par le même système comme TerminoWeb. Enfin, dans un contexte d'extraction d'information [SCU 05] [MUK 06], les patrons sont parfois projetés sur des phrases contenant deux occurrences de termes associés à des concepts fixés. Les termes mis en relation sont donc des noms possibles d'instances des concepts ou des termes désignant ces concepts.

[JAC 08] montre que la difficulté est plus grande lorsqu'on explore le web, où les variations lexicales sont plus nombreuses que dans des corpus spécialisés : on sort d'une approche terminologique pour travailler sur la langue générale. La difficulté à associer les bons concepts aux termes rend difficile le calcul de critères de qualité, le rappel ou la précision. En sélectionnant les termes retrouvés selon des critères numériques, certains systèmes automatisent la définition de relations conceptuelles et s'affranchissent de l'évaluation de relations lexicales, comme dans TextToOnto. Cette option radicale écarte toute possibilité de variation sémantique dans l'interprétation des phrases retournées par le patron.

4.5. Enrichissement d'un modèle de connaissances

Pour atteindre le niveau conceptuel et identifier une relation conceptuelle, l'approche par patron suppose soit de fixer des catégories sémantiques parmi les éléments des patrons (et de connaître les termes rattachés à ces catégories), soit de mener une deuxième interprétation des phrases retournées. Dans ce deuxième cas, une fois identifiée une relation lexicale, le choix des concepts en relation peut être automatisé si la relation terme-concept est stable et univoque ou si on implémente un algorithme de désambiguïsation des termes en cas de polysémie. Dans le premier cas, l'identification des concepts associés aux termes se fait en amont, soit parce que le corpus a été annoté à l'aide de catégories sémantiques, soit en projetant des listes de termes connus associés à des concepts. La difficulté se situe alors dans les phrases utilisant des termes nouveaux, qu'il est difficile d'associer à des concepts. Poser la relation entre concepts suppose également de fixer la nature de la relation conceptuelle associée au patron. Or l'efficacité de cette étape se heurte à la polysémie de la relation d'une part, et à la diversité de la place des termes autour du patron lui-même, dont la localisation n'est pas toujours stable. A ce stade, la plupart

des approches d'extraction de relation font l'hypothèse que le type de la relation est celui fixé pour le patron reconnu, et peu d'outils autorisent de le remettre en question pour un couple de concept donné. Or la variation existe aussi à ce niveau. La plupart des systèmes font l'hypothèse que des ajustements pourront être réalisés sur l'ontologie construite dans un éditeur d'ontologie.

Les *compétences* en jeu sont plus des connaissances du domaine et un savoir-faire en matière de modélisation de connaissances et de langage de représentation des connaissances. La décision de modélisation doit dépasser la disparité de niveau entre l'expression trouvée dans le texte et celui auquel doit se situer la relation dans le modèle. Il est difficile d'établir des règles générales en dehors de contextes très contraints comme en extraction d'information, et il est souvent nécessaire de procéder au cas par car pour trouver les « bons concepts » à mettre en relation. TextToOnto [MAE 02] guide l'identification des « bons » concepts en relation à partir de plusieurs relations entre instances, en se basant sur la place de ces instances dans la hiérarchie des concepts. Les auteurs insistent sur la nécessité d'une intervention humaine à cette étape, et cela se retrouve dans de nombreux travaux [AUS 08] [BAR 06]. Suivant les systèmes, les relations conceptuelles trouvées ainsi viennent alimenter une base de données, enrichir ou peupler une ontologie, ou encore constituer des annotations sémantiques des textes analysés. L'approche peut consister à procéder soit progressivement, en validant les résultats phrase par phrase approches, soit d'un coup, en générant automatiquement une partie de modèle à l'aide de l'ensemble des relations identifiées.

5. Conclusion et perspectives

La question du repérage des relations sémantiques en corpus selon une approche par patrons constitue un point d'entrée particulièrement intéressant sur les collaborations possibles entre analyses linguistiques et analyses automatiques. Bien que semblant relever d'un objet unique, cette question demande de s'interroger sur les complémentarités, le renouvellement mais aussi les irréductibilités entre les deux approches. En effet, elles relèvent de deux paradigmes différents, avec des histoires et des objectifs ciblés. Ainsi, la question des marqueurs de relation s'inscrit, pour la linguistique, dans celle du fonctionnement des discours, de la référence, de la sémantique textuelle, des rapports entre discours et situation. Pour l'informatique, l'utilisation de marqueurs s'intègre dans différents types d'applications (extraction d'information, systèmes question-réponse, annotation sémantique ou construction d'ontologies), alors que la mise au point de cette approche constitue une problématique riche, impliquant différents niveaux d'analyse de langue, et pouvant faire appel à des procédés plus ou moins automatiques. Elle soulève des questions fondamentales sur l'automatisation des chaînes de traitement, la validation des résultats intermédiaires et finaux, ou encore la généricité et la réutilisabilité.

La linguistique s'attache à décrire le fonctionnement des marqueurs en prenant en compte la variation. La rencontre avec l'informatique, et en particulier le Traitement Automatique de la Langue, intervient au moins en trois points. Tout d'abord, les outils mis à disposition des linguistes permettent d'accéder plus rapidement à certaines régularités voire à les faire émerger (par exemple, apparition fréquente de tel « marqueur » dans tel ou tel texte). Un autre point concerne les besoins de l'informatique pour définir des patrons réutilisables ; ce besoin nécessite de la part des linguistes de proposer des descriptions qui correspondent au bon niveau de réutilisabilité. Par exemple, ils peuvent s'interroger sur la pertinence de décrire finement les déterminants à retenir dans un patron ou bien ignorer la nécessaire présence de ce déterminant qui relève du fonctionnement syntaxique de la langue (en général, un nom est précédé d'un déterminant). Enfin, le fait que l'informatique vise une utilisation réelle peut intervenir aussi dans la description puisque le choix et l'interprétation des portions de texte comme étant ou non des marqueurs de relation peut tenir compte non seulement du fonctionnement dans les textes mais aussi de l'objectif visé par la construction de la ressource. Ces éléments ne sont pas du tout négligeables même dans la perspective d'une analyse linguistique et peuvent interpeler une linguistique descriptive plus classique mais peut être plus éloignée du fonctionnement réel en corpus. Toutefois, les possibilités de la mise en œuvre des éléments descriptifs mis au jour par les linguistes peuvent aussi se heurter aux limites des outils, ce qui peut avoir un effet de limite artificielle parce que seulement technique, qui ne correspond pas à la réalité des fonctionnements langagiers.

L'automatisation de la projection de patrons bénéficie bien sûr des descriptions linguistiques pour mieux en caractériser la forme autant que la sémantique. En revanche, l'analyse informatique a du mal à intégrer toutes les nuances et finesses que le linguiste introduit au moment de fixer le comportement de ces patrons et d'en interpréter les résultats. En cherchant à automatiser le processus d'exploitation des phrases retournées par les patrons, que ce soit pour enrichir un modèle conceptuel, l'instancier, retrouver des informations précises ou organiser des données, l'analyse informatique est obligée de simplifier les hypothèses linguistiques, de supposer une forte stabilité, et de combiner d'autres critères linguistiques pour repérer les termes en relation. Au-delà des problèmes d'interprétation, l'analyse informatique soulève des difficultés propres liées à la projection des patrons ou l'évaluation des résultats.

Nous avons vu qu'il était souhaitable et possible de donner un nouveau souffle à ces travaux en révisant le processus dans le but de mieux prendre en compte des phénomènes de variation. Une première perspective consiste à différencier des éléments centraux et forts du patron par rapport à des éléments variables et plus optionnels, et d'en tirer des conséquences au moment de projeter les patrons. Ceci revient à développer et implémenter la notion d'indice comme alternative à celle de marqueur : au lieu de rechercher une structure figée et complexe dont on doit

spécifier toutes les variantes acceptables, il s'agit de caractériser des indices premiers, à rechercher en priorité, puis des éléments à tester dans un deuxième temps, des indices supplémentaires qui viennent soit confirmer la présence d'un patron, soit nuancer la signification de la relation qu'il révèle, soit encore faciliter le repérage des éléments qu'il met en relation. Cette notion d'indice est prometteuse, tant au niveau linguistique que pour l'opérationnalisation informatique, car elle autorise un filtrage progressif, en plusieurs étapes, et permet de nuancer les interprétations possibles en fonction des spécificités des corpus, de leur domaine thématique ou de leur genre.

6. Références

[AGI 00] AGICHTEIN E., GRAVANO L., « Snowball : Extracing relations from large plain text collections», *Proceedings of the 5th ACM Conference on Digital Libraries*, p. 85-94, San Antonio, Texas, 2000.

[AGU 08] AGUADO de Cea G., GÓMEZ PÉREZ A., MONTIEL PONSODA E., SUÁREZ FIGUEROA M.C, « Natural Language-Based Approach for Helping in the Reuse of Ontology Design Patterns», A. Gangemi and J. Euzenat (Eds.) *Knowledge Engineering: Practice and Patterns, Proc. of the 16th EKAW Conference on Knowledge Engineering and Knowledge Management*, Acitrezza (Italy), Sept 2008, Berlin: Springer, LNAI 5268, p. 32-47, 2008.

[AMA 01] AMARDEILH F., Extraction d'Information : Etude de Faisabilité appliquée au domaine Boursier, Mémoire de DEA RACCPOR de l'Université Technologique de Troyes, 2001.

[AMA 07] AMARDEILH F., Web Sémantique et Informatique Linguistique : propositions méthodologiques et réalisation d'une plateforme logicielle, Mémoire de thèse de l'université Paris X, 2007.

[AUG 08] AUGER A., BARRIERE C., « Pattern based approaches to semantic relation extraction: a state-of-the-art », *Terminology*, Auger A. and Barriere C (Eds.), special issue on "Pattern-based approaches to semantic relation extraction", Amsterdam/Philadelphia, John Benjamins Publishing Company, 14-1, p. 1-19, 2008.

[AUS 07] AUSSENAC-GILLES, CONDAMINES A., « Corpus et terminologie ». R.T. Pédauque (ed.) : *La redocumentarisation du monde* Toulouse, Cepadues Editions, p.131-147.

[AUS 08a] AUSSENAC-GILLES N., CHAGNOUX M., HERNANDEZ N., « An Interactive Pattern-Based Approach for Extracting Non-Taxonomic Relations from Texts», *Pacific Graphics*, *Patras, Greece*, *22/07/2008*, P. Buitelaar, P. Cimiano, G. Paliouras, M. Spilliopoulou (Eds.), *proceedings of the ECCAI workshop OntoLex08 - From Text to Knowledge: The Lexicon/Ontology Interface*, p. 1-6, 2008.

[AUS 08b] AUSSENAC-GILLES N., DESPRES S., SZULMAN S. , « The TERMINAE Method and Platform for Ontology Engineering from texts », *Bridging the Gap between Text and Knowledge - Selected Contributions to Ontology Learning and Population from Text*. P. Buitelaar, P. Cimiano (Eds.), IOS Press, p. 199-223, 2008.

[AUS 08c] AUSSENAC-GILLES N., JACQUES M.-P., « Designing and Evaluating Patterns for Relation Acquisition from Texts with CAMÉLÉON », Auger A. and Barriere C. (Eds.), *Terminology* 14-1, Pattern-based approaches to semantic relation extraction, Amsterdam/Philadelphia, John Benjamins Publishing Company, 14-1, p. 45-73, 2008.

[BAR 04] BARRIÈRE C., « Building a concept hierarchy from corpus analysis», *Terminology* 10-2, p. 241-263, 2004.

[BAR 06] BARRIERE C., AGBADO A., « TerminoWeb: a software environment for term study in rich contexts », *International Conference on Terminology, Standardization and Technology Transfert (TSTT 2006), Beijing (China)*, p. 103-113, 2006.

[BLO 07] BLOHM S., CIMIANO P., « Using the Web to Reduce Data Sparseness in Pattern-Based Information Extraction », Proc. of *PKDD 2007*, 11th European Conf. on Principles and Practice of Knowledge Discovery in Databases, Warsaw, Poland, 18-29, 2007.

[BUD 07] BUDIN G., « L'apport de la philosophie autrichienne au développement de la théorie de la terminologie : ontologie, théories de la connaissance et de l'objet », D. Savatovsky, D. Candel (eds.) : Genèses de la terminologie contemporaine, *Langages* 168**,** p. 11-23, 2007.

[BUI 05] BUITELAAR P., CIMIANO P., MAGNINI B., Ontology Learning From Text: Methods, Evaluation and Applications, IOS Press, 2005.

[BYR 99] BYRD R., RAVIN, Y., « Identifying and Extracting Relations in Text », *Proceedings of NLDB* 99 *(Application of Natural Language to dataBase*), Klagenfurt, Austria, 1999.

[CHA 96] CHARLET J., BACHIMONT B., BOUAUD J., ZWEIGENBAUM P., « Ontologie et réutilisabilité : expérience et discussion », N. Aussenac-Gilles, P. Laublet, C. Reynaud (eds) : *Acquisition et Ingénierie des Connaissances*, Toulouse : Cépaduès-Editions, p.69-88, 1996.

[CON 00] CONDAMINES A., « *Chez* dans un corpus de sciences naturelles : un marqueur de méronymie ? ». *Les Cahiers de Lexicologie* 77, p. 165-187, 2000.

[CON 05] CONDAMINES A., « Anaphore nominale infidèle et hyperonymie : le rôle du genre textuel », *Revue de Sémantique et Pragmatique* 18, p.23-42, 2005.

[CON 06] CONDAMINES A., « *Avec* et l'expression de la méronymie : l'importance du genre textuel ». G. Kleiber, C. Schnedecker, A. Thyssen (eds.) : *La relation «Partie - Tout».* Leuven : Peeters, p.633-650, 2006.

[CON 07] CONDAMINES A., « L'interprétation en sémantique de corpus : le cas de la construction de terminologies », *Revue Française de Linguistique Appliquée*, Corpus : état des lieux et perspectives. Vol.XII-1. p. 39-52, 2007.

[CON 09] CONDAMINES A., « Taking *genre* into account for analyzing conceptual relation patterns », à paraître dans *Corpora*. 2009.

[CRU 02] CRUSE A., « Hyponymy and its Varieties», Green R., Bean C.A., Myaeng S.-H (eds.), *The semantics of relationships*, Dordrecht/Boston/London, Kluwer Academic Publishers. p. 3-22, 2002.

[DES 91] DESCLES J.P., JOUIS C., OH H.G., MAIREREPPERT D., « Exploration contexuelle et sémantique : un système expert qui trouve les valeurs sémantiques des temps de l'indicatif dansun texte », Herin-Aimé D., Dieng R., Regourd J.-P., Angoujard J.-P. (Eds.) *Knowledge Modelling and Expertise Transfer (KEMT'91),* Amsterdam, p. 371-400, 1991.

[FAU 98] FAURE, D., NEDELLEC, C.: « A corpus-based conceptual clustering method for verb frames and ontology », Velardi, P. (ed.): *Proceedings of the LREC Workshop on Adapting lexical and corpus resources to sublanguages and applications*, p. 5–12, 1998.

[FEL 04] FELIU J., Relacions conceptuals i terminologia: analisis i proposta de deteccio semiautomatica, Doctoral thesis, Universitat Pompeu Fabra, Barcelona, Espagne. 2004.

[FIS 07] FISH S., *Quand lire c'est faire,* Paris, Les Prairies ordinaires, 2007 (1è parution : Harvard University Press, 1980).

[GAI 03] GAIZAUSKAS R., DEMETRIOU G., ARTYMIUK P., WILLETT P., « Protein structure an dinformation extraction from biological texts : The PASTA system », *Bioinformatics* 19(1), p. 135-143, 2003.

[GAR 98] GARCIA, D., Analyse automatique des textes pour l'organisation causale des actions, Réalisation du système COATIS, Thèse de Doctorat en informatique, Université Paris IV-Sorbonne, 1998.

[GIR 02] GIRJU, R., MOLDOVAN D, « Text mining for causal relations », *Proceedings of FLAIRS 2002*, Pensacola Beach Florida, p. 360–364, 2002.

[GUA 96] GUARINO N., PRIBBENOW S., VIEU L. (eds) : Special issue on Part-whole relations. *Data and Knowledge Engineering Journa*l, 20(3), 1996.

[GRA 04] GRABAR N., HAMMON T., « Les relations dans les terminologies structurées : de la théorie à la pratique », *Revue d'Intelligence Artificielle (*RIA), 18(1), Paris : Hermès, p. 57-85, 2004.

[GRE 02] GREEN R., BEAN C.A., MYAENG S.-H., « Introduction », Green R., Bean C.A., Myaeng S.-H (eds.), T*he semantics of relationships*, Dordrecht/Boston/London : Kluwer Academic Publishers, p. vi-xviii, 2002.

[HAD 02] HADDAD M., Extraction et impact des connaissances sur les performances des systemes d'information, Thèse de doctorat en Informatique de l'Université de Grenoble1, 2002.

[HAL 08] HALSKOV J., BARRIERE C., « Web based extraction of semantic relation instances dor terminology work », Auger A. and Barriere C (Eds.), *Terminology* 14-1, special issue on "Pattern-based approaches to semantic relation extraction", Amsterdam/Philadelphia, John Benjamins Publishing Company, 14-1, p. 20-44, 2008.

[HEA 92] HEARST, M.A.: « Automatic acquisition of hyponyms from large text corpora». *Proceedings of the 14th conference on Computational linguistics*, Morristown, NJ, USA, Association for Computational Linguistics, p. 539–545, 1992.

[JAC 96] JACKIEWICZ, A., « L'expression lexicale de la relation d'ingrédience (partie-tout) », *Faits de langues***,** 7, p. 53-62, 1996.

[JAC 06] JACQUES M.-P., AUSSENAC-GILLES N. « Variabilité des performances des outils de TAL et genre textuel : Cas des patrons lexico-syntaxiques », *Traitement Automatique des Langues (TAL)*, Association pour le Traitement Automatique des Langues (ATALA), 47 (1), p. 11-32, 2006. Accès : http://www.atala.org/article.php3?id_article=320

[KEE 08] KEET M., ALESSANDRO A., « Representing and reasoning over a taxonomy of part–whole relations », *Applied Ontology*, 3(1-2), Special issue on Ontological Foundations of Conceptual Modelling, Giancarlo Guizzardi and Terry Halpin (Eds.), 2008.

[KHE 07] KHELIF K., DIENG-KUNTZ R., BARBRY P., « An ontology-based approach to support text mining and information retrieval in the biological domain », *Journal of Universal Computer Science* (JUCS), 13 (12), Special Issue on Ontologies and their Applications, p. 1881-1907, 2007.

[KHO 02] KHOO C., CHAN S., NIU Y. « The Many Facets of the Cause-Effect Relation», Green R., Bean C.A., Myaeng S.-H (eds.) *The semantics of relationships*. Dordrecht/Boston/London : Kluwer Academic Publishers, p. 51-70, 2002.

[LEE 01] LEE D., « Genres, Registers, Text Types, Domain and Styles: Clarifying the concepts and navigating a path through the BNC jungle », *Language Learning and Technology*, 5(3), p.37-72, 2001.

[MAE 00] MAEDCHE A., STAAB S., « Discovering conceptual relations from text », *W. Horn (ed.): ECAI 2000. Proceedings of the 14th European Conference on Artificial Intelligence*. IOS Press, Amsterdam, p. 321-325, 2000.

[MAE 02] MAEDCHE A., *Ontology learning for the Semantic Web*, volume 665. Kluwer Academic Publisher, 2002.

[MAR 06] MARSHMAN E., L'HOMME M.-C., « Portabilité des marqueurs de la relation causale : étude sur deux corpus spécialisés », *Actes des Journées du CRTT : corpus et dictionnaires de langues de spécialité. Lyon, France*, p. 87-110, 28-29 September 2006.

[MAR 07] MARSHMAN E., Lexical Knowledge Patterns for the Semi-automatic Extraction of Cause-effect and Association Relations from Medical Texts: A Comparative Analysis of English and French. Thèse de doctorat, Département de linguistique et de traduction, Université de Montréal. 2007.

[MAR 08] MARSHMAN E., VAN BOLDEREN, « Interlinguistic variation and Lexical Knowledge patterns », B. Nistrup Madsen, H. E. Thomasen (Eds), *Terminology and Knowledge Engineering (TKE 2008), Managing Ontologies and Lexical Resources*, August 2008 Copenhague : ISV, p. 263-278, 2008.

[MAY 08] MAYNARD D., LI Y., PETERS W., « NLP techniques dor Term Extraction and Ontology Population », Buitelaar, P., Cimiano P. (Eds.), *Ontology Learning and Population: Bridging the Gap between Text and Knowledge*, p. 107-127. IOS Press, Amsterdam, 2008.

[MEY 01] MEYER I., « Extracting Knowledge-rich Contexts for Terminography : A Conceptual and methodological Framework », D. Bourigault, M.C. L'homme, C.Jacquemin (eds) : *Recent Advances in Computational Terminology*, John Benjamins. p. 279-302, 2001.

[MIN 02] MINEL J.-L., *Filtrage Sémantique : du résumé automatique à la fouille de textes*. Paris : Hermès, 2002.

[MOR 99] MORIN E., « Des patrons lexico-syntaxiques pour aider au dépouillement terminologiques », *Traitement Automatique des Langues*, 40-1, 143-166. 1999.

[MUK 06] MUKHERJEA S., SAHAY, S, « Discovering biomedical relations utilizing the world-wide web », *Proceedings of Pacific Symposium on Bio-Computing. Maui*, Hawaii, p.164-175. 2006.

[NED 04] NEDELLEC C., « Machine Learning for Information Extraction in Genomics - State of the Art and Perspectives », *Text Mining and its Applications: Results of the NEMIS Launch Conference Series: Studies in Fuzziness and Soft Computing Sirmakessis*, Spiros (Ed.), 2004.

[PAN 06] PANTEL P., PENNACHIOTTI M., « Expresso: leveraging generic patterns for automatically harvesting semantic relations », *Proceedings ot ACL 2006*, Sidney Australia, p. 113-120, 2006.

[PAN 08] PANTEL P., PENNACHIOTTI M., « Automatically Harvesting and Ontologizing Semantic Relations », Buitelaar, P., Cimiano P. (Eds.), *Ontology Learning and Population: Bridging the Gap between Text and Knowledge*, p. 171-195. IOS Press, Amsterdam, 2008.

[RAM 08] RAMAKRISHNAN C., MENDES P., JUN WANG S., SHETH A., « Unsupervised discovery of compound entities for relationship extraction », A. Gangemi, J. Euzenat (Eds.) *Knowledge Engineering: Practice and Patterns, Proceedings of the 16th EKAW Conference on Knowledge Engineering and Knowledge Management*, Acitrezza (Italy), Sept/Oct 2008, Berlin: Springer, LNAI 5268, p. 146-155, 2008.

[RAS 95] RASTIER F., « Le terme : Entre ontologie et Linguistique », *La Banque des Mots* 7, Numéro spécial, p. 35-64, 1995.

[RAS 01] RASTIER F., *Arts et Sciences du texte*. Paris : PUF, formes sémiotiques, 2001.

[REB 00] REBEYROLLE J., TANGUY L., « Repérage automatique de structures linguistiques en corpus : le cas des énoncés définitoires », *Cahiers de Grammaire*, 25, p. 153-174, 2000.

[ROG 00], ROGERS, M., « Genre and Terminology » A. Trosborg (ed.) *Analysing Professional Genre*, Amsterdam/Philadelphia: John Benjamins, p. 3-19, 2000.

[RUI 07] RUIZ-CASADO M., ALFONSECA E., CASTELLS P., « Automatising the learning of lexical patterns: An application to the enrichment of wordnet by extracting semantic relationships from wikipedia », *Data Knowledge Engineering,* 61(3), p. 484–499, 2007.

[RUI 08] RUIZ-CASADO M. ALFONSECA E., OKUMURA M., CASTELLS P., « Information extraction and Semantic Annotation of Wikipedia », Buitelaar, P., Cimiano P. (Eds.), *Ontology Learning and Population: Bridging the Gap between Text and Knowledge*, p. 145-169. IOS Press, Amsterdam, 2008.

[SCH 05] SCHUTZ A., BUITELAAR P., « RelExt: A tool for relation extraction from text in ontology extension », Gil,Y., Motta, E., Benjamins V.R., Musen, M., (eds), *The Semantic Web – Proceedings of ISWC 2005: 4th International Semantic Web Conference, ISWC 2005*, Galway, Ireland, Berlin: Springer Verlag, LNAI 3729, p. 593-606, 2005.

[SEG 01] SEGUELA P., Construction de modèles de connaissances par analyse linguistique de relations lexicales dans les documents techniques. Thèse de doctorat en Informatique, Université Toulouse III, mars 2001.

[SIE 08] SIERRA G., ALARCON R., AGUILAR C. , BACH C., « Definitional verbal patterns for semantic relation extraction », Auger A. and C. Barrière (eds.), Pattern-based Approaches to Semantic Relation Extraction: Special issue of *Terminology* 14:1, p. 74–98, 2008.

[SIL 99] SILBERZTEIN M., « INTEX: a Finite State Transducer toolbox », *Theoretical Computer Science*, 231-1, Elsevier Science. 1999.

[SLO 95] SLODZIAN M., « La doctrine terminologique, nouvelle théorie du signe au carrefour de l'universalisme et du logicisme », *ALFA Actes de Langue Française et de Linguistique : Terminologie et langues de spécialité*, 7/8, Dalhousiana, Halifax, Nova Scotia, Canada., p. 121-136, 1995.

[SMA 08] SMART P., ENGELBRECHT P., « An Analysis of the Origin of Ontology Mismatches on the Semantic Web », A. Gangemi, J. Euzenat (Eds.) *Knowledge Engineering: Practice and Patterns, Proc. of the 16th EKAW Conf. on Knowledge Engineering and Knowledge Management*, Acitrezza (Italy), Sept 2008, Springer, LNAI 5268, p 120-135, 2008.

[SOL 08] SOLER V., ALCINA A.**,** « Patrones léxicos para la extracción de conceptos vinculados por la relación parte-todo en español », *Terminology 14-1*, Auger A. & Barrière C. (eds.), Pattern-based Approaches to Semantic Relation Extraction, p. 99–123, 2008.

[STA 01] STAAB S., MAEDCHE A., « Ontology Learning for the Semantic Web », *IEEE Intelligent Systems,* Special Issue on the Semantic Web, 16(2), p. 72-79, 2001.

[VEL 06] VELARDI P., NAVIGLI R., CUCHIARELLI A., NERI R., « Evaluation of Ontolearn, a methodology for automatic learning of domain ontologies », Buitelaar, P., Cimiano P. and Magnini B. (eds.), *Ontology Learning from Text: Methods, evaluation and applications*. p. 92–106. Amsterdam: IOS Press, 2006.

[VIE 07] VIEU L., AURNAGUE M., « Part-of relations, functionality and dependence », M. Aurnague, M. Hickmann, L. Vieu (Eds.), *The Categorization of Spatial Entities in Language and Cognition*. John Benjamins Publishing Company, p. 307-336, 2007.

[WIN 87] WINSTON M., CHAFFIN R., HERRMANN D., « A Taxonomy of Part-Whole Relations », *Cognitive Science*, 11, p. 417-441, 1987.

[WUS 81] WÜSTER E., « L'étude scientifique générale de la terminologie, zone frontalière entre la linguistique, la logique, l'ontologie, l'informatique et les sciences des choses », G.Rondeau et H.Felber (eds), *Textes choisis de terminologie*, GIRSTERM, Université de Laval, Québec. p. 55-108, 1981.