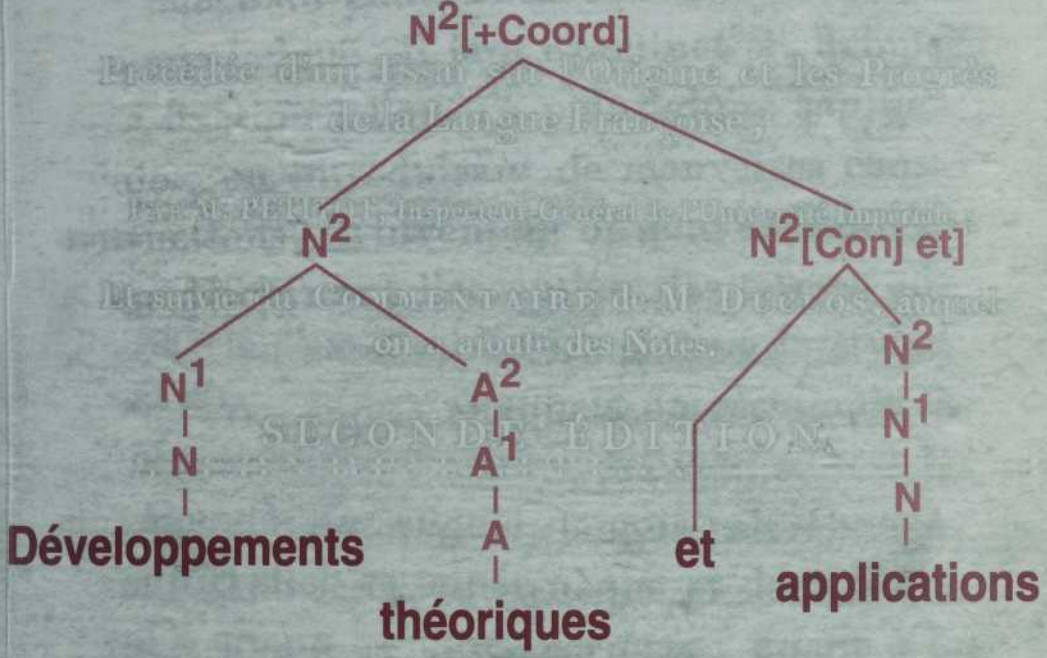
itement automatique du français écrit



Sous la direction de
Louisette Emirkanian
et
Lorne H. Bouchard

Acfas LES CAHIERS SCIENTIFIQUES 86

ÉDITION:

Association canadienne-française pour l'avancement des sciences
Acfas

COMMANDES TÉLÉPHONIQUES:
Acfas : (514) 849-0045
Les cartes MasterCard et Visa sont acceptées.
Veuillez allouer trois semaines pour la livraison.

COMMANDES POSTALES:
Acfas : 425, rue De La Gauchetière Est
Montréal (Québec)
H2L 2M7
Télécopieur : (514) 849-5558
(les individus doivent joindre le paiement à leur commande)

DISTRIBUTION EN LIBRAIRIE:
Diffusion Prologue
1650, boul. Lionel-Bertrand
Boisbriand (Québec)
J7H 1N7
Téléphone: (514) 434-0306
Ext.: 1-800-363-2864
Télécopieur: (514) 434-2627
Ext.: 1-800-361-8088

MISE EN PAGE DU TEXTE :
Julie Hudon

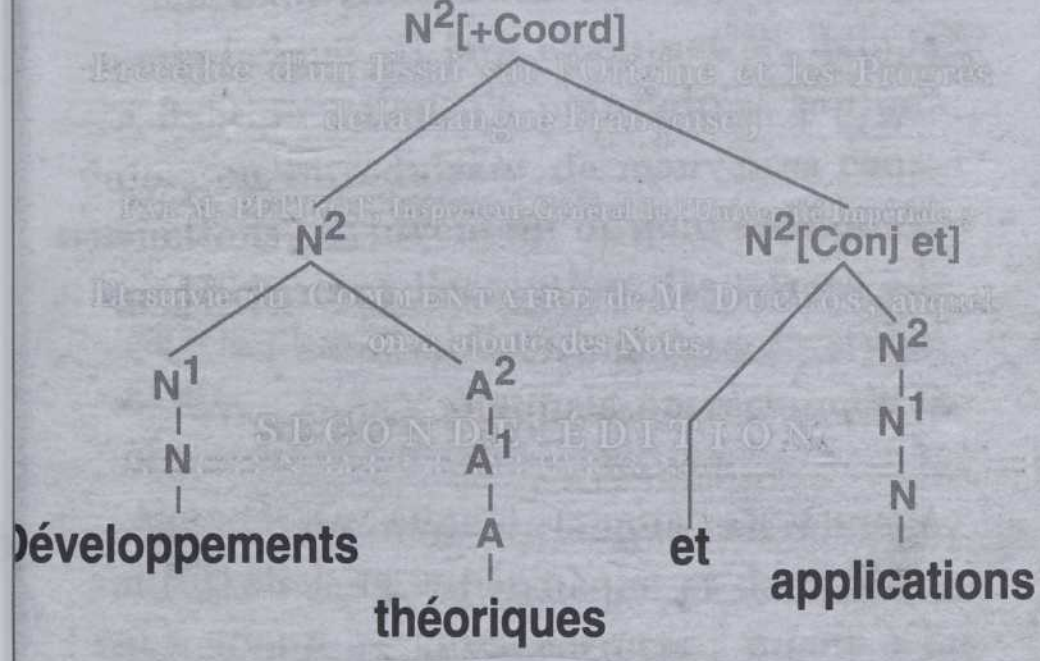
GRAPHISME DE LA PAGE COUVERTURE:
Nathalie Proulx

IMAGE DE LA COUVERTURE:
Adaptation de la page de garde de la
Grammaire générale et raisonnée de Port-Royal, 1810

Dépôt légal 2e trimestre 1996
Bibliothèque nationale du Québec
Bibliothèque nationale du Canada

Voir données de catalogage avant publication (CIP) au couvert III

Traitement automatique du français écrit



Sous la direction de
Louisette Emirkanian
et
Lorne H. Bouchard

Acfas *LES CAHIERS SCIENTIFIQUES* 86

Table des matières

Remerciements

Nous voudrions d'abord remercier tous les collaborateurs de ce recueil grâce auxquels le colloque que nous avons eu le plaisir d'organiser, le 17 mai 1994 à l'Université du Québec à Montréal, a été un succès.

Nous voudrions exprimer toute notre gratitude à Jean-Guy Meunier pour son appui constant et la confiance qu'il nous a témoignée. Nous remercions également Monique Charbonneau, Présidente Directrice Générale du CEFRIO, d'avoir prononcé l'allocution d'ouverture du colloque qui a donné le ton à cette journée.

Une table ronde intitulée "La recherche en linguistique informatique et les industries de la langue" clôturait cette journée. Nous remercions très sincèrement les participants à cette table ronde : Suzanne Bertrand-Gastaldy (Université de Montréal), Claude Coulombe (Machina Sapiens), Louis Lamothe (Secrétariat du conseil du trésor), Jean-Guy Meunier (Université du Québec à Montréal), Normand Montour (Bell Canada), Richard Parent (Ministère des communications du Québec), Silvia Pavel (Travaux publics et services gouvernementaux) et Michèle Valiquette (Travaux publics et services gouvernementaux).

La tenue de ce colloque ainsi que cette publication ont été rendues possibles grâce à l'aide financière que nous ont apportée le Centre ATO.CI, le CEFRIO, l'Observatoire Québécois des Industries de la Langue, la société ALEX-Informatique, les Départements de linguistique et d'informatique de l'UQAM, l'ACFAS, le Programme d'aide financière à la recherche et à la création de l'UQAM ainsi que le Ministère des Affaires internationales, de l'Immigration et des Communautés culturelles (Direction générale France). Qu'il nous soit permis de remercier ici plus particulièrement Josiane Ayoub, Paula Bouffard, Monique Charbonneau, Philippe Gabrini, Édith Girard, Christian Gohel, Patricia Legault, Jean-Guy Meunier, Robert Papen et Robert Proulx.

Finalement, nous remercions les trois évaluateurs anonymes du manuscrit dont les commentaires et les remarques ont contribué à l'amélioration de la présentation et du contenu de ce volume.

Julie Hudon a effectué le fastidieux travail de mise en page de la version finale. Nous avons apprécié sa compétence, sa gentillesse, sa disponibilité et surtout sa patience. Nous la remercions très vivement.

Présentation

Les textes réunis dans cet ouvrage ont d'abord été présentés comme communications lors d'un colloque, *Traitement automatique du français écrit : développements théoriques et applications* que nous organisions, au soixante-deuxième Congrès de l'ACFAS, sous l'égide du Centre ATO.CI, des Départements de linguistique et d'informatique de l'Université du Québec à Montréal et de l'Observatoire Québécois des Industries de la Langue. Ce colloque a eu lieu le 17 mai 1994 à l'Université du Québec à Montréal.

Le traitement automatique de la langue écrite est un domaine de recherche multidisciplinaire qui prend sa source essentiellement dans les recherches en linguistique et en informatique.

Compte tenu du rôle central de la faculté langagière parmi les processus cognitifs d'ordre supérieur, l'analyse et la génération de l'écrit constituent des domaines de recherche fondamentaux pour le développement des systèmes à base de connaissances ; de plus, ce domaine est le moteur de toutes les applications dans les industries de la langue, un domaine en plein essor. Plus récemment, la possibilité du traitement parallèle de l'information a eu un impact considérable non seulement sur les applications mais aussi sur la recherche

fondamentale dans ce domaine. En effet, le traitement parallèle permet d'envisager une interaction effective des savoirs, jusqu'alors cloisonnés dans des sous-disciplines plus ou moins étanches.

En ciblant les connaissances de la langue, le traitement automatique de la langue naturelle s'inscrit dans le courant général de la recherche sur les systèmes à base de connaissances. Du point de vue du développement de ces systèmes, l'écrit est une source précieuse de connaissances que les concepteurs souhaiteraient pouvoir exploiter dans les systèmes évolués de traitement de l'information.

Le volume de l'information écrite suggère que le traitement automatique de la langue est un domaine d'application qui peut avoir une importance économique considérable. Mais aussi, et cela nous semble peut-être encore plus important, l'étude de la problématique du traitement automatique de l'écrit montre que l'évolution des connaissances dans ce domaine a un impact qualitatif sur la nature de ce qui peut être automatisé et sur le partage effectif des efforts entre personne et machine.

Les articles de ce recueil font état de travaux sur le traitement automatique du français écrit.

De nombreux fils tissent la trame complexe de ce domaine. En premier lieu, la direction qui nous apparaît la plus saillante est celle de l'axe théorie/pratique, sur laquelle on peut situer les articles selon qu'ils traitent d'un problème de nature plus fondamentale, du développement d'outils ou encore de l'intégration de divers outils existants lors de la mise en œuvre d'une application. Cette direction n'identifie que deux pôles d'une réalité qui demeure entière. En effet, le développement théorique sans la perspective des applications risque de devenir stérile et, en contrepartie, le développement d'applications qui ne sont pas construites sur des bases théoriques solides peut être aveugle. En second lieu, les articles peuvent être situés selon une dimension horizontale/verticale, selon que l'objectif visé est de développer un système à large spectre ou un système pointu qui cherche, par une étude en profondeur, à résoudre un problème précis.

L'article de Laurence Danlos étudie le problème de la génération d'un discours comportant deux phrases juxtaposées sans qu'un mot de liaison n'indique la nature du rapport entre les phrases (une figure de style connue sous le nom de *parataxe*) dans le cas particulier où il y a relation causale directe entre ces phrases. Une étude minutieuse des données linguistiques met en relief une asymétrie surprenante dans le comportement de ces discours selon que la cause précède le résultat ou l'inverse et souligne le comportement abérant des discours CAUSE < RÉSULTAT. Une analyse de ces faits linguistiques dans la perspective de la structure aspectuelle des événements sous-jacents permet cependant d'identifier deux types de coréférence évenementielle, à savoir la globalisation et la particularisation, et d'expliquer le comportement des discours exprimant une relation causale directe grâce à ces relations de coréférence évenementielle. La contribution de cet article est d'illustrer concrètement que le problème de la génération est extrêmement complexe et qu'il existe, sur ce sujet, d'importantes lacunes dans nos connaissances.

L'article de Louisette Emirkanian, Lyne Da Sylva et Lorne H. Bouchard documente une partie de la grammaire computationnelle du français qui a été développée à l'UQAM. L'article étudie d'une part la façon de rendre compte de l'accord du participe passé dans le cadre de la théorie de la Grammaire syntagmatique généralisée (GSG) et d'autre part l'implantation de l'analyse sur ordinateur à l'aide de l'outil *Grammar Development Environment* (GDE). La formalisation précise dans la GSG des conditions de cet accord aboutit à une première proposition d'analyse. Une étude attentive des principes de fonctionnement et de la façon d'en rendre compte avec l'outil GDE débouche sur une nouvelle formalisation qui permet de traiter le problème de façon plus générale. La contribution de cet article est de montrer que des phénomènes syntaxiques particuliers à une langue naturelle donnée sont un terrain d'exploration fertile pour la validation des théories linguistiques.

Alors que les deux articles précédents exploraient de façon verticale certains problèmes liés au traitement du français

écrit, l'article de Christophe Fouqueré argumente en faveur de l'utilisation d'un formalisme unifié pour étudier les propriétés des grammaires des langues naturelles formulées dans le paradigme des grammaires d'unification, LFG, GSG et HPSG notamment. L'originalité de son approche est que le formalisme qu'il propose tâche de conserver les structures fondamentales de la théorie grammaticale d'origine et évite donc le défaut majeur de vouloir tout niveler. Le problème de la validation d'une grammaire est utilisé pour illustrer l'utilité de ce formalisme. Une économie de moyens est mise en œuvre pour résoudre le problème et éviter les écueils simultanés de l'explosion combinatoire et de l'indécidabilité : analyse et validation de la partie réécriture hors contexte, extraction et solutions des contraintes par les techniques de résolution de contraintes. L'article se termine par une étude du formalisme unifié à l'aide de la logique linéaire. L'intérêt de la logique linéaire est qu'elle permet non seulement un traitement naturel de la non-monotonicité mais aussi un traitement uniforme selon les deux perspectives calcul et déduction.

Les deux articles suivants sont issus de recherches sur le traitement parallèle de l'information linguistique qui ont été effectuées à l'UQAM dans le cadre du projet ALEX/ATO.CI/UQAM coordonné par Robert Proulx.

L'article signé Lyne Da Sylva et al. présente une modélisation parallèle du traitement du problème de la désambiguïsation structurelle dans la chaîne de traitement du français et décrit un prototype qui a été développé sur ordinateur VOLVOX. La résolution des problèmes d'ambiguïté structurelle fait appel à plusieurs sources de connaissances qui se situent dans des domaines souvent cloisonnés : morphologie, syntaxe, sémantique et prosodie, par exemple. Un langage commun uniforme, basé sur les structures d'attribut/valeur, a été choisi pour représenter l'expertise dans les différents domaines. Un compilateur écrit en LISP permet de traduire ce langage commun en un ensemble de règles de production, un formalisme essentiellement parallèle. Ces règles de production sont ensuite compilées en un ensemble d'objets qui sont distribués sur les ordinateurs fonctionnant en parallèle. Les

étapes discrètes dans le traitement ont pour avantage de permettre de suivre le flux de l'expertise et de voir où et comment il est utilisé.

L'article de François Daoust et Fernande Dupuis est issu d'un projet plus vaste visant à évaluer la possibilité de paralléliser SATO, un logiciel d'analyse de texte qui est en développement à l'UQAM depuis plus de vingt ans. Cet article, comme le précédent, traite du problème de la désambiguïsation, mais ici il s'agit de désambiguïsation catégorielle. L'article compare les résultats obtenus en utilisant une procédure de désambiguïsation basée sur des règles classiques avec ceux obtenus en utilisant une procédure basée sur un modèle associatif réalisé à l'aide de réseaux neuromimétiques. Un protocole expérimental a été mis au point pour comparer les deux approches. Une étude des résultats obtenus a pu mettre en lumière des lacunes dans la grille catégorielle.

Les quatre derniers articles traitent d'applications du traitement automatique de l'écrit. Les trois premiers visent le développement d'outils spécialisés permettant la prise en charge par l'ordinateur de certains traitements alors que le dernier fait appel à une orchestration de toute une panoplie d'outils existants.

L'article de Massimo Fasciano et Guy Lapalme décrit le prototype d'un outil permettant de présenter des données numériques soit sous forme de texte soit sous forme de graphique, selon ce qui est le plus efficace dans une situation donnée. En effet, les deux moyens d'expression sont complémentaires : le graphique permet de montrer une vue d'ensemble alors que le texte permet de décrire et de souligner les détails intéressants.

L'article de Christine Leonhardt fait le point sur les recherches en terminotique poursuivies activement depuis plusieurs années à la Direction de la terminologie et de la documentation du Bureau de la traduction à Travaux publics et Services gouvernementaux Canada, recherches qui ont porté fruit. En effet, la banque de terminologie TERMIUM®, qui a évolué depuis 1987, est maintenant disponible soit en direct soit sur CD-ROM et le poste de travail LATTER est

couramment utilisé par les terminologues de la Direction de la terminologie et de la documentation.

Les deux derniers articles illustrent le paradigme des systèmes à base de connaissances où des connaissances, qui ont d'abord été identifiées, sont ensuite stockées et exploitées dans les applications.

L'article signé Pierre Isabelle et al. défend le point de vue selon lequel l'analyse des traductions est le point de départ d'une nouvelle génération d'aides à la traduction. L'idée essentielle consiste à exploiter la mémoire traductionnelle contenue dans les textes bilingues, une richesse dont le potentiel n'a pas jusqu'à présent été exploité. Cette opération est rendue possible techniquement par la mise au point d'une procédure permettant d'apparier les phrases contenues dans un texte avec celles contenues dans le texte de sa traduction. Basé sur cet appariement, un prototype permettant de détecter certaines erreurs de traduction a été mis au point et les résultats sont évalués. Finalement, un projet en cours de réalisation, vise le développement d'un système de dictée de traduction qui combine l'utilisation de la mémoire traductionnelle comme source additionnelle de connaissances et le modèle de la parole utilisé dans les systèmes de reconnaissance vocale qui se fondent sur les modèles de Markov cachés. Les connaissances représentées dans ce système sont essentiellement des connaissances des langues de départ et d'arrivée.

L'article signé Suzanne Bertrand-Gastaldy, Louis-Claude Paquin, Gracia Pagola et François Daoust présente une application dont le but est d'assister les conseillers juridiques dans la sélection, la classification, l'indexation et la condensation des jugements de la cour. Prenant comme point de départ que le texte est un objet sémiotique, on en conclut que l'automatisation complète de la tâche est impossible, au mieux on ne peut qu'automatiser une aide à l'interprétation basée sur des indices textuels. Les textes d'origine sur support informatique sont enrichis d'annotations qui permettent de calculer des indices. Un logiciel d'analyse statistique des données est utilisé pour effectuer une analyse de discrimination de ces indices. Les résultats obtenus sont alors confrontés avec les

résultats recueillis dans la première phase de l'enquête cognitive et soumis aux experts du domaine. La solution fait appel à de nombreux outils : analyseur de texte (SATO), analyseur de données statistiques (SPSS), coquille de système expert (GSE), programmes développés sur mesure (en ICON). Cette recherche attire l'attention sur l'importance de l'identification des sources de connaissances requises et sur celle du repérage systématique des indices qui sont utilisés pour les désigner dans le texte.

RELATIONS CAUSALES DIRECTES : DISCOURS, STRUCTURE ÉVÉNEMENTIELLE ET CORÉFÉRENCE ÉVÉNEMENTIELLE

Laurence Danlos

Résumé*

Le Traitement Automatique du Langage Naturel (TALN), que ce soit en compréhension ou en génération de textes, repose sur l'analyse du discours principalement basée sur des considérations de pragmatique ou des Sciences Cognitives. A l'inverse, les études linguistiques, principalement morpho-syntaxiques ou sémantiques, reposent sur l'analyse de la phrase isolée. Il existe donc de fait un fossé entre la communauté TALN / discours / pragmatique/ Sciences Cognitives et la communauté Linguistique / phrase isolée / sémantique / morpho-syntaxe. Nous pensons que ce fossé est tout à fait regrettable et nous voulons montrer à travers cet article que l'étude du discours (et donc du TALN) ne saurait se passer des analyses de la phrase isolée et que, réciproquement, l'étude de la phrase isolée aurait tout à gagner en osant plonger une phrase isolée dans un discours (de deux phrases !). Pour atteindre notre objectif, nous disséquerons un cas particulier, les discours exprimant une "relation causale directe". Cette étude emprunte les méthodes de la linguistique (e.g. principe de la paire

* Je remercie Michel Cosse, Danièle Godard et Jean-Claude Milner pour leur relecture attentive et leurs commentaires pertinents.

minimale) pour les appliquer aux discours. Il s'agit donc de contrôler l'explosion combinatoire des paramètres linguistiques qui surgit dès qu'un ensemble de phrases (ne serait-ce que de cardinalité 2) est à l'étude, sans se jeter à corps perdu dans des considérations pragmatico-cognitives dont la formalisation et la vérificabilité sont sujettes à caution.

L'étude "linguistique" du discours que nous allons proposer est d'une longueur telle que nous avons dû faire l'impasse sur ses conséquences en génération de textes, notre domaine du TALN préféré. Ces conséquences sont examinées dans (Danlos 1995a, 1995b). Soulignons cependant le point suivant : les données et l'analyse linguistique qui vont être présentées concernent le français mais elles se révèlent identiques (à quelques modifications près) en anglais, en italien et en coréen, langues pour lesquelles nous avons bénéficié des jugements d'acceptabilité de linguistes natifs[1]. Est-ce à dire que les phénomènes discursifs que nous avons mis en avant relèvent de principes universaux ?

Introduction

Les verbes "causatifs" tels que *casser* ou *tuer* ont été largement étudiés dans la littérature, entre autres depuis le fameux KILL = CAUSE BECOME NOT ALIVE de McCawley (1968). Nous examinerons ici des discours exprimant une relation causale dont la première phrase exprime la cause et la seconde le résultat au moyen d'un verbe causatif construit soit à l'actif avec une valeur aspectuelle d'accomplissement[2] soit à la forme neutre avec une valeur aspectuelle d'achèvement :

(1) Luc a cogné la carafe contre l'évier. Il l'a cassée.

(2) Luc a cogné la carafe contre l'évier. Elle s'est cassée.

Nous montrerons que ces discours ont un comportement aberrant, par exemple, ils ne permettent pas d'insérer un complément de date dans la seconde phrase, bien que ce complément soit autorisé dans cette phrase prise isolément :

---

[1] Les données sur l'anglais viennent de Owen Rambow, celles sur l'italien de Fiammetta Namer, et celles sur le coréen de Manghyu Pak (Pak, à paraître). Je les remercie d'avoir effectué ce travail de linguistique contrastive.

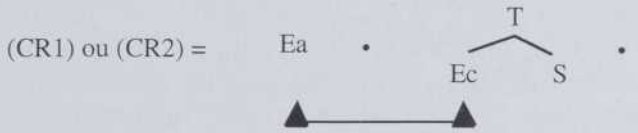
[2] Les termes aspectuels que nous employons seront précisés dès le début de la Section 1.

* Luc a cogné la carafe contre l'évier. Il l'a cassée à minuit.

* Luc a cogné la carafe contre l'évier. Elle s'est cassée à minuit.

Pour expliquer ce comportement aberrant, nous aurons recours à la structure événementielle proposée par Pustejovsky (1991). La décomposition des transitions (accomplissement ou achèvement) qu'il met en évidence permet de proposer l'analyse suivante des discours comme (1) ou (2) : la première phrase réfère à un événement noté Ea (e.g. *Luc a cogné la carafe contre l'évier*) ; la seconde phrase réfère à une transition T (e.g. *Luc a cassé la carafe* ou *La carafe s'est cassée*) décomposable en un événement noté Ec et un état noté S, selon le schéma

; l'interprétation de (1) ou (2) comme relation causale implique que les événements Ea et Ec soient interprétés comme coréférents, ce que nous schématisons de la façon suivante :



À partir de cette analyse, nous allons défendre l'hypothèse suivante : la relation de coréférence entre Ea et Ec est d'un type particulier que nous appelons "généralisation" où Ec généralise Ea, soit Ec = Généra (Ea). Ce type de coréférence événementielle s'observe dans des discours mettant en jeu une relation de "paraphrase généralisante" comme :

(a) Luc a violé Marie. Il a donc commis un crime.

La généralisation impose des contraintes sur les phrases décrivant les événements coréférentiels, e.g. interdiction d'apporter une information nouvelle dans la seconde phrase :

* Luc a violé Marie. Il a donc commis un crime à minuit.

Nous montrerons que la plupart des contraintes qui pèsent sur les discours comme (1) ou (2) ne sont rien d'autres que des contraintes de généralisation, ramenant ainsi des contraintes

apparemment *ad hoc* pesant sur des discours exprimant une relation causale à des contraintes de coréférence événementielle.

Nous évoquerons aussi les discours obtenus en inversant l'ordre des phrases :

(1') Luc a cassé la carafe. Il l'a cognée contre l'évier.
(2') ? La carafe s'est cassée. Luc l'a cognée contre l'évier.[3]

Ceux-ci ont un comportement régulier, par exemple, ils acceptent normalement l'insertion d'un complément de date dans la première ou dans la seconde phrase :

Luc a cassé la carafe à minuit. Il l'a cognée contre l'évier.
Luc a cassé la carafe. Il l'a cognée contre l'évier à minuit.

Nous assimilerons leur comportement à celui observé dans les relations de "paraphrase particularisante" :

(b) Luc a commis un crime. Il a violé Marie.
Luc a commis un crime à minuit. Il a violé Marie.
Luc a commis un crime. Il a violé Marie à minuit.

Les paraphrases particularisantes mettent en jeu un autre type de coréférence événementielle dit de "particularisation". Les relations de généralisation et de particularisation ne sont pas symétriques l'une de l'autre, en particulier nous montrerons les formules suivantes :

Ej = Généra (Ei) => Ei = Parti (Ej)
Ei = Parti (Ej) ≠> Ej = Généra (Ei)

Cette asymétrie entre généralisation et particularisation reflète que les discours (1) ou (2) ont un comportement aberrant contrairement aux discours (1') ou (2'). Nous compléterons cette étude sur les discours exprimant les relations causales par un bref examen des discours non parataxiques.

Après avoir ainsi montré que l'analyse de certains discours peut profiter pleinement d'analyses de sémantique lexicale complétées par des relations particulières de coréférence événementielle, nous conclurons en montrant que les

[3] L'acceptabilité de ce discours sera discutée dans la Section 1.6.

phénomènes discursifs peuvent à leur tour apporter une contribution intéressante à la sémantique lexicale.

La première partie de cet article est consacrée à un examen minutieux des faits empiriques concernant les discours tels que (1) et (2). La seconde partie présente notre analyse de ces faits.

1 Présentation des faits empiriques

Nous allons étudier deux types de discours parataxiques[4] illustrés par les exemples (1) et (2) :

(1) Luc a cogné la carafe contre l'évier. Il l'a cassée.

(2) Luc a cogné la carafe contre l'évier. Elle s'est cassée.

Ces discours diffèrent par leur seconde phrase : celle-ci est construite autour de *casser* conjugué au passé composé, mais ce verbe est employé dans une construction transitive en (1) tandis qu'il est à la forme (pronominale) dite de "neutralité" (Ruwet 1972, Boons et alii 1976) en (2). Malgré cette différence formelle, ces deux discours sont équivalents sur le plan de la sémantique dite véri-conditionnelle ou référentielle : ils réfèrent à la même situation dans le monde réel. Leur sémantique commune met en jeu une "relation causale directe" qui se définit de la façon suivante (Schank 1975, Danlos 1985) : la cause est une action effectuée par un agent humain H et affectant directement une entité X, le résultat est un changement d'état physique de X, ce changement n'est pas intrinsèquement impliqué par l'action. Ci-dessous d'autres exemples de relation causale directe :

Luc a cogné la carafe contre l'évier. Elle s'est fêlée.[5]
[H = Luc, X = la carafe]

Luc a posé son cartable sur mes chemises. Il les a froissées.
[H = Luc, X = mes chemises]

---

[4] Rappelons qu'une parataxe est une construction par juxtaposition sans qu'un mot de liaison indique la nature du rapport entre les phrases.

[5] Cet exemple montre qu'il n'est pas obligatoire qu'une carafe se casse quand elle est cognée contre un évier, contrairement à la première intuition de certains locuteurs. Ajoutons qu'elle peut aussi être ébréchée ou n'être pas endommagée du tout.

Luc a tiré une balle dans le crâne de Marie. Il l'a tuée.
[H = Luc, X = Marie]

Les Japonais ont envoyé une bombe sur le porte-avions. Il a coulé.
[H = les Japonais, X = le porte-avions]

Certains verbes d'action affectant directement une entité X impliquent intrinsèquement un changement d'état physique de X, par exemple :

Luc a entassé mes chemises dans le tiroir ==> Mes chemises sont entassées dans le tiroir

Par définition, ces cas ne sont pas considérés comme relevant d'une relation causale directe. Autrement dit, une relation causale directe met en jeu deux prédicats, l'un exprimant la cause, l'autre le résultat. Nous n'étudierons que les relations causales dont le résultat est un changement d'état physique, en écartant les changements d'état psychique, moral ou intellectuel. Cette position sera justifiée dans la conclusion.

Les discours (1) et (2) ne sont que des illustrations des discours que nous allons étudier. Avant de présenter ceux-ci, deux remarques.

(i) Les notations que nous allons utiliser, e.g. H Vc X, mélangent la forme et le sens : le symbole H représente l'entité dénotée par le sujet, X l'entité dénotée par l'objet, et Vc un verbe causatif, donc une réalisation de surface. Néanmoins les nombreux paradigmes que nous allons présenter sont d'une complexité telle qu'il nous a paru nécessaire d'aider le lecteur en ayant recours à cette notation abusive mais mnémonique. Signalons dans le même ordre d'idées que nous présenterons les construction transitives sous la forme canonique H Vc X même si l'entité X est réalisée sous forme de particule pré-verbale (réfléchie ou non). Les pronoms ne seront pas discutés dans cet article. Nous les utiliserons quand c'est nécessaire, sans commentaire.

(ii) Nous nous appuierons sur la classification aspectuelle de Vendler (1967) pour les différents types d'événements. La terminologie française ou anglo-saxonne variant énormément d'un auteur à l'autre, précisons les termes que nous allons

employer (qui sont adaptés de ceux de Pustejovsky (1991)) en les illustrant par des exemples :

- procès : *Luc a couru dans la forêt (pendant 1 heure), Luc buvait de la bière.*
- transition :
  - accomplissement : *Luc a couru à la boulangerie (en 1 heure), Luc a bu une bière (en 1 heure), Luc a cassé la carafe.*
  - achèvement : *La carafe s'est cassée, Luc mourut, Luc est arrivé.*
- état : *Luc est malade, Luc croit que Marie est partie.*

Signalons aussi que les discours étudiés ont tous leurs verbes conjugués au passé composé : ceci nous permet de limiter cette étude (déjà assez complexe) sans entrer dans le vaste domaine du temps.

Présentons maintenant les structures de discours que nous allons étudier :

(CR1) H Va (passé composé) ... X .... . H Vc (passé composé) X
=: Luc a cogné la carafe contre l'évier. Il l'a cassée.
[Transitive, Accomplissement]

(CR2) H Va (passé composé) ... X .... . X (se) Vc (passé composé).
=: Luc a cogné la carafe contre l'évier. Elle s'est cassée.
[Neutre, Achèvement]

Le symbole Va désigne un verbe d'action construit à l'actif avec l'agent humain H comme sujet, l'entité X figurant dans la séquence des compléments de Va, à une profondeur quelconque. Le symbole Vc désigne un verbe causatif. Par définition, un verbe causatif comme *casser* a une construction transitive, *Luc a cassé la carafe*, qui désigne un accomplissement et qui reçoit une décomposition lexicale glosée de la façon suivante : Luc a agi sur la carafe, ce qui a entraîné que la carafe est devenue cassée. Un verbe causatif a en général une construction neutre qui désigne un achèvement, celle-ci peut être pronominale ou non[6] :

[6] Toutefois, certains verbes causatifs n'ont pas de construction neutre. C'est le cas de *tuer* : la phrase *Marie a tué* ne peut pas être interprétée

La carafe s'est cassée
Le porte-avions a coulé

Certains verbes ont les deux constructions neutres, e.g. *La carafe s'est cassée* versus *La carafe a cassé* [7]. Pour les discours (CR2), le fait que la construction neutre soit ou non pronominale n'intervient pas. Nous regroupons donc les formes *X se Vc* et les formes *X Vc* (regroupement condensé en *X (se) Vc*), rejoignant par là-même les positions de Ruwet (1972) et de Boons et alii (1976) sur l'assimilation de ces formes. Rappelons que les faits empiriques que nous allons mettre en évidence sont valides en anglais comme en français[8]. Or l'anglais n'a qu'une seule réalisation de la construction neutre, la forme non pronominale :

The carafe broke (= La carafe s'est cassée)
The aircraft carrier sank (= Le porte-avions a coulé)

Ceci constitue donc un argument supplémentaire pour regrouper sous une seule structure, (CR2), les discours dont le résultat est exprimé par un verbe causatif employé dans une construction neutre.

Les discours de structure (CR1) et (CR2) seront mis en parallèle avec la structure (CR3) où le verbe causatif a un aspect statique :

(CR3) H Va (passé composé) ... X .... . X *être* (présent) Vc-pp.
(3) =: Luc a cogné la carafe contre l'évier. Elle est cassée.
[Participe passé adjectival, état]

---

avec *Marie* dans le rôle de patient ; la forme pronominale n'a qu'une interprétation réfléchie (*Marie s'est tuée*) ou une interprétation à agent fantôme (Boons et alii 1976) *Les mouches se tuent facilement cette année.*

7 Cependant, pour *casser*, nous travaillerons exclusivement sur les discours (CR2) construits avec la forme *La carafe s'est cassée.* Nous laissons le soin au lecteur de vérifier que les faits décrits ne changent pas avec la forme *La carafe a cassé.*

8 Les faits empiriques sont valides en français et en anglais pour (CR1) et (CR2) mais aussi pour la structure (CR3) présentée ci-dessous.

1.1 Modifieurs interdits

Les secondes phrases de (CR1) ou (CR2) prises isolément acceptent un ensemble de modifieurs. Or nous allons voir que certains de ces modifieurs sont interdits dans les discours qui nous intéressent.

1.1.1 Compléments circonstanciels de lieu et de date[9]

On devrait s'attendre à pouvoir adjoindre aux secondes phrases de (CR1) et (CR2) décrivant des transitions des compléments circonstanciels de lieu et de date, comme on peut le faire quand elles sont employées isolément :

Luc a cassé la carafe (chez Paul + à minuit)
La carafe s'est cassée (chez Paul + à minuit)

Or on constate que tel n'est pas le cas :

(1a) ¶ Luc a cogné la carafe contre l'évier. Il l'a cassée (chez Paul + à minuit).

(2a) ¶ Luc a cogné la carafe contre l'évier. Elle s'est cassée (chez Paul + à minuit).

Les discours (1a) et (2a) peuvent être considérés comme elliptiques à cause d'une contraction de deux informations : l'état résultant de la carafe (elle est cassée) et le lieu ou la date[10]. Ce type de contraction gêne plus ou moins les locuteurs[11], mais

---

9 Nous appelons complément de "date" des compléments comme *à minuit, hier, jeudi dernier,* etc. Les informations temporelles concernant les durées seront discutées dans la Section 1.2 ci-dessous.

10 Cette contraction d'informations ne s'observe pas dans des discours contenant trois phrases :

Luc a cogné la carafe contre l'évier. Il l'a cassée. Cela s'est passé (chez Paul + à minuit).

Luc a cogné la carafe contre l'évier. Elle s'est cassée. Cela s'est passé (chez Paul + à minuit).

11 Une contraction d'informations s'observe aussi dans une phrase isolée comme *Ma montre est réparée sur l'établi* comportant un participe passé adjectival et un complément de lieu. Cette phrase est jugée comme inacceptable par certains locuteurs et acceptable par d'autres, mais ces derniers identifient quand même une contraction d'informations entre *Ma montre est réparée* et *elle est sur l'établi* (Ch. Leclère, communication personnelle).

tout le monde s'accorde sur le caractère elliptique de (1a) et (2a) ainsi que sur le contraste entre ces discours elliptiques et ceux obtenus en insérant les compléments de lieu ou de date dans la première phrase qui sont parfaits :

(1b) Luc a cogné la carafe contre l'évier (chez Paul + à minuit). Il l'a cassée.

(2b) Luc a cogné la carafe contre l'évier (chez Paul + à minuit). Elle s'est cassée.

Les discours présentant un caractère elliptique sont précédés du signe ¶. Nous évitons ainsi le signe * réservé à l'inacceptabilité grammaticale ou interprété comme tel. Le signe ¶ précède aussi les discours qui n'ont pas la sémantique voulue, ici une sémantique de relation causale directe. On notera d'ailleurs que (1a) peut à la rigueur être interprété comme une succession dans le temps de deux événements indépendants. Dans cette interprétation, la seconde phrase ne met pas en jeu un phénomène de contraction d'informations. Les discours précédés du signe ¶ seront qualifiés d'inacceptables, étant bien entendu que cette notion d'inacceptabilité n'a rien de syntaxique.

L'inacceptabilité de (1a) et (2a) est inattendue, mais pas celle de (3a) où la seconde phrase est construite autour d'un participe passé adjectival à aspect statique :

(3a) * Luc a cogné la carafe contre l'évier. Elle est cassée (chez Paul + à minuit).

qui peut s'expliquer par l'inacceptabilité de la seconde phrase prise isolément :

* La carafe est cassée (chez Paul + à minuit)

En effet, on peut affirmer que l'inacceptabilité de (3a) relève de la règle suivante :

(R1) *Si un discours contient au moins une phrase inacceptable hors contexte, alors il est inacceptable.*

Cette règle, de portée très générale, ne souffre pas d'exception si on met de côté certains discours oraux, éventuellement accompagnés de gestes, qui s'apparentent à du pur charabia. En revanche, les exemples (1a) et (2a) infirment l'implication réciproque de la règle (R1) soit :

(R2) *Si un discours ne contient que des phrases acceptables hors contexte, alors il est acceptable (modulo les questions de cohérence et de sémantique).*

Nous dirons qu'une structure de discours a un comportement "régulier" si elle débouche sur des discours respectant les règles (R1) et (R2), un comportement "incongru" si la règle (R2) est violée. L'interdiction d'insérer un complément de lieu ou de date dans la seconde phrase de (CR1) et (CR2) montre que ces structures de discours sont incongrues.

La seule façon d'apporter des informations sur le lieu et la date dans les structures (CRi) consiste donc à les insérer dans la première phrase de ces discours. Précisons cependant le point suivant : dans un discours (CR2), il est possible d'adjoindre dans la seconde phrase un complément de date si tant est qu'il en existe un dans la première phrase (référant à une date antérieure) :

(2c) Hier, les Japonais ont envoyé une bombe sur le porte-avions. Il a coulé ce matin.

(2d) Les Japonais ont envoyé une bombe sur le porte-avions à 3h. Il a coulé à 5h.

Il est donc possible avec (CR2) d'exprimer un décalage de temps entre l'action décrite dans la première phrase et le résultat décrit dans la seconde. Ceci n'est pas autorisé avec les discours (CR1). En effet, les exemples

(1c) ¶ Hier, les Japonais ont envoyé une bombe sur le porte-avions. Ils l'ont coulé ce matin.

(1d) ¶ Les Japonais ont envoyé une bombe sur le porte-avions à 3h. Il l'ont coulé à 5h.

n'ont pas la sémantique voulue : ils ne s'interprètent pas comme une relation causale mais comme une succession de deux événements indépendants. D'où le signe ¶ devant ces exemples.

1.1.2 Modifieurs orientés vers l'agent

Une phrase comportant un agent, comme *Luc a cassé la carafe*, permet normalement l'adjonction d'adverbiaux

(adverbes, groupes prépositionnels ou subordonnées) orientés vers cet agent :

> Luc a cassé la carafe (désinvoltement + de manière désinvolte + avec désinvolture + pour attirer l'attention sur lui + en rêvant à sa fiancée + par esprit de vengeance).

Or ce type d'adverbiaux ne peut pas apparaître dans la seconde phrase de (CR1) :

(1e) ¶ Luc a cogné la carafe contre l'évier. Il l'a cassée (désinvoltement + de manière désinvolte + avec désinvolture + pour attirer l'attention sur lui + en rêvant à sa fiancée + par esprit de vengeance).

Apportons une précision. Un discours comme le suivant extrait des exemples (1e) :

($1e_1$) ¶ Luc a cogné la carafe contre l'évier. Il l'a cassée désinvoltement.

devient acceptable si on insère dans la seconde phrase un connecteur comme *ainsi* ou *donc* :

($1e_2$) ¬ø Luc a cogné la carafe contre l'évier. Il l'a (ainsi + donc) cassée désinvoltement.

Outre la présence d'un tel connecteur, ($1e_1$) et ($1e_2$) diffèrent par leur contexte d'énonciation : ($1e_2$) est parfait dans un contexte où le fait que Luc a cassé la carafe est déjà connu par l'interlocuteur ; les informations apportées par ($1e_2$) sont alors les suivantes : la manière dont Luc a cassé la carafe, information exprimée dans la première phrase, et le jugement du locuteur sur cette manière, à savoir désinvolte. Or nous ne nous intéressons aux discours (CRi) que dans un contexte gauche nul, ce qui suppose que toutes les informations véhiculées sont nouvelles pour l'interlocuteur. Les discours qui ne sont acceptables que dans un contexte gauche non nul sont précédés du signe ¬ø, comme c'est le cas pour ($1e_2$).

L'inacceptabilité de (1e) est inattendue - c'est une exception à la règle (R2) - mais pas celle de (2e) ou (3e) ci-dessous qui peut s'expliquer par l'inacceptabilité de la seconde phrase prise isolément, donc par la règle (R1) :

(2e) * Luc a cogné la carafe contre l'évier. Elle s'est cassée (désinvoltement + de manière désinvolte + avec désinvolture + pour attirer l'attention sur lui + en rêvant à sa fiancée + par esprit de vengeance).

(3e) * Luc a cogné la carafe contre l'évier. Elle est cassée (désinvoltement + de manière désinvolte + avec désinvolture + pour attirer l'attention sur lui + en rêvant à sa fiancée + par esprit de vengeance).

La seule façon d'introduire ces adverbiaux dans les discours (CRi) consiste à les faire apparaître dans la première phrase :

(1f) Luc a cogné la carafe contre l'évier (désinvoltement + de manière désinvolte + avec désinvolture + pour attirer l'attention sur lui + en rêvant à sa fiancée + par esprit de vengeance). Il l'a cassée.

(2f) Luc a cogné la carafe contre l'évier (désinvoltement + de manière désinvolte + avec désinvolture + pour attirer l'attention sur lui + en rêvant à sa fiancée + par esprit de vengeance). Elle s'est cassée.

(3f) Luc a cogné la carafe contre l'évier (désinvoltement + de manière désinvolte + avec désinvolture + pour attirer l'attention sur lui + en rêvant à sa fiancée + par esprit de vengeance). Elle est cassée.

Parmi les adverbes orientés vers le sujet, il est classique (voir entre autres Molinier 1990) de distinguer les adverbes de phrases :

Cruellement, Luc a démoli le château de cartes de sa petite sœur.
= Luc a été cruel de démolir le château de cartes de sa petite sœur.

des adverbes de verbe :

Luc a tué cette mouche cruellement.
= Luc a tué cette mouche de manière cruelle.

Ces deux types d'adverbes sont interdits dans la seconde phrase des discours (CR1) :

($1e_3$) ¶ Luc a soufflé sur le château de cartes de sa petite sœur. Cruellement, il l'a démoli.

($1e_4$) ¶ Luc a brûlé les pattes de cette mouche. Il l'a tuée cruellement.

Nous verrons cependant dans la Section 2.3 que ces deux interdictions reçoivent des explications différentes. Précisons aussi que les instrumentaux sont interdits dans la seconde phrase :

¶ Luc a tapé sur la carafe. Il l'a cassée avec un couteau.
Luc a tapé sur la carafe avec un couteau. Il l'a cassée.

tout comme l'adverbe *volontairement* ou un adverbial sémantiquement proche (e.g. *en le faisant exprès*) :

¶ Luc a cogné la carafe contre l'évier. Il l'a cassée volontairement.[12]

Notons le point suivant : la phrase *Luc a cogné la carafe contre l'évier* est ambiguë dans la mesure où Luc peut avoir agi volontairement ou non (voir *Luc a cogné la carafe contre l'évier pour attirer l'attention sur lui* versus *Luc a cogné la carafe contre l'évier en la lavant*). Néanmoins, l'interdiction d'insérer volontairement dans la seconde phrase de (CR1) est sans rapport avec cette ambiguïté : cette interdiction reste valide si la première phrase décrit une action effectuée volontairement par l'agent, comme dans :

¶ Luc a lancé la carafe contre le mur. Il l'a cassée volontairement.

Par contre, nous verrons dans la section suivante que l'adverbe *involontairement* ou un adverbial sémantiquement proche (e.g. *sans le faire exprès)* est éventuellement autorisé dans un contexte approprié :

Bébé a cogné la carafe contre l'évier. Bien involontairement, il l'a cassée.

1.2 Modifieurs autorisés

Nous venons de passer en revue les modifieurs autorisés dans la seconde phrase de (CR1) ou (CR2) prise isolément mais interdits dans les discours (CR1) ou (CR2). Nous proposerons dans la Section 2 des explications à ces interdictions. Pour que nos explications soient valides, il faut qu'elles bloquent les modifieurs interdits mais qu'elles ne bloquent pas les modifieurs autorisés. Nous allons donc examiner quels sont les modifieurs qui sont autorisés dans la seconde phrase de (CRi) que celle-ci soit prise isolément ou plongée dans (CRi). Pour

---

[12] Comme pour ($1e_1$), cet exemple devient acceptable si l'on insère un connecteur comme *donc* ou *ainsi* et si l'on sait déjà que Luc a cassé la carafe (contexte gauche non nul) :

¬Ø Luc a cogné la carafe contre l'évier. Il l'a (ainsi + donc) cassée volontairement.

cela, nous suivrons la classification des adverbes présentée dans Molinier (1990).

Parmi les adverbes de phrase non conjonctifs[13], les évaluatifs comme *malheureusement* et les modaux comme *naturellement* sont autorisés dans les trois types de discours qui nous intéressent :

(1g) Luc a cogné la carafe contre l'évier. (Malheureusement + naturellement), il l'a cassée.

(2g) Luc a cogné la carafe contre l'évier. (Malheureusement + naturellement), elle s'est cassée.

(3g) Luc a cogné la carafe contre l'évier. (Malheureusement + naturellement), elle est cassée.

Une remarque : considérons les exemples suivants qui sont tous trois maladroits :

(1h) ? Luc a lancé la carafe contre un mur. Il l'a cassée.

(2h) ? Luc a lancé la carafe contre un mur. Elle s'est cassée.

(3h) ? Luc a lancé la carafe contre un mur. Elle est cassée.

Le côté maladroit de ces discours s'explique par un effet de redondance pragmatique : la seconde phrase est inférable à partir de la première grâce à notre connaissance du monde. On sait en effet que si une carafe (en verre) est lancée contre un mur, il y a de fortes chances pour qu'elle soit cassée. Cette redondance pragmatique est pardonnée si elle est annoncée par un adverbial modal comme *naturellement*, *comme de bien entendu, comme on pouvait s'y attendre* :

Luc a lancé la carafe contre un mur. Naturellement, il l'a cassée.

Luc a lancé la carafe contre un mur. Naturellement, elle s'est cassée.

Luc a lancé la carafe contre un mur. Naturellement, elle est cassée.

---

[13] Les adverbes conjonctifs, et plus généralement les connecteurs, seront étudiés dans la section suivante. Rappelons que cette section-ci concerne les modifieurs qui sont acceptables dans la seconde phrase de (CRi) prise isolément ou plongée dans (CRi). Elle ne concerne donc pas les connecteurs qui n'interviennent que dans un discours.

Soulignons que c'est le fait que le locuteur sache qu'en lançant une carafe contre un mur on la casse qui rend les discours (h) maladroits et non le fait que l'agent ait cette connaissance. En effet, on observe cette même maladresse avec des agents qui n'ont pas cette connaissance, par exemple un bébé ou un adulte en pleine crise de somnambulisme :

? Luc a lancé la carafe contre un mur en pleine crise de somnambulisme. Il l'a cassée.
? Luc a lancé la carafe contre un mur en pleine crise de somnambulisme. Elle s'est cassée.
? Luc a lancé la carafe contre un mur en pleine crise de somnambulisme. Elle est cassée.

Ces discours présentent le même caractère maladroit que les exemples (h), qui disparaît en ajoutant un adverbe comme *naturellement* :

Luc a lancé la carafe contre un mur en pleine crise de somnambulisme. Naturellement, il l'a cassée.
Luc a lancé la carafe contre un mur en pleine crise de somnambulisme. Naturellement, elle s'est cassée.
Luc a lancé la carafe contre un mur en pleine crise de somnambulisme. Naturellement, elle est cassée.

Parmi les adverbes de verbe, les quantifieurs sont autorisés :

Luc a cogné la carafe contre l'évier. Il l'a sérieusement fêlée.
Luc a cogné la carafe contre l'évier. Elle s'est sérieusement fêlée.
Luc a cogné la carafe contre l'évier. Elle est sérieusement fêlée.

(1i) Luc a cogné la carafe contre l'évier. Il l'a cassée (en deux + en mille morceaux).
(2i) Luc a cogné la carafe contre l'évier. Elle s'est cassée (en deux + en mille morceaux).
(3i) Luc a cogné la carafe contre l'évier. Elle est cassé (en deux + en mille morceaux).

Passons aux adverbiaux de durée. Les secondes phrases de (CR1) et (CR2) prises isolément permettent l'adjonction d'un complément de durée introduit par *en* contrairement à la seconde phrase de (CR3) :

Les Japonais ont coulé le porte-avions hier en 1 heure
Le porte-avions a coulé en 1 heure
* Le porte-avions est coulé en 1 heure[14]

Ces compléments de durée sont autorisés dans la seconde phrase de (CR1) ou (CR2) :

($1j_1$) Les Japonais ont envoyé des bombes sur le porte-avions. Ils l'ont coulé en 1 heure.
($2j_1$) Les Japonais ont envoyé des bombes sur le porte-avions. Il a coulé en 1 heure.

On notera cependant le contraste suivant entre (CR1) et (CR2) :

($1j_2$) ¶ Les Japonais ont envoyé une bombe sur le porte-avions. Ils l'ont coulé en 1 heure.
($2j_2$) Les Japonais ont envoyé une bombe sur le porte-avions. Il a coulé en 1 heure.

Les exemples (j_1) et (j_2) diffèrent par le fait que *bombe* est au pluriel dans le premier cas et au singulier dans le second, ce qui implique sur le plan aspectuel que l'action dans (j_1) est un procès duratif tandis qu'elle est ponctuelle dans (j_2). Or on constate que les exemples ($2j_1$) et ($2j_2$) ont tous deux une sémantique causale[15] tandis que seul ($1j_1$) a cette sémantique : ($1j_2$) est déviant, en tout cas n'a pas une sémantique causale, d'où notre signe ¶. Le contraste entre ($1j_2$) et ($2j_2$) est à rapprocher de celui que nous avons observé entre la distribution des dates dans les discours (CR1) et (CR2) (voir Section 1.1.1), contraste que nous rappelons :

¶ Hier, les Japonais ont envoyé une bombe sur le porte-avions. Ils l'ont coulé ce matin.
Hier, les Japonais ont envoyé une bombe sur le porte-avions. Il a coulé ce matin.

---

[14] Si le participe passé adjectival est construit avec le verbe *être* conjugué à l'imparfait, un complément de durée introduite par *en* est à la limite acceptable :
? Les Japonais ont envoyé des bombes sur le porte-avions. Il était coulé en 1 heure.
Néanmoins, nous laissons de côté de tels discours.

[15] Nous verrons en fait dans la Section 2.3 que ($2j_1$) et ($2j_2$) ont tous deux une sémantique causale mais que ($2j_1$) exprime une relation causale "directe", tandis que ($2j_2$) exprime une relation causale "indirecte".

Ceci confirme la règle que nous avions avancée : il est possible dans (CR2), mais pas dans (CR1) d'exprimer un décalage de temps entre l'action décrite dans la première phrase et le résultat décrit dans la seconde. Plus précisément, les discours (j1) ont tous deux la même sémantique de relation causale qui se glose de la façon suivante : l'agent (i.e. les Japonais) a commencé son action (i.e. envoyer des bombes sur le porte avion) à l'instant T0 et il a continué son action jusqu'à l'instant T1, instant où le porte-avions a coulé (changement d'état considéré comme ponctuel), avec l'équation T1 - T0 = 1 heure. Il n'y a donc pas de décalage dans le temps entre la fin de l'action et le résultat (le changement d'état) et les discours (CR1) et (CR2) sont tous deux acceptables. Par contre, le discours ($2j_2$) a une sémantique qui se glose de la façon suivante : l'agent (i.e. les Japonais) a effectué une action ponctuelle (i.e. envoyer une bombe sur le porte-avions) à l'instant T0 ; cette action ponctuelle a eu comme conséquence un changement d'état qui a eu lieu à l'instant T1 avec l'équation T1 - T0 = 1 heure. Il y a donc un décalage d'une heure entre la fin de l'action et le résultat. Un tel décalage n'est pas exprimable dans un discours (CR1) comme en témoigne le caractère déviant de ($1j_2$). Ces faits empiriques seront à nouveau confirmés dans la section suivante sur les connecteurs et recevront une explication dans la Section 2.

Enfin, examinons l'insertion d'un adverbe comme *involontairement* ou *sans le faire exprès* dans la seconde phrase de (CR1), insertion qui est acceptable dans une phrase isolée :

Luc a cassé la carafe involontairement.

Rappelons que l'insertion de *volontairement* ou *en le faisant exprès* est exclue dans la seconde phrase de (CR1) :

($1k_1$) ¶ Luc a tapé sur la carafe avec un couteau. Il l'a cassée volontairement.

Si l'on substitue *involontairement* à *volontairement* dans ($1k_1$), on obtient un discours d'une acceptabilité encore douteuse quoique meilleure :

($1l_1$) ?¶ Luc a tapé sur la carafe avec un couteau. Il l'a cassée involontairement.

Par contre, si l'on ajoute un complément de but dans la première phrase, le contraste entre *volontairement* et *involontairement* devient plus net :

($1k_2$) ¶ Luc a tapé sur la carafe avec un couteau pour attirer l'attention sur lui. Il l'a cassée volontairement.

($1l_2$) ? Luc a tapé sur la carafe avec un couteau pour attirer l'attention sur lui. Il l'a cassée involontairement.

Le discours ($1k_2$) ne fait carrément pas sens, tandis que le discours ($1l_2$) est presque naturel. Ce dernier discours devient tout à fait naturel s'il y a une emphase sur le côté involontaire de la chose, grâce au déplacement en tête de phrase de *involontairement* et au renforcement par l'adverbe *bien* [16]:

($1l_3$) Luc a tapé sur la carafe avec un couteau pour attirer l'attention sur lui. Bien involontairement, il l'a cassée.

Notons aussi l'acceptabilité de discours comme :

Bébé a soufflé sur le château de cartes de Zoé. Bien involontairement, il l'a démoli.
Luc a étranglé sa femme en pleine crise de *delirium tremens*. Bien involontairement, il l'a tuée.

On retiendra donc que *volontairement* ou un adverbial de sens proche est totalement exclu dans la seconde phrase de (CR1) mais que *involontairement* ou un adverbial de sens proche peut y apparaître dans un contexte approprié. Ces faits recevront une explication dans la Section 1.4

1.3 Connecteurs

Parmi les connecteurs ou les adverbes de phrase conjonctifs (dans la terminologie de Molinier (1990), les connecteurs "consécutifs" sont en général autorisés pour introduire les secondes phrases de (CRi) :

---

[16] Il est envisageable de considérer que *involontairement* est un adverbe de verbe dans ($1l_2$) et un adverbe de phrase exprimant le jugement du locuteur dans ($1l_3$). Néanmoins cette distinction n'est pas très nette. Voir aussi la remarque B) dans la Conclusion.

(1m) Luc a cogné la carafe contre l'évier. (De ce fait + Par là-même + En conséquence de quoi), il l'a cassée.

(2m) Luc a cogné la carafe contre l'évier. (De ce fait + Par là-même + En conséquence de quoi), elle s'est cassée.

(3m) Luc a cogné la carafe contre l'évier. (De ce fait + Par là-même + En conséquence de quoi), elle est cassée.

Il n'est pas vrai pour autant que tout connecteur consécutif soit valide pour chaque discours (CRi). Par exemple, le connecteur *ceci a eu comme conséquence (pour X) que* semble adapté pour les discours (CR2) mais malheureux pour les discours (CR1) :

Luc a cogné la carafe contre l'évier. Ceci a eu comme conséquence (pour la carafe) qu'elle s'est cassée.

?¶ Luc a cogné la carafe contre l'évier. Ceci a eu comme conséquence (pour la carafe) qu'il l'a cassée.

À l'inverse, le connecteur *ce faisant (= en faisant cela)* est autorisé en (CR1) et (syntaxiquement[17]) interdit en (CR2) et (CR3) :

Luc a cogné la carafe contre l'évier. Ce faisant, il l'a cassée.

* Luc a cogné la carafe contre l'évier. Ce faisant, elle s'est cassée.

* Luc a cogné la carafe contre l'évier. Ce faisant, elle est cassée.

Les conjonctifs "reformulatifs" comme *autrement dit* sont autorisés et même parfois recommandés. En effet, considérons les discours suivants qui sont maladroits :

? Luc a coupé l'annulaire gauche de Marie. Il l'a mutilée.

? Luc a coupé l'annulaire gauche de Marie en pleine crise de somnambulisme. Il l'a mutilée.

Le côté maladroit de ces deux discours s'explique par un effet de redondance lexicale : le premier prédicat, *couper un annulaire*, correspond à la définition de *mutiler*. Cette redondance lexicale est pardonnée si elle est annoncée par un reformulatif comme *autrement dit* :

Luc a coupé l'annulaire gauche de Marie. Autrement dit, il l'a mutilée.

Luc a coupé l'annulaire gauche de Marie en pleine crise de somnambulisme. Autrement dit, il l'a mutilée.

---

[17] À cause de la (non) coréférence des sujets.

Les conjonctifs reformulatifs effacent donc la redondance lexicale comme les adverbes de phrases modaux effacent la redondance pragmatique (voir Section 1.2).

Tournons-nous maintenant vers les connecteurs temporels comme *immédiatement (= immédiatement après cela)* ou *1 heure après (= 1 heure après cela).* Ces connecteurs sont autorisés dans (CR2) mais interdits dans (CR1) et (CR3) :

(1n) ¶ Les Japonais ont envoyé (une + des) bombe(s) sur le porte-avions. (Ils l'ont immédiatement coulé + Ils l'ont coulé immédiatement/1 heure après).

(2n) Les Japonais ont envoyé (une + des) bombe(s) sur le porte-avions. (Il a immédiatement coulé + Il a coulé immédiatement/1 heure après).

(3n) ¶ Les Japonais ont envoyé (une + des) bombe(s) sur le porte-avions. Il est immédiatement coulé + Il est coulé immédiatement/1 heure après).[18]

Le contraste entre (1n) et (2n) renforce l'observation déjà faite dans les Sections 1.1.1 et 1.2 : il est impossible dans (CR1) - mais possible dans (CR2) - d'exprimer un décalage même infinitésimal (voir *immédiatement*) entre la fin de l'action (quelle soit durative ou ponctuelle) et le résultat.

1.4 Contraintes lexicales

Comparons les discours suivants dont la seconde phrase est construite respectivement autour de *se tuer* et *se suicider* (avec H = X) :

(1o) Beregovoy s'est tiré une balle dans la tête. Il s'est tué.

(1p) ¶ Beregovoy s'est tiré une balle dans la tête. Il s'est suicidé.

On constate que (1o) est acceptable mais que (1p) ne l'est pas (Danlos 1985) : il est impossible qu'un commentateur à la radio ou à la télévision apprenne à ses auditeurs le suicide de Beregovoy par le discours (1p). Par contre, si la mort de

---

[18] L'acceptabilité de ce discours est amélioré si *être* est conjugué à l'imparfait (voir note 14) :

? Les Japonais ont envoyé une bombe sur le porte-avions. (Il était immédiatement coulé + Il était coulé immédiatement/1 heure après.)

Beregovoy est un fait déjà connu (contexte gauche non nul, ce qui n'est pas dans le champ de notre étude), les discours suivants sont acceptables :

¬ø Beregovoy s'est tiré une balle dans la tête. Il s'est donc suicidé.
¬ø Beregovoy s'est tiré une balle dans la tête. Il s'est ainsi suicidé d'une manière spectaculaire.

Signalons qu'en inversant l'ordre des phrases de (1p), on obtient un discours parfait dans un contexte gauche nul :

Beregovoy s'est suicidé. Il s'est tiré une balle dans la tête.

La différence d'acceptabilité entre (1o) et (1p) ne peut être expliquée que par la différence sémantique entre *se tuer* et *se suicider*. On peut poser que la sémantique de *se suicider* est "se tuer en agissant délibérément et en voulant sa propre mort". Dans (1o) et (1p), on sait que Beregovoy s'est tué en agissant délibérément (i.e. en se tirant une balle dans la tête), mais on sait de plus dans (1p), grâce a la sémantique de *se suicider*, que Beregovoy voulait sa propre mort. La seule façon d'expliquer la différence d'acceptabilité entre (1o) et (1p) consiste donc à poser la règle suivante :

(r1) le discours (1p) est inacceptable parce qu'il véhicule l'information que Beregovoy voulait sa propre mort,
le discours (1o) est acceptable parce rien n'indique que Beregovoy voulait sa propre mort (qu'il l'ait voulue ou pas).

On observe exactement le même phénomène avec la paire *tuer/assassiner*. En effet, on constate que des discours suivants :

(1q) Luc a poussé (volontairement) Marie par la fenêtre. Il l'a tuée.
(1r) ¶ Luc a poussé (volontairement) Marie par la fenêtre. Il l'a assassinée.

seul le discours (1q) est acceptable dans un contexte gauche nul. On peut poser que la sémantique de *assassiner quelqu'un* est "tuer cette personne en agissant délibérément et en voulant sa mort". On peut donc émettre la règle suivante, similaire à la règle (r1) :

(r2) le discours (1r) est inacceptable parce qu'il véhicule l'information que Luc voulait la mort de Marie,
le discours (1q) est acceptable parce que rien n'indique que Luc voulait la mort de Marie (qu'il l'ait voulue ou pas).

Les paires comme *se tuer / se suicider* ou *tuer / assassiner* sont malheureusement à notre connaissance assez rares. Ainsi il n'existe pas de verbe qui signifierait "casser un objet en agissant délibérément et en voulant que cet objet soit cassé". Néanmoins, au vu des règles (r1) et (r2), il semble que l'on peut avancer l'hypothèse que des discours comme :

> Luc a cogné la carafe contre l'évier (volontairement + involontairement + E[19]). Il l'a cassée.

n'indiquent nullement que Luc voulait que la carafe soit cassée, que l'action exprimée dans la première phrase soit volontaire, involontaire ou ambiguë. Plus généralement, on peut avancer l'hypothèse suivante :

(H1) *Un discours de structure (CR1) ne peut être acceptable avec une sémantique de relation causale directe que si rien n'indique que l'agent H voulait le résultat (qu'il l'ait voulu ou pas).*

Cette hypothèse, par laquelle le résultat est présenté comme une simple conséquence de l'action décrite dans la première phrase sans aucune relation avec l'éventuel but de l'agent, est confirmée par les faits suivants :

- possibilité d'introduire un complément de but de l'agent dans la première phrase, ce qui montre à l'évidence que le résultat décrit dans la seconde phrase n'était pas le but de cet agent :

> Luc a cogné la carafe contre l'évier pour attirer l'attention sur lui. Il l'a cassée.
> Luc a sauté du premier étage pour épater Marie. Il s'est tué.[20]

---

[19] Le symbole E représente la séquence vide.

[20] En substituant *se suicider* à *se tuer* dans ce dernier exemple :
* Luc a sauté du premier étage pour épater Marie. Il s'est suicidé.
on obtient un discours non pas "elliptique" comme (1p) mais sémantiquement incohérent. Dans le même ordre d'idées, on contrastera les exemples suivants où la seconde phrase est à la forme négative :
Luc a sauté du premier étage pour épater Marie. Heureusement, il ne s'est pas

- interdiction d'introduire un adverbe comme *volontairement* dans la seconde phrase de (CR1) (voir Section 1.1.2), que l'action exprimée dans la première phrase soit volontaire, involontaire *(a fortiori)* ou ambiguë :

¶ Luc a lancé la carafe contre le mur. Il l'a cassée volontairement.
¶ Luc a renversé son café sur la nappe sans le faire exprès. Il l'a salie volontairement.
¶ Luc a cogné la carafe contre l'évier. Il l'a cassée volontairement.

- possibilité d'introduire un adverbe comme *involontairement* dans un contexte approprié (voir Section 1.2) :

(1s) Luc a tapé sur la carafe avec un couteau pour attirer l'attention sur lui. Bien involontairement, il l'a cassée.
Bébé a soufflé sur le château de cartes de Zoé. Bien involontairement, il l'a démoli.

Notons que le fait que l'introduction de *involontairement* donne un résultat douteux dans un contexte "non approprié" ;

(1t) ?¶ Luc a tapé sur la carafe avec un couteau. Il l'a cassée involontairement.

renforce l'hypothèse (H1). En effet, cette hypothèse prédit le caractère douteux de (1t) par un effet de redondance. Cette redondance est pardonnée si elle est soulignée comme en (1s).

- caractère maladroit des deux discours suivants (voir Section 1.2) :

(1u) ? Luc a lancé la carafe contre le mur. Il l'a cassée.

(1v) ? Luc a lancé la carafe contre le mur en pleine crise de somnambulisme. Il l'a cassée.

Il y a une différence dans l'intentionnalité de Luc dans ces deux discours puisque Luc n'a aucune velléité particulière en (1v) tandis qu'il peut en avoir en (1u), entre autres que la carafe soit cassée. Or, malgré cette différence d'intentionnalité, ces deux discours dégagent la même impression de redondance

---

tué.
* Luc a sauté du premier étage pour épater Marie. Heureusement, il ne s'est pas suicidé.

pragmatique, qui est pardonnée si elle est annoncée par un modal comme *naturellement* :

> Luc a lancé la carafe contre le mur. Naturellement, il l'a cassée.
> Luc a lancé la carafe contre le mur en pleine crise de somnambulisme. Naturellement, il l'a cassée.

Ceci confirme que le résultat dans (CR1) est présenté comme une conséquence de l'action décrite dans la première phrase, quel que soit l'(éventuel) but de l'agent.

En conclusion, nous considérons que nous pouvons valider l'hypothèse (H1). Cette hypothèse nous permettra d'expliquer d'autres phénomènes (voir Section 2.3). On remarquera que les discours (CR2) et (CR3) n'induisent aucune notion de but : le résultat exprimé sous forme d'un achèvement ou d'un état ne peut être qu'une conséquence de l'action de la première phrase. A ce titre là, les discours (CR1), (CR2) et (CR3) ont la même sémantique.

Faisons à ce propos une remarque sur les relations rhétoriques. Il est classique de considérer que les discours (CRi) avec l'ordre CAUSE < RÉSULTAT mettent en jeu la relation rhétorique dite "résultative". Les chercheurs qui travaillent dans le cadre de la "Rhethoric Structure Theory" (RST) et qui définissent les relations rhétoriques de façon purement sémantique (e.g. Hovy & Maier 1995) considèrent généralement que la relation résultative doit être affinée en relation résultative volontaire et relation résultative non volontaire. Knott & Dale (1994) contestent cette distinction en faisant remarquer qu'il n'existe aucun connecteur anglais qui la réalise. L'hypothèse (H1), ainsi que les faits mis en avant dans la Section 3, abonde dans le sens de Knott & Dale puisqu'elle revient à poser le postulat suivant : la relation résultative volontaire est définissable sur des bases purement sémantiques, mais elle n'est pas exprimable dans la langue (du moins en français et en anglais). En se basant sur des faits linguistiques, il ne faut définir qu'une seule relation résultative, celle-ci étant non marquée sur l'aspect voulu ou non du résultat, seul l'aspect involontaire pouvant être éventuellement indiqué dans un contexte marqué.

1.5 Présence ou absence d'un argument

Jusqu'à présent, nous n'avons étudié les discours parataxiques CAUSE < RÉSULTAT que lorsque la cause est exprimée à l'actif, donc mentionnant l'agent H, le résultat mentionnant cet agent dans (CR1) mais pas dans (CR2) ni dans (CR3) :

(1w) Luc a cogné la carafe contre l'évier. Il l'a cassée.
(2w) Luc a cogné la carafe contre l'évier. Elle s'est cassée.
(3w) Luc a cogné la carafe contre l'évier. Elle est cassée.

Considérons les discours suivants où la CAUSE est exprimée au passif sans agent, le verbe causatif étant construit de différentes façons :

(1x) ¶ La carafe a été cognée contre l'évier. Luc l'a cassée.
(2x) La carafe a été cognée contre l'évier. Elle s'est cassée.
(3x) La carafe a été cognée contre l'évier. Elle est cassée.

Le discours (1x), qui ne mentionne pas d'agent dans la première phrase mais qui en mentionne un dans la seconde, n'a pas une sémantique de relation causale directe : il a une sémantique de succession temporelle de deux événements, le premier étant une action effectuée par un agent non spécifié mais probablement différent de Luc, le second étant une action effectuée par Luc. Par contre, les discours (2x) et (3x), qui ne mentionnent d'agent ni dans la première phrase ni dans la seconde, ont une sémantique de relation causale directe.

Pour compléter notre propos, considérons les discours dont la cause est exprimée au passif avec agent :

(1y) La carafe a été cognée contre l'évier par Luc. Il l'a cassée.
(2y) La carafe a été cognée contre l'évier par Luc. Elle s'est cassée.
(3y) La carafe a été cognée contre l'évier par Luc. Elle est cassée.

Ces trois discours ont une sémantique de relation causale directe comme les discours (w) où la cause est à l'actif.

Les paradigmes (w), (x) et (y) que nous venons de présenter conduisent à poser la règle suivante :

(R3) *Si l'agent est mentionné dans la CAUSE (structure à l'actif ou au passif avec agent), le RÉSULTAT peut mentionner ou non cet agent, mais si l'agent n'est pas mentionné dans la CAUSE (structure au passif sans agent), le RÉSULTAT ne doit pas*

mentionner d'agent si on veut une sémantique de relation causale directe.

Cette règle semble assez *ad hoc*. Nous présenterons dans la Section 2.3 une hypothèse qui permet de s'en passer. Notons que cette discussion sur l'agent H n'a pas lieu d'être pour l'entité X : celle-ci figure obligatoirement dans la première et la seconde phrase de (CRi), quelles que soient les constructions syntaxiques de la seconde phrase.

Enfin apportons la précision suivante : les discours (CRi) gardent une sémantique de relation causale si les éléments H et X de la seconde phrase sont des groupes nominaux anaphorisant les éléments correspondants de la première phrase :

Luc a cogné la carafe contre l'évier. Cet imbécile a cassé cet objet de valeur.
Un parisien a cogné la carafe contre l'évier. Ce français l'a cassée.
Luc a cogné la carafe contre l'évier. Ce beau récipient (s'est + est) cassé.

Il n'est donc pas nécessaire que les éléments H et X de la seconde phrase soit des pronoms. Mais bien entendu, il faut qu'ils soient en relation de coréférence avec des éléments de la première phrase : les discours suivants n'ont clairement pas une sémantique de relation causale :

¶ Luc a cogné la carafe contre l'évier. Max l'a cassée.
¶ Les Japonais ont bombardé la ville. Le porte-avions a coulé.
¶ Luc a renversé son café sur la nappe. La moquette est salie.

1.6 Bilan

Les faits empiriques que nous avons mis en évidence sur les discours parataxiques respectant l'ordre CAUSE < RÉSULTAT présentent les discours (CR1) et (CR2) comme incongrus :

- ils violent la règle (R2) (i.e. *discours avec que des phrases acceptables isolément ==> discours acceptable*) ;

- de plus, les discours (CR1) présentent, d'une part, des contraintes lexicales, d'autre part, des restrictions sur la distribution des arguments qui conduisent à la règle *ad hoc*

(R3) (i.e. *pas d'agent dans la cause ==> pas d'agent dans le résultat*).

Les contraintes lexicales pesant sur (CR1) ont été expliquées par l'hypothèse (H1) (i.e. *pas de notion de but*). Nous allons avancer dans la section suivante une autre hypothèse qui explique le comportement incongru de (CR1) et (CR2) et qui permet de se passer de la règle (R3). Auparavant, deux remarques.

A) Nous avons étudié dans les sections précédentes les structures (CR1), (CR2) et (CR3) en laissant de côté d'autres constructions syntaxiques du verbe causatif, à savoir le passif sans et avec agent :

(CR4) H Va (passé composé) ... X .... . X *être* (passé composé) Vc-pp.
[Passif sans agent]

(4) =: Luc a cogné la carafe contre l'évier. Elle a été cassée.

(CR5) H Va (passé composé) ... X .... . X *être* (passé composé) Vc-pp par H.
[Passif avec agent]

(5a) ?* Luc a cogné la carafe contre l'évier. Elle a été cassée par lui.

(5b) ? Luc a cogné la carafe contre l'évier. Elle a été cassée par cet imbécile.

Le discours (5a) où la seconde occurrence de l'agent H apparaît sous forme de pronom fort est maladroit. Cette maladresse est atténuée dans (5b) où la seconde occurrence de H apparaît sous un nom de qualité. Nous allons montrer brièvement que (CR4) et (CR5) ont le même comportement que (CR1) en nous limitant aux points cruciaux, à savoir :

- impossibilité d'insérer un complément de lieu ou de date dans la seconde phrase de (CR4) ou (CR5) :

(4a) ¶ Luc a cogné la carafe contre l'évier. Elle a été cassée (chez Paul + à minuit).

(5a) ¶ Luc a cogné la carafe contre l'évier. Elle a été cassée par cet imbécile (chez Paul + à minuit).

- impossibilité d'insérer un modifieur orienté vers l'agent dans la seconde phrase de (CR4) ou (CR5) :

(4c) ¶ Luc a cogné la carafe contre l'évier. Elle a été cassée (désinvoltement + par esprit de vengeance).

(5c) ¶ Luc a cogné la carafe contre l'évier. Elle a été cassée par cet imbécile (désinvoltement + par esprit de vengeance).

- possibilité d'insérer un adverbe quantifieur dans la seconde phrase de (CR4) ou (CR5) :

(4i) Luc a cogné la carafe contre l'évier. Elle a été cassée en mille morceaux.

(5i) Luc a cogné la carafe contre l'évier. Elle a été cassée par cet imbécile en mille morceaux.

- possibilité d'insérer un adverbe de durée dans la seconde phrase de (CR4) ou (CR5) :

($4j_1$) Les Japonais ont envoyé des bombes sur le porte-avions. Il a été coulé en 1 heure.

($5j_1$) Les Japonais ont envoyé des bombes sur le porte-avions. Il a été coulé par ces imbéciles en 1 heure.

- impossibilité d'insérer un connecteur temporel dans la seconde phrase de (CR4) ou (CR5) :

(4n) ¶ Les Japonais ont envoyé des bombes sur le porte-avions. Il a été coulé 1 heure après.

(5n) ¶ Les Japonais ont envoyé des bombes sur le porte-avions. Il a été coulé par ces imbéciles 1 heure après.

- impossibilité de construire la seconde phrase de (CR4) ou (CR5) avec *assassiner* :

(4r) ¶ Luc a poussé (volontairement) Marie par la fenêtre. Elle a été assassinée.

(5r) ¶ Luc a poussé (volontairement) Marie par la fenêtre. Elle a été assassinée par cet imbécile.

- enfin, le discours (CR5), où l'agent est mentionné dans le résultat, ne permet pas que la cause soit au passif sans agent contrairement à (CR4) :

(4x) La carafe a été cognée contre l'évier. Elle a été cassée.

(5x) ¶ La carafe a été cognée contre l'évier. Elle a été cassée par Luc.

Ces faits confirment la règle (R3) avancée dans la Section 1.5 : si l'agent n'est pas mentionné dans la cause, il ne peut pas être mentionné dans le résultat.

Au total, on peut affirmer que les discours (CR4) et (CR5) ont le même comportement que les discours (CR1). Gardant en tête :

(i) cette affirmation,
(ii) le postulat suivant : le passage, pour une phrase donnée, d'une construction active à une construction passive (avec ou sans agent) ne change pas ses propriétés aspectuelles,
(iii) le fait que les discours (CR1) et (CR2), dont la seconde phrase décrit une transition, ont un comportement incongru contrairement à (CR3) dont la seconde phrase décrit un état,

on peut avancer la conclusion suivante : le comportement d'un discours (CRi) dépend de la valeur aspectuelle de sa seconde phrase, mais il ne dépend pas de la construction syntaxique choisie pour exprimer cette valeur aspectuelle. Il est donc logique d'expliquer le comportement d'un discours (CRi) d'après la valeur aspectuelle de sa seconde phrase, et c'est ce que nous allons faire dans la section suivante.

Dans la suite de cet article, les affirmations avancées sur (CR1) concernent aussi (CR4) et (CR5), sans que nous le mentionnions à chaque fois explicitement.

B) Contrastons les discours (CRi) respectant l'ordre CAUSE < RÉSULTAT avec ceux obtenus en inversant l'ordre des phrases, discours notés (RCi) (RÉSULTAT < CAUSE) :

(RC1) H Vc (passé composé) X. H Va (passé composé) ... X .... .
(1z) Luc a cassé la carafe. Il l'a cognée contre l'évier.

(RC2) X (se) Vc (passé composé). H Va (passé composé) ... X .... .
(2z) ? La carafe s'est cassée. Luc l'a cognée contre l'évier.

(RC3) X *être* (présent) Vc-pp. H Va (passé composé) ... X .... .
(3z) La carafe est cassée. Luc l'a cognée contre l'évier.

(RC4) X *être* Vc-pp (passé composé). H Va (passé composé) ... X .... .
(4z) La carafe a été cassée. Luc l'a cognée contre l'évier.

(RC5) X *être* Vc-pp (passé composé) par H. H Va (passé composé)... X .... .
(5z) La carafe a été cassée par Luc. Il l'a cognée contre l'évier.

D'abord notons l'acceptabilité douteuse de (RC2). Nous n'avons pas à offrir de réelle explication à la différence d'acceptabilité entre (2z) et (2) ou entre les éléments de la paire suivante :

Marie a lancé de l'eau sur le feu. Il s'est éteint.
($2z_1$) ? Le feu s'est éteint. Marie a lancé de l'eau dessus.

Rappelons simplement que différents chercheurs (Ruwet 1972, Boons et alii 1976, van Voorst 1995) ont mis en avant que les constructions neutres *X (se) V* désignaient préférentiellement une "activité indépendante" de X qui se déroule sans le "contrôle d'un agent instigateur". On pourrait donc avancer l'intuition que la première phrase de (RC2), une construction neutre, fait croire de préférence à une activité indépendante de X, mais que cette prévision est contredite dans la seconde phrase, d'où une acceptabilité douteuse. Par contre, l'interprétation préférentielle de la construction neutre dans (CR2) serait gommée par le contenu de la première phrase. Il est bien entendu nécessaire d'étayer cette intuition par des considérations formelles sur la neutralité, ce qui sort du cadre de cette étude.

Les discours RÉSULTAT < CAUSE ne présentent pas les mêmes propriétés que les discours CAUSE < RÉSULTAT, en particulier, à l'inverse de ces derniers, ils ont un comportement régulier dans la mesure où l'insertion d'adverbiaux dans la première phrase est acceptable si et seulement si elle est acceptable hors contexte, autrement ils obéissent et à la règle (R1) et à la règle (R2) :

(1zz) Luc a cassé la carafe (volontairement + chez Paul). Il l'a cognée contre l'évier.
(2zz) ? La carafe s'est cassée (*volontairement + chez Paul). Luc l'a cognée contre l'évier.
(3zz) La carafe est cassée (*volontairement + *chez Paul). Luc l'a cognée contre l'évier.
(4zz) La carafe a été cassée (volontairement + chez Paul). Luc l'a cognée contre l'évier.
(5zz) La carafe a été cassée par Luc (volontairement + chez Paul). Il l'a cognée contre l'évier.

De plus, les discours RÉSULTAT < CAUSE n'ont pas tous la même sémantique : ainsi, les discours (RC2) et (RC3) n'induisent aucune notion de but, mais ceci n'est pas le cas pour (RC1),

(RC4) et (RC5) comme en témoigne le fait que les discours obtenus avec *assassiner* sont acceptables [21]:

(1zzz) Luc a assassiné Marie. Il l'a poussée par la fenêtre.
(4zzz) Marie a été assassinée. Luc l'a poussée par la fenêtre.
(5zzz) Marie a été assassinée par Luc. Il l'a poussée par la fenêtre.

On retiendra que les discours RÉSULTAT < CAUSE n'ont pas un comportement aberrant à l'inverse des discours CAUSE < RÉSULTAT. Dans la section suivante, nous verrons que cette dissymétrie s'explique par le fait que ces deux types de discours mettent en jeu deux notions différentes de coréférence événementielle.

2 Analyse des faits empiriques

Nous allons présenter une analyse des faits empiriques décrits dans la section précédente. Elle s'articule autour de deux points :

- la structure événementielle des transitions proposée par Pustejovsky (1991) qui sera présentée dans la Section 2.1,

- une relation de coréférence événementielle d'un type particulier que nous avons appelée "généralisation" et qui sera présentée dans la Section 2.2.

Notre analyse consiste à avancer que l'interprétation de relation causale directe dans (CR1) et (CR2) passe par une relation de coréférence événementielle de type généralisation. Elle sera présentée dans la Section 2.3 où nous montrerons que les contraintes apparemment *ad hoc* qui pèsent sur les discours (CR1) et (CR2) ne sont rien d'autres que des contraintes de généralisation.

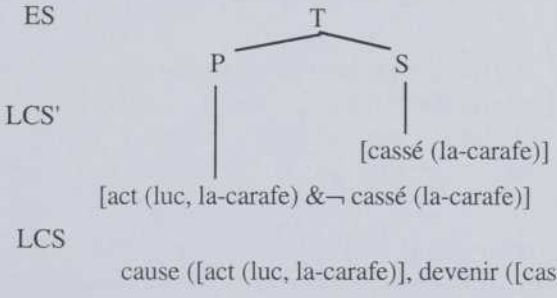
---

[21] Nous avons vu dans la Section 1.4 que la relation rhétorique résultative (discours CAUSE < RÉSULTAT) ne pouvait pas être affinée sur des bases linguistiques en relation résultative volontaire versus involontaire. Ces exemples montrent, par contre, que la relation rhétorique causale (discours RÉSULTAT < CAUSE) peut être affinée sur la base de faits linguistiques en relation causale volontaire versus non volontaire.

2.1 Structure événementielle des transitions

Dans le cadre de la sémantique lexicale, Pustejovsky (1991) introduit un niveau de représentation, "la structure événementielle" (en anglais "event structure" abrégé en ES), qui vient compléter d'autres niveaux de représentation sémantique, entre autres, la structure conceptuelle lexicale (en anglais "lexical conceptual structure" abrégé en LCS) (Jackendoff 1983, 1990). La structure événementielle, qui repose sur la classification vendlerienne - i.e. trois types d'événements : état, procès et transition (accomplissement et achèvement) -, permet de décomposer certains types d'événements en sous-événements. En particulier, les accomplissements qui sont des transitions (notées T) sont décomposés en un procès (noté P) et un état noté S. Intuitivement, cette décomposition repose sur le fait qu'un accomplissement décrit le passage pour une entité X d'un état non-S (noté ¬ S) à un état S comme le résultat d'un procès P, action exercée sur X par un agent humain. L'analyse sémantique de *Luc a cassé la carafe* est présentée ci-dessous :

(I)

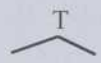


Le niveau LCS' encapsule la décomposition sémantique de chaque événement de la structure événementielle. Le niveau LCS est calculé à partir de la structure événementielle et des informations de LCS'. Nous renvoyons à Pustejovsky (1991) pour plus de détail. Signalons simplement que cette analyse sémantique lui permet de rendre compte de la portée des adverbes, du rôle de la structure argumentale et de la correspondance entre lexique et syntaxe. Son analyse de la portée des adverbes sera présentée dans la Section 2.3.

Revenons aux discours (CR1) :

(CR1) H Va (passé composé) ... X .... . H Vc (passé composé) X.
=: Luc a cogné la carafe contre l'évier. Il l'a cassée.

La structure événementielle proposée par Pustejovsky nous permet d'avancer l'analyse suivante : la première phrase réfère à un événement que nous notons Ea. Cet événement dénote une action qui peut être un procès ou une transition. La seconde phrase réfère à une transition T que nous décomposons en



,en suivant Pustejovsky modulo le fait que nous ne contraignons pas le premier sous-événement à être un procès. L'interprétation de (CR1) comme relation causale directe implique que les événements Ea et Ec soient interprétés comme coréférents, ce que nous schématisons de la façon suivante :

Soulignons bien que si les événements Ea et Ec ne sont pas interprétés comme coréférents, le discours (CR2) n'a pas une sémantique de relation causale mais, par exemple, une sémantique de succession dans le temps de deux événements indépendants.

Nous allons défendre dans la Section 2.3 l'hypothèse suivante : la relation de coréférence entre Ea et Ec est d'un type particulier que nous appelons "généralisation". Ce type de coréférence événementielle s'observe dans des discours mettant en jeu une relation de "paraphrase généralisante" comme :

Luc a violé Marie. Il a donc commis un crime.

La généralisation impose des contraintes sur les phrases décrivant les événements coréférentiels, e.g. interdiction d'apporter une information nouvelle dans la seconde phrase d'une relation de paraphrase généralisante :

¶ Luc a violé Marie. Il a donc commis un crime à minuit.

Nous montrerons que les contraintes qui pèsent sur les discours (CR1) et qui ne s'expliquent pas par l'hypothèse (H1) (i.e. *pas de notion de but*), e.g. interdiction d'ajouter un complément de date dans la seconde phrase :

¶ Luc a cogné la carafe contre l'évier. Il l'a cassée à minuit.

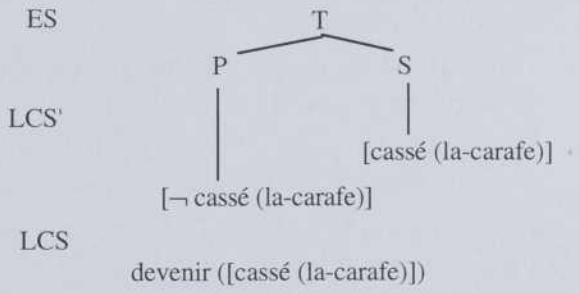
ne sont rien d'autres que des contraintes de généralisation.

Avant de présenter la généralisation, examinons les discours (CR2) dont la seconde phrase est un achèvement :

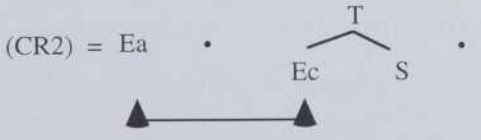
(CR2) H Va (passé composé) ... X .... . X (se) Vc (passé composé).
=: Luc a cogné la carafe contre l'évier. Elle s'est cassée.

Pustejovsky donne la même structure événementielle aux transitions que celles-ci soient des accomplissements ou des achèvements. L'analyse sémantique de *La carafe s'est cassée* est présentée ci-dessous :

(II)



La comparaison entre (I) et (II) montre que les accomplissements et les achèvements ne diffèrent qu'aux niveaux LCS' et LCS. La position de Pustejovsky rejoint celle de Dowty (1979) où les achèvements et accomplissements ne diffèrent que par des questions d'agentivité. Elle permet d'expliquer les propriétés aspectuelles communes à ces deux types d'événements. Elle nous permet aussi de proposer une analyse sémantique des discours (CR2) identique à celle de (CR1) soit :



où Ea est l'événement décrit dans la première phrase, T la transition décrite dans la seconde avec une relation de coréférence entre Ea et Ec. Nous montrerons dans la Section 2.3 que la relation de coréférence entre Ea et Ec est du type généralisation dans (CR2) comme dans (CR1) et que les contraintes qui pèsent sur (CR2), e.g. interdiction d'insérer un complément de date dans la seconde phrase

¶ Luc a cogné la carafe contre l'évier. Elle s'est cassée à minuit.

ne sont autres que des contraintes de généralisation.

2.2 Relation de généralisation

Considérons les paires suivantes dont les éléments ne diffèrent que par l'ordre des phrases juxtaposées (et les différences de pronominalisation qui en découlent) et par la présence de *donc* dans les discours (a) :

(1) a Luc a violé Marie. Il a donc commis un crime.
b Luc a commis un crime. Il a violé Marie.

(2) a Luc a englouti un whisky. Il a donc bu de l'alcool.
b Luc a bu de l'alcool. Il a englouti un whisky.

(3) a Luc a dit bonjour à Marie. Il lui a donc adressé la parole.
b Luc a adressé la parole à Marie. Il lui a dit bonjour.

Les discours (a) peuvent paraître plus naturels si on les complète par d'autres informations :

Luc a violé Marie. Il a donc commis un crime qui sera jugé aux assises.
Luc a englouti un whisky. Il a donc bu de l'alcool alors qu'il est sous antibiotiques.
Luc a dit bonjour à Marie. Il lui a donc adressé la parole après avoir juré de ne plus le faire.

Néanmoins, nous étudierons les discours (a) tels quels, car ce qui nous intéresse c'est la relation de coréférence événementielle : dans ces discours de forme canonique[22] *P1. Donc P2 ...* .

[22] Dans la forme canonique, nous plaçons *donc* en tête de phrase, même

les phrases P1 et P2 réfèrent au même événement E. Nous introduirions donc du bruit en ajoutant des éléments qui n'ont rien à voir avec cette relation de coréférence, i.e. des éléments qui ne concernent pas la description de E.

Les discours (b) de forme *P1. P2.* sont éventuellement ambigus, ayant soit l'interprétation où P1 et P2 réfèrent au même événement, soit l'interprétation où ces phrases réfèrent à deux événements distincts. L'ambiguïté est levée si l'on remplace le point séparant les deux phrases par un deux-points. Néanmoins, ne voulant pas introduire de bruit par l'intermédiaire d'une variante sur le signe de ponctuation, nous maintenons le signe point et nous ne retenons que l'interprétation des discours (b) mettant en jeu un seul événement.

Dans (a) et (b), les deux phrases réfèrent donc au même événement, mais il n'y a aucun élément qui marque explicitement cette coréférence (en mettant de côté *donc* dans (a) qui apporte une éventuelle marque de coréférence). On contrastera les discours (4a) et (4c) d'une part, (4b) et (4d) d'autre part :

(4a) Luc a commis un viol. Il a donc commis **un** crime.
(4c) Luc a commis un viol. Il a commis **ce** crime (*E + à minuit).

(4b) Luc a commis un crime. Il a commis **un** viol.
(4d) * Luc a commis un crime. Il a commis **ce** viol (E + à minuit).

Le discours (4c) met en jeu une anaphore démonstrative : la coréférence est explicitement marquée par *ce*. L'inacceptabilité de (4d) relève d'un phénomène connu (Milner 1982) : une anaphore démonstrative ne peut pas anaphoriser un hyperonyme (*crime* est un hyperonyme de *viol*). Précisons que la seconde phrase de (4c) doit obligatoirement apporter une information nouvelle par rapport à la première phrase (e.g. un complément de date). Nous verrons, par contre, qu'un tel ajout d'information est interdit dans (4a).

Dans les exemples (1a)-(4a) et (1b)-(4b), le fait que les deux phrases réfèrent au même événement repose sur une relation d'hyponymie entre les prédicats et sur une relation d'anaphore

---

si ce connecteur apparaît dans le groupe verbal de P2.

ou de coréférence par hyponymie entre les arguments. Précisément :

- dans un discours (a), le prédicat de la première phrase est en relation d'hyponymie (notée <) avec celui de la seconde, tandis que c'est la situation inverse dans (b) :

  - violer (x, y) < commettre un crime (x, y)[23]
  - engloutir (x, y) < boire (x, y)
  - dire bonjour (x, y) < adresser la parole (x, y)

- dans un discours (a), un argument de la première phrase est soit en relation d'anaphore (notée ≡) ou de coréférence par hyponymie avec l'argument de la seconde phrase ayant le même rôle thématique, soit n'a pas de correspondant dans la seconde phrase ; dans un discours (b), c'est la situation inverse :

  - dans (1a), *Il* anaphorise *Luc*, et *Marie* n'a pas de correspondant dans la seconde phrase ;
  - dans (2a), *Il* anaphorise *Luc*, et *un whisky* < *de l'alcool* ;
  - dans (3a), *Il* anaphorise *Luc*, et *lui* anaphorise *Marie*.

Précisons qu'une non coréférence entre deux arguments ayant le même rôle thématique exclut que les deux phrases réfèrent au même événement (les discours suivants sont précédés du signe ¶ puisqu'ils n'ont pas la sémantique voulue, i.e. celle où les deux phrases décrivent le même événement) :

¶ Luc a violé Marie. Max a donc commis un crime.
¶ Max a commis un crime. Luc a violé Marie.

Les données sur les discours (a) *P1. Donc P2* sont résumées dans le Tableau 1 ci-dessous qui se lit de la façon suivante : la seconde colonne indique le type de constituant que l'on peut observer étant donné le type de constituant indiqué dans la première colonne.

[23] Le second argument (facultatif) de *commettre un crime* est introduit par la préposition *contre* : *Luc a commis un crime contre l'humanité*.

	P1	P2
Prédicat	pred1	pred2 , avec pred1 < pred2
Argument avec un rôle thématique donné	arg1	ø ou arg2 , avec arg1 ≡ arg2 ou arg2 , avec arg1 < arg2

Tableau 1

Avant d'approfondir ces données, donnons les définitions suivantes :

Relation de généralisation : relation de coréférence événementielle d'un type particulier, à savoir : un événement Ei généralise un événement Ej, noté Ei = Généra (Ej), si et seulement si Ei et Ej réfèrent au même événement E et si les informations concernant Ej sont plus "spécifiques" que celles concernant Ei.

Relation de "paraphrase généralisante" : relation qui lie les deux phrases des discours (a) et plus généralement relation qui lie les deux phrases P1 et P2 d'un discours *P1. Donc P2 ...* . tel que P1 et P2 réfèrent au même événement E, sans marque explicite de coréférence autre que celles que nous venons de préciser.

Relation de "paraphrase particularisante" : relation qui lie les deux phrases des discours (b) et plus généralement relation qui lie les deux phrases P1 et P2 d'un discours *P1. P2* . tel que P1 et P2 réfèrent au même événement E, sans marque explicite de coréférence autre que celles que nous venons de préciser.

A priori, les relations de paraphrases généralisante et particularisante mettent toutes deux en jeu une coréférence événementielle de type généralisation si on adopte (ce que nous faisons) les postulats suivants :

- un hyponyme est plus spécifique qu'un hyperonyme,
- un argument plein est plus spécifique qu'un argument vide,
- deux arguments en relation anaphorique sont aussi spécifiques l'un que l'autre.

En effet, on peut vérifier sur le Tableau 1 que les éléments de la première colonne sont plus spécifiques que ceux de la seconde, quels que soient ceux-ci.

En fait, nous allons montrer qu'une paraphrase généralisante met effectivement en jeu une généralisation, mais que ceci n'est pas le cas pour une paraphrase particularisante[24].

2.2.1 Paraphrase généralisante

On ne peut observer une relation de paraphrase généralisante que si la seconde phrase n'apporte **aucune** information nouvelle sur l'événement E par rapport à la première phrase. Commençons par examiner les modifieurs en prenant le cas d'un complément de date : il peut apparaître dans la première phrase des discours (a) mais pas dans la seconde :

(5a) Luc a violé Marie à minuit. Il a donc commis un crime.[25]

(6a) ¶ Luc a violé Marie. Il a donc commis un crime à minuit.

Cependant, si un complément de date apparaît dans la première phrase (e.g. *à minuit*), un hyperonyme de ce complément (e.g. *pendant la nuit*) peut apparaître dans la seconde :

(7a) Luc a violé Marie à minuit. Il a donc commis un crime pendant la nuit. [à minuit < pendant la nuit]

Un tel hyperonyme n'apporte pas d'information nouvelle. Les faits sont les mêmes pour un complément de lieu[26] :

Luc a violé Marie chez Paul. Il a donc commis un crime.

¶ Luc a violé Marie. Il a donc commis un crime chez Paul .

---

[24] Je remercie Anne Guyon d'avoir attiré mon attention sur la dissymétrie entre paraphrases globalisante et particularisante.

[25] Ce discours peut éventuellement être perçu comme ambigu avec une seconde interprétation analogue à celle de :

Luc a joué du saxo à minuit. Il a donc fait du tapage nocturne.

où la relation d'hyponymie entre les deux prédicats accompagnés de modifieurs est : *jouer du saxo à minuit (x) < faire du tapage nocturne (x)*.

[26] Si la première phrase comporte un complément de lieu, on peut éventuellement observer une anaphore de ce complément dans la seconde phrase :

? Luc a violé Marie chez Paul. Il y a donc commis un crime.

Les discours obtenus étant douteux, nous les laisserons de côté.

Luc a violé Marie chez Paul. Il a donc commis un crime chez un particulier. [chez Paul < chez un particulier]

On observe donc une situation radicalement opposée à celle d'un discours comme (4c) qui comporte une anaphore démonstrative et où la seconde phrase doit obligatoirement apporter une information nouvelle par rapport à la seconde :

(4c) Luc a commis un viol. Il a commis **ce** crime (*E + à minuit + chez Paul).

Un tel discours ne met donc pas en jeu une relation de généralisation : le prédicat de la première phrase est plus spécifique que celui de la seconde, mais cette dernière doit apporter une information nouvelle sur E.

Examinons les modifieurs orientés vers l'agent. Apparemment, un adverbe de verbe peut figurer dans la seconde phrase d'une relation de paraphrase généralisante, comme l'atteste l'acceptabilité du discours suivant :

(8a) Luc a dit bonjour à Marie. Il lui a donc adressé la parole (poliment + d'une manière polie).[27]

Mais en fait l'acceptabilité de ce discours repose sur la relation d'hyponymie : *dire bonjour (x, y)* < *adresser la parole poliment (x, y)*. En effet, contrastons (8a) avec (9a) où *impoliment* remplace *poliment* :

(9a) ¶ Luc a dit bonjour à Marie. Il lui a donc adressé la parole (impoliment + d'une manière impolie).

Ce discours est inacceptable dans la mesure où on ne considère pas que *dire bonjour (x, y)* est en relation d'hyponymie avec *adresser la parole impoliment (x, y)*. D'ailleurs, si pour une raison quelconque, on considère que *dire bonjour (x, y)* est en relation d'hyponymie avec *adresser la parole impoliment (x, y)*, le discours (9a) devient acceptable et (8a) inacceptable. On comparera aussi les paires suivantes :

---

[27] Nous indiquons la paraphrase *d'une manière polie* pour ne retenir que l'interprétation de *poliment* comme adverbe de verbe. Les adverbes de phrase seront examinés ultérieurement.

¶ Luc a bu un whisky. Il a donc bu de l'alcool (imprudemment + de manière imprudente).
Luc a bu un whisky à jeun. Il a donc bu de l'alcool (imprudemment + de manière imprudente).

¶ Luc a pris sa voiture pour aller faire une course. Il a donc conduit (imprudemment + de manière imprudente).
Luc a roulé à 160 km/h pour aller faire une course. Il a donc conduit (imprudemment + de manière imprudente).

Dans ces paires, seul le second exemple est acceptable et présente une relation d'hyponymie entre les prédicats éventuellement accompagnés de modifieurs : par exemple, *prendre sa voiture* n'est pas un hyponyme de *conduire imprudemment* mais *rouler à 160km/h* l'est. On retiendra donc qu'un adverbe de verbe est interdit dans la seconde phrase d'une relation de paraphrase généralisante s'il apporte une information nouvelle : un adverbe de verbe n'est autorisé que dans la mesure où il forme avec le prédicat verbal un prédicat complexe qui est un hyperonyme du prédicat verbal de la première phrase éventuellement complété par des modifieurs.

Tournons-nous vers les adverbes de phrase. Ceux-ci peuvent figurer sans problème dans la seconde phrase des discours (a) :

Luc a dit bonjour à Marie. Poliment, il lui a donc adressé la parole.
Luc a bu un whisky. Imprudemment, il a donc bu de l'alcool.
Luc a pris sa voiture pour aller faire une course. Imprudemment, il a donc conduit.

Ces adverbes de phrase n'apportent aucune information nouvelle sur l'événement E : ils n'indiquent qu'un jugement du locuteur par rapport au comportement de l'agent, e.g. *Luc a été poli d'adresser la parole à Marie*.

En conclusion, un modifieur n'est autorisé dans la seconde phrase des discours (a) si et seulement si il n'apporte aucune information nouvelle sur l'événement E.

Après cet examen des modifieurs, revenons aux arguments. Rappelons que, pour un rôle thématique donné, un argument peut apparaître dans la première phrase de (a) et pas dans la seconde, c'est le cas de *Marie* dans (1a) et de *Luc* dans (10a) :

(1a) Luc a violé Marie. Il a donc commis un crime.
(10a) Marie a été violée par Luc. Elle a donc été victime d'un crime.[28]

Par contre, la situation inverse ne permet pas d'obtenir une paraphrase généralisante, comme le montrent les discours suivants (dans les deux premiers, *Luc* n'apparaît que dans la seconde phrase, dans les deux derniers, c'est *Marie* qui n'apparaît que dans la seconde phrase) :

(11a) ¶Marie a été victime d'un viol. Luc a donc commis un crime contre elle.
(12a) ¶Marie a été victime d'un viol. Luc a donc commis un crime.
(13a) ¶Luc a commis un viol. Il a donc commis un crime contre Marie.
(14a) ¶ Luc a commis un viol. Marie a donc été victime d'un crime.

Ces discours (douteux) ne peuvent pas recevoir une interprétation telle que les deux phrases réfèrent au même événement. Or leur seconde phrase apporte une information nouvelle par rapport à la première phrase. En revanche, si un argument n'apparaît ni dans la première phrase ni dans la seconde, comme l'agent et le patient respectivement dans (15a) et (16a) :

(15a) Marie a été victime d'un viol. Elle a donc été victime d'un crime.
(16a) Luc a commis un viol. Il a donc commis un crime.

on a bien une relation de paraphrase généralisante dans laquelle il est sous-entendu que l'agent ou le patient (non spécifié) est le même dans les deux phrases.

Les données sur les paraphrases généralisantes sont récapitulées dans le tableau ci-dessous[29] qui complète le Tableau 1 et qui se lit comme ce dernier (i.e. la seconde colonne indique le type de constituant que l'on peut observer étant donné le type de constituant indiqué dans la première colonne).

---

[28] Cet exemple met en jeu la relation d'hyponymie *être violé (x) = être victime d'un viol (x) < être victime d'un crime (x)*, bien que l'équivalence *être violé (x) = être victime d'un viol (x)* sera remise en question lors de l'étude des paraphrases particularisantes.

[29] La relation d'hyponymie "étendue" que nous avons utilisée, e.g. *dire bonjour (x,y) < adresser la parole poliment (x, y)*, n'est pas représentée dans ce tableau : elle demande une formalisation qui sort largement du cadre de cette étude.

Paraphrase généralisante	P1	P2
Prédicat	pred1	pred2 , avec pred1 < pred2
Argument avec un rôle thématique donné	arg1 --- ø	ø ou arg2 , avec arg2 ≡ arg1 ou arg2 , avec arg1 < arg2 --- ø
Modifieur de verbe d'un type donné	modif1 --- ø	ø ou modif2, avec modif1 < modif2 --- ø

Tableau 2

Au total, on peut affirmer que la relation de paraphrase généralisante met en jeu un phénomène de généralisation : l'événement E2 décrit dans la seconde phrase généralise l'événement E1 décrit dans la première. En effet, E1 et E2 réfèrent au même événement et les informations concernant E1 sont plus spécifiques que celles concernant E2.

2.2.2 Paraphrase particularisante

On n'observe pas de contrainte aussi stricte pour une paraphrase particularisante. Ainsi, un complément de date peut apparaître dans la première ou dans la seconde phrase des relations de paraphrase particularisante :

(6b) Luc a commis un crime à minuit. Il a violé Marie.

(5b) Luc a commis un crime. Il a violé Marie à minuit.

De même, pour un complément de lieu ou un modifieur orienté vers l'agent :

Luc a commis un crime (chez Paul + par esprit de vengeance). Il a violé Marie.

Luc a commis un crime. Il a violé Marie (chez Paul + par esprit de vengeance).

Il semble que l'adjonction d'un adverbial dans la première phrase des discours (b) soit autorisée dès qu'elle l'est hors contexte. Ajoutons que si la première phrase comporte un

modifieur d'un type donné, la seconde peut comporter un hyponyme de ce modifieur :

(7b) Luc a commis un crime pendant la nuit. Il a violé Marie à minuit. Luc a commis un crime chez un particulier. Il a violé Marie chez Paul.

Passons aux arguments. Les discours suivants :

(1b) Luc a commis un crime. Il a violé Marie.
(10b) Marie a été victime d'un crime. Elle a été violée par Luc.
(15b) Marie a été victime d'un crime. Elle a été victime d'un viol.
(16b) Luc a commis un crime. Il a commis un viol.

sont tout aussi acceptables que les discours (a) équivalents. Examinons les discours (b) équivalents aux discours inacceptables (11a)-(14a) :

(11b) ? Luc a commis un crime contre Marie. Elle a été victime d'un viol.
(12b) ¶ Luc a commis un crime. Marie a été victime d'un viol.
(13b) ? Luc a commis un crime contre Marie. Il a commis un viol.
(14b) ¶ Marie a été victime d'un crime. Luc a commis un viol.

Les discours (12b) et (14b) excluent l'interprétation telle que les deux phrases réfèrent au même événement. Par contre, les discours (11b) et (13b) peuvent éventuellement être interprétés comme une relation de paraphrase particularisante. On peut noter la différence suivante : (11b) et (13b) présentent une relation anaphorique pour un argument (*Marie* dans (11b) et *Luc* dans (13b)), tandis que (12b) et (14b) ne présentent aucune relation anaphorique entre arguments, chacune des phrases ne mentionnant qu'un argument (soit *Luc* soit *Marie*). Cette différence pourrait être reliée aux différences d'interprétation entre (12b) et (14b) d'une part et (11b) et (13b) d'autre part. La situation est en fait plus complexe. Contrastons (12b) avec (12'b) obtenu en substituant le prédicat *être violé (x)* à *être victime d'un viol (x)* :

(12b) ¶ Luc a commis un crime. Marie a été victime d'un viol.
(12'b) ? Luc a commis un crime. Marie a été violée.

À l'inverse de (12b), le discours (12'b) peut éventuellement être interprété comme une relation de paraphrase particularisante. Cette différence entre (12b) et (12'b) montre que les prédicats *être victime d'un viol (x)* et *être violé (x)* ne peuvent pas être

considérés comme équivalents, en tout cas pour une paraphrase particularisante. Pour une paraphrase généralisante, ils sont équivalents au niveau qui nous intéresse dans la mesure où (11'a) est aussi inacceptable que (11a) et (12'a) aussi inacceptable que (12a) :

(11a) ¶ Marie a été victime d'un viol. Luc a donc commis un crime contre elle.

(11'a) ¶ Marie a été violée. Luc a donc commis un crime contre elle.

(12a) ¶ Marie a été victime d'un viol. Luc a donc commis un crime.

(12'a) ¶Marie a été violée. Luc a donc commis un crime.

Quoi qu'il en soit, on retiendra qu'il existe des contraintes sur la distribution des arguments dans les deux phrases d'une relation particularisante, mais que celles-ci ne sont pas immédiates et demandent d'approfondir les questions d'hyponymie entre prédicats. Les données sur les paraphrases particularisantes sont résumées dans le Tableau 3 ci-dessous.

Paraphrase particularisante	**P1**	**P2**
Prédicat	pred1	pred2 , avec pred2 < pred1
Argument avec un rôle thématique donné	arg1	ø ? ou arg2 , avec arg2 ≡ arg1 ou arg2 , avec arg2 < arg1
	ø	ø ou arg2
Modifieur de verbe d'un type donné	modif1	ø ou modif2, avec modif2 < modif1
	ø	ø ou modif2

Tableau 3

La comparaison entre les Tableaux 2 et 3 montre clairement que les relations de paraphrase généralisante et particularisante ne sont pas symétriques. En particulier, une paraphrase particularisante ne met pas en jeu un phénomène de généralisation comme en témoigne le discours (6b) : dans ce discours, le prédicat de la première phrase est moins spécifique que celui de la seconde, mais la première phrase apporte une information nouvelle (i.e. un complément de date) par rapport à la seconde.

Une paraphrase particularisante met en jeu un autre type de coréférence événementielle que l'on peut appeler "particularisation" et dont la définition complète demande de formaliser la distribution des arguments et la notion d'hyponymie entre prédicats. La dissymétrie entre généralisation et particularisation se traduit par les formules suivantes :

Ej = Généra (Ei) => Ei = Parti (Ej)
Ei = Parti (Ej) ≠> Ej = Généra (Ei)

En effet, si un discours (a) exprime une paraphrase généralisante, alors le discours (b) obtenu en inversant l'ordre des phrases exprime une paraphrase particularisante ; par contre, l'inverse n'est pas vraie, voir (6b) acceptable et (6a) inacceptable.

2.2.3 Conclusion

Nous avons mis en évidence deux phénomènes linguistiques non encore décrits à notre connaissance et que nous avons appelé "généralisation" et "particularisation". Ces phénomènes de coréférence événementielle s'observent dans la langue respectivement dans les "relations de paraphrase généralisante" et dans les "relations de paraphrase particularisante". Dans la section suivante, nous allons faire appel à ces phénomènes pour expliquer le comportement des discours (CRi) et (RCi).

Auparavant, une remarque. La relation de généralisation se formalise aisément dans un système de représentation des connaissances reposant sur des structures de traits typés avec hiérarchie des types, comme celle de Aït-Kaci (1986, 1989) dont nous nous inspirons pour l'entrée d'un système de génération (Danlos 1994). Les structures de traits typés sont appelées des Ψ–termes chez Aït-Kaci. Sans entrer dans les détails techniques, l'idée est la suivante :

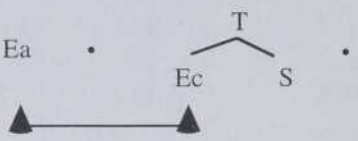
Ei = Généra (Ej) si et seulement si les représentation de Ei et Ej sont deux Ψ–termes ti et tj tels que tj est subsumé par ti.

2.3 Relation causale directe et généralisation

En premier lieu, nous allons défendre l'hypothèse suivante :

(H2) *Un discours de structure (CR1) ou (CR2) avec l'analyse sémantique*

(CR1) ou (CR2) =



n'a une interprétation de relation causale directe que lorsque Ec = Génera (Ea).

Commençons par (CR1) :

(CR1) H Va (passé composé) ... X .... . H Vc (passé composé) X
=: Luc a cogné la carafe contre l'évier. Il l'a cassée.
[Transitive, Accomplissement]

Cette hypothèse est vraie pour un discours (CR1) sans aucun modifieur dans la seconde phrase. En effet, l'interprétation de relation causale directe implique que les événements Ea et Ec soient coréférents et les informations concernant Ea sont plus spécifiques que celles concernant Ec car les seules information sur Ec sont les suivantes :

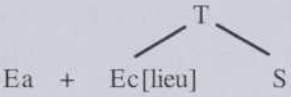
- c'est une action non spécifiée, alors que cette action est spécifiée dans Ea,
- c'est une action effectuée par l'agent H, comme dans Ea.

On a donc Ec = Génera (Ea) et une sémantique de relation causale directe, conformément à l'hypothèse (H2). Examinons maintenant si cette hypothèse est valide lorsque la seconde phrase de (CR1) autorise ou interdit un modifieur d'un type donné. Pour cela, nous allons reprendre un à un les faits décrits dans la Section 1 :

(i) *Interdiction d'ajouter un complément de lieu ou de date dans la seconde phrase :*

(1a) ¶ Luc a cogné la carafe contre l'évier. Il l'a cassée (chez Paul + à minuit).

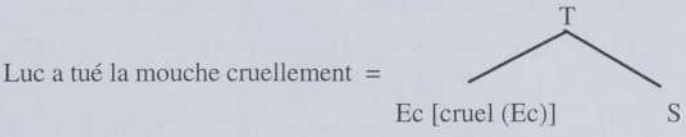
Un tel complément de lieu ou de date porte sur Ec avec une analyse sémantique schématisée de la façon suivante pour un complément de lieu :



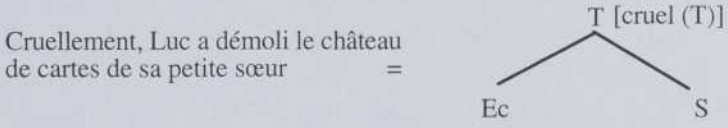
Ec comporte donc une information non présente dans Ea. De ce fait, il n'y a pas de relation de généralisation entre Ea et Ec. Or il n'y a pas de sémantique de relation causale directe.

(ii) *Interdiction d'ajouter un modifieur orienté vers le sujet*

Dans Pustejovsky (1991), la différence entre les adverbes de verbe et les adverbes de phrase se traduit par une différence de portée dans les structures événementielles. Un adverbe de verbe porte sur Ec :



tandis qu'un adverbe de phrase porte sur la transition :



L'interdiction d'ajouter un adverbe de verbe dans la seconde phrase de (CR1) :

(1e$_4$) ¶ Luc a brûlé les pattes de cette mouche. Il l'a tuée cruellement.

reçoit donc la même explication que celle présentée pour les compléments de lieu et de date : Ec comporte une information non présente dans Ea, il n'y a pas de généralisation entre Ea et Ec et pas de sémantique de relation causale directe. Par contre, on ne peut pas expliquer ainsi l'interdiction d'ajouter des adverbes de phrase :

(1e$_3$) ¶ Luc a soufflé sur le château de cartes de sa petite sœur. Cruellement, il l'a démoli.

puisque ceux-ci portent sur T et non sur Ec. En revanche, on peut avoir recours à une observation de Molinier (1990), à savoir que les adverbes de phrase impliquent que l'agent a accompli une action volontaire : la phrase

Luc a démoli le château de cartes de sa petite sœur.

est ambiguë selon que Luc a agi volontairement ou pas, mais la phrase

Cruellement, Luc a démoli le château de cartes de sa petite sœur.

n'est pas ambiguë : elle signifie que Luc a volontairement démoli le château (et que ceci était cruel de sa part). Par conséquent, le discours ($1e_3$) indique que l'agent H voulait le résultat. Or dans la Section 1.4, nous avons mis en avant l'hypothèse (H1) :

(H1) *Un discours de structure (CR1) ne peut être acceptable avec une sémantique de relation causale directe que si rien n'indique que l'agent H voulait le résultat (qu'il l'ait voulu ou pas).*

L'inacceptabilité de ($1e_3$) s'explique donc par le fait que ce discours est interdit par l'hypothèse (H1).

Nous venons d'avancer deux types d'explication, l'une pour l'interdiction des adverbes de verbe - hypothèse (H2) -, l'autre pour l'interdiction des adverbes de phrase - hypothèse (H1). L'interdiction d'insérer d'autres modifieurs orientés vers l'agent relève de l'une ou l'autre explication[30]. Ainsi, l'inacceptabilité de :

¶ Luc a cogné la carafe contre l'évier. Il l'a cassée en rêvant à sa fiancée.

relève de l'hypothèse (H2) : le gérondif apporte une information sur ce que faisait l'agent pendant Ec, information non présente dans Ea, d'où non généralisation entre Ea et Ec. En revanche, l'inacceptabilité de :

¶ Luc a cogné la carafe contre l'évier. Il l'a cassée pour ennuyer Marie.

---

[30] On peut envisager de relier ces deux hypothèses en posant le postulat suivant : indiquer explicitement dans la seconde phrase de (CR1) que l'agent voulait le résultat implique que l'on apporte une information sur Ec, information évidemment non présente dans Ea ; ce qui revient à Ec ≠ Généra (Ea).

s'explique par le fait que le complément de but dans la seconde phrase implique que Luc voulait le résultat, ce qui est interdit par l'hypothèse (H1).

(iii) *Possibilité d'insérer un adverbe "évaluatif"* :

(1g) Luc a cogné la carafe contre l'évier. Malheureusement, il l'a cassée.

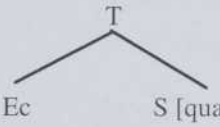
Un tel adverbial, qui indique un jugement du locuteur, ne porte pas sur Ec et n'introduit pas de notion de but. Il est donc autorisé sans problème, conformément aux hypothèses (H1) et (H2).

(iv) *Possibilité d'insérer un adverbial quantifieur* :

(1i) Luc a cogné la carafe contre l'évier. Il l'a cassée en mille morceaux.

Un quantifieur porte sur l'état S et non sur Ec :

Luc a cassé la carafe en mille morceaux =



On a donc Ec = Généra (Ea) et une sémantique de relation causale directe.

(v) *Possibilité d'insérer un complément de durée* :

($1j_1$) Les Japonais ont envoyé des bombes sur le porte-avions. Ils l'ont coulé en 1 heure.

D'après Pustejovsky (1991), un complément de durée dans une phrase isolée comme *Les Japonais ont coulé le porte-avions en 1 heure* exprime l'intervalle de temps entre le début de Ec et le changement d'état. Dans (CR1), ce complément de durée exprime le temps écoulé entre Ea = Ec et le changement d'état. Il ne concerne pas à proprement parler Ec. On a donc Ec = Généra (Ea) et une sémantique de relation causale directe.

(vi) *Possibilité d'insérer un connecteur "consécutif" ou "reformulatif"* :

($1m_1$) Luc a cogné la carafe contre l'évier. Ce faisant, il l'a cassée.

Un tel connecteur ne porte pas sur Ec et n'introduit pas de notion de but. Il est donc autorisé sans problème, conformément aux hypothèses (H1) et (H2).

(vii) *Impossibilité d'exprimer un décalage entre la fin de l'action et le changement d'état :*

(1r) ¶ Les Japonais ont envoyé des bombes sur le porte-avions. Ils l'ont coulé 1 heure après.

Les événements Ea et Ec ne peuvent pas être coréférentiels dans (1r) puisqu'ils sont décalés d'une heure dans le temps. Il n'y a pas de généralisation entre Ea et Ec et pas de sémantique de relation causale directe.

(viii) *Impossibilité de mentionner l'agent H dans la seconde phrase s'il n'est pas mentionné dans la première*

Pour des raisons évidentes, nous mentionnons ici l'exemple (5x) de structure (CR5) présentée dans la Section 1.6) :

(1x) ¶ La carafe a été cognée contre l'évier. Luc l'a cassée.
(5x) ¶ La carafe a été cognée contre l'évier. Elle a été cassée par cet imbécile.

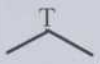
L'événement Ec indique qui est l'agent contrairement à Ea. On n'a donc pas de généralisation entre Ea et Ec et pas de sémantique de relation causale directe. Soulignons que l'hypothèse (H2) (i.e. *Ec = Généra (Ea)*) permet de se passer de la règle (R3) (i.e. *pas d'agent dans la cause ==> pas d'agent dans le résultat*) qui présentait un caractère *ad hoc* (voir Section 1. 5).

Nous sommes donc arrivée à la conclusion suivante : l'hypothèse (H2) sur la relation de généralisation ainsi que l'hypothèse (H1) sur l'absence de notion de but permettent de prédire toutes les propriétés de (CR1) (et de (CR4) et (CR5)).

Tournons-nous vers les discours (CR2) :

(CR2) H Va (passé composé) ... X .... . X (se) Vc (passé composé).
=: Luc a cogné la carafe contre l'évier. Elle s'est cassée.
[Neutre, Achèvement]

Si l'on suit Pustejovsky (1991) qui donne la même structure événementielle aux achèvements qu'aux accomplissements soit



,on peut expliquer certaines des propriétés des discours (CR2) comme nous l'avons fait pour les discours (CR1), i.e. par l'hypothèse (H2). Ainsi, l'interdiction d'ajouter un complément de lieu dans la seconde phrase de (CR2) :

(2a) ¶ Luc a cogné la carafe contre l'évier. Elle s'est cassée chez Paul.

relève d'une violation des contraintes de généralisation comme pour (CR1). Les différences entre (CR1) et (CR2) pertinentes pour notre propos sont de deux ordres :

(i) *Possibilité d'exprimer un décalage dans le temps entre la fin de l'action et le changement d'état :*

(2c) Hier, les Japonais ont envoyé une bombe sur le porte-avions. Il a coulé ce matin.

(2n) Les Japonais ont envoyé une bombe sur le porte-avions. Il a coulé 1 heure après.

($2j_2$) Les Japonais ont envoyé une bombe sur le porte-avions. Il a coulé en 1 heure.

Rappelons que les mêmes exemples dans un discours de structure (CR1) n'induisent pas une sémantique de relation causale directe :

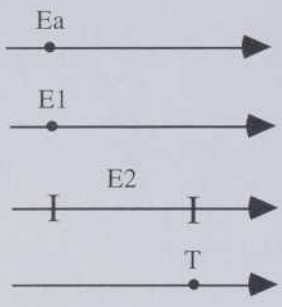
(1c) ¶ Hier, les Japonais ont envoyé une bombe sur le porte-avions. Ils l'ont coulé ce matin.

(1n) ¶ Les Japonais ont envoyé une bombe sur le porte-avions. Ils l'ont coulé 1 heure après.

($1j_2$) ¶ Les Japonais ont envoyé une bombe sur le porte-avions. Ils l'ont coulé en 1 heure.

Un discours comme ($2j_2$) abrège une chaîne causale à plus de deux éléments dont les extrémités sont les deux phrases de ($2j_2$), e.g. :

Ea Les Japonais ont envoyé une bombe sur le porte-avions
=> E1 La bombe a fait une brèche dans le porte-avions
=> E2 Le porte-avions a pris l'eau (pendant 1 heure)
=> T Le porte-avions a coulé



Sur l'axe des temps, Ea et E1 sont ponctuels et quasi simultanés (avec Ea < E1), E2 succède immédiatement à E1 et dure 1 heure, jusqu'au moment de la transition T.

Dans la décomposition de T en sous-événements, il semble logique de poser que Ec = E2. Il n'y a alors pas de relation de généralisation entre Ea et Ec puisque ces événements sont non coréférentiels : Ea est ponctuel, et Ec = E2 est duratif.

La question qui se pose est la suivante : est-ce qu'un discours comme ($2j_2$) a une sémantique de relation causale directe ? Nous pensons que non : une relation causale directe ne peut abréger une chaîne causale à plus de deux éléments. Les discours (2c), (2n) ou ($2j_2$) avec un décalage temporel entre le cause et le résultat ont plutôt une sémantique de relation causale indirecte (Danlos 1985). On notera d'ailleurs que ces discours peuvent être paraphrasés par des constructions factitives comme :

> Le fait que les Japonais ont envoyé une bombe sur le porte-avions hier a fait qu'il a coulé ce matin.

et qu'il est connu que ces constructions factitives peuvent décrire des relations causales non directes, comme dans :

> Le fait que la foudre est tombée hier sur un groupe de fillettes a fait que deux pompiers sont morts ce matin.

On rappellera aussi le contraste mis en avant entre autres par Fodor (1970), McCawley (1978) et repris dans Jackendoff (1990) :

> *? Bill killed Mary on Tuesday by giving him poison on Monday.
> Bill caused Mary to die on Tuesday by giving him poison on Monday.

Ce contraste souligne le caractère direct des relations causales exprimées par un verbe causatif *(kill)*, contrairement au caractère moins direct des relations causales exprimées par une construction causale périphrastique *(cause to die)* qui permet d'exprimer un décalage temporel entre la cause et le résultat. Il

va de pair avec le contraste observé entre les discours (CR1) et (CR2) présentant un décalage entre la cause et le résultat. Au total, dans les discours (CR2) avec un décalage entre la fin de l'action et le changement d'état, on n'a pas de généralisation entre Ea et Ec et on n'a pas une sémantique de relation causale directe, mais une sémantique de relation causale indirecte. Autrement dit, ces discours n'infirment pas l'hypothèse (H2). Il reste à expliquer pourquoi les discours (CR2) peuvent exprimer une relation causale directe ou indirecte, tandis que les discours (CR1) ne peuvent exprimer qu'une relation causale directe, tout en donnant la même structure événementielle aux accomplissements et aux achèvements.

(ii) *Impossibilité d'insérer un adverbe orienté vers le sujet dans la seconde phrase de (CR2) prise isolément* (ce qui n'est pas le cas pour la seconde phrase de (CR1)) :

* La carafe s'est cassée (désinvoltement + de manière désinvolte + avec désinvolture + pour attirer l'attention sur lui + en rêvant à sa fiancée + par esprit de vengeance).

On n'a donc pas besoin de faire appel aux hypothèses (H1) et (H2) pour expliquer l'inacceptabilité des discours suivants :

(2e) * Luc a cogné la carafe contre l'évier. Elle s'est cassée (désinvoltement + de manière désinvolte + avec désinvolture + pour attirer l'attention sur lui + en rêvant à sa fiancée + par esprit de vengeance)

La règle (R1) (i.e. *phrase isolée inacceptable ==> discours inacceptable*) de portée très générale suffit.

Notons que les règles (R1) et (R2) suffisent pour expliquer le comportement de (CR3) qui est régulier dans la mesure où un modifieur est accepté dans la seconde phrase si et seulement si il est accepté dans cette phrase prise isolément.

Ayant examiné toutes les contraintes qui pèsent sur les discours (CR1) et (CR2), nous pouvons maintenant énoncer la condition nécessaire et suffisante qui suit :

Un discours de structure (CR1) ou (CR2) est acceptable (dans un contexte gauche nul) avec une sémantique de relation causale directe si et seulement si :

- *il ne contient que des phrases acceptables isolément (et telles que la première phrase puisse être sur le plan pragmatique la cause de la seconde),*
- *il n'induit aucune notion de but [hypothèse (H1) pour (CR1) et* de facto *pour (CR2)*[31]*],*
- *Ec = Généra (Ea) [hypothèse (H2)].*

Deux remarques :

A) La relation de généralisation dans une relation causale directe met en jeu plus de contraintes que dans une relation de paraphrase généralisante. En effet, si la première phrase de (CR1) ou (CR2) comporte un complément de date ou de lieu, tout hyperonyme de ce complément est interdit dans la seconde phrase :

¶ Luc a cogné la carafe contre l'évier à minuit. Il l'a cassée pendant la nuit.

¶ Luc a cogné la carafe contre l'évier chez Paul. Il l'a cassée chez un particulier.

De plus, les éléments H et X de la seconde phrase de (CR1) ou (CR2) peuvent être des pronoms ou des groupes nominaux anaphorisant les éléments correspondants de la première phrase (voir Section 1.5), mais non des hyperonymes indéfinis :

Luc a cogné une carafe contre l'évier. (Il + Cet imbécile) a cassé **cet** objet de valeur.

¶ Luc a cogné une carafe contre l'évier. (Il + Cet imbécile) a cassé **un** objet de valeur.

Un parisien a cogné la carafe contre l'évier. **Ce** français l'a cassée.

¶ Un parisien a cogné la carafe contre l'évier. **Un** français l'a cassée.

[31] Rappelons qu'un discours (CR2) ne peut induire aucune notion de but puisque la seconde phrase décrit un achèvement.

Luc a cogné une carafe contre l'évier. **Ce** beau récipient (s'est + est) cassé.

¶ Luc a cogné une carafe contre l'évier. **Un** beau récipient (s'est + est) cassé.

Les données sur les relations causales directes de structure (CRi) sont récapitulées dans le Tableau 4 ci dessous.

Relation causale directe de structure (CRi)	**P1**	**P2**
Argument avec un rôle thématique donné	arg1	ø ou arg2 , avec arg2 ≡ arg1
	ø	ø
Modifieur de verbe d'un type donné	modif1	ø
	ø	ø

Tableau 4

La comparaison entre les Tableaux 2 et 4 montre clairement que les relations causales directes imposent des contraintes plus strictes que les relations de paraphrase généralisante. Cette différence pourrait s'expliquer par le fait que la coréférence événementielle est implicite pour les relations causales directes, alors qu'elle est explicite, lexicale et soulignée par *donc* pour les relations de paraphrase généralisante.

B) Il semble que les discours RÉSULTAT < CAUSE obtenus en inversant les phrases

(1z) Luc a cassé la carafe. Il l'a cognée contre l'évier.

(2z) ? La carafe s'est cassée. Luc l'a cognée contre l'évier.

(3z) La carafe est cassée. Luc l'a cognée contre l'évier.

(4z) La carafe a été cassée. Luc l'a cognée contre l'évier.

(5z) La carafe a été cassée par Luc. Il l'a cognée contre l'évier.

mettent en jeu une coréférence événementielle analogue à celle observée dans les relations de paraphrase particularisante (voir Section 2.2). Pour montrer cela, commençons par noter que

l'enchaînement de deux phrases avec des agents qui sont non coréférents :

¶ Max a cassé la carafe. Luc l'a cognée contre l'évier.

exclut toute interprétation de relation causale directe : il n'y a pas de coréférence événementielle, comme il n'y a pas de coréférence événementielle dans les discours suivants où les agents sont non coréférents :

¶ Max a commis un crime. Luc a violé Marie.

Passons à l'adjonction d'adverbiaux. Rappelons (voir Section 1.6) que les discours RÉSULTAT < CAUSE ont un comportement régulier dans la mesure où l'insertion d'adverbiaux dans la première phrase est acceptable si et seulement si elle est acceptable hors contexte :

(1zz) Luc a cassé la carafe (volontairement + chez Paul). Il l'a cognée contre l'évier.

(2zz) ? La carafe s'est cassée (*volontairement + chez Paul). Luc l'a cognée contre l'évier.

(3zz) La carafe est cassée (*volontairement + *chez Paul). Luc l'a cognée contre l'évier.

(4zz) La carafe a été cassée (volontairement + chez Paul). Luc l'a cognée contre l'évier.

(5zz) La carafe a été cassée par Luc (volontairement + chez Paul). Il l'a cognée contre l'évier.

Or, ce comportement régulier s'observe aussi dans une relation de paraphrase particularisante (voir Tableau 3). Ajoutons que les deux phrases des discours RÉSULTAT < CAUSE peuvent comporter des adverbiaux en relation d'hyperonymie (si la première phrase prise isolément admet l'adjonction d'un adverbial) :

Luc a cassé la carafe pendant la nuit. Il l'a cognée contre l'évier à minuit.

Ceci est aussi le cas pour les discours exprimant une relation de paraphrase particularisante:

Luc a commis un crime pendant la nuit. Il a violé Marie à minuit.

Enfin, examinons la distribution des arguments dans les deux phrases. Dans les discours (1z)-(5z) où la CAUSE est à

l'actif, les arguments H et X sont mentionnés dans la seconde phrase, X est mentionné dans la première, et selon les exemples, H est mentionné ou non dans la première. Au niveau argumental, la seconde phrase est donc "plus spécifique" que la première, ce qui est cohérent avec une relation de paraphrase particularisante. Examinons maintenant les discours où la CAUSE est au passif sans agent, elle ne mentionne donc pas H :

(1yz) ?¶ Luc a cassé la carafe. Elle a été cognée contre l'évier.
(2yz) ? La carafe s'est cassée. Elle a été cognée contre l'évier.
(3yz) La carafe est cassée. Elle a été cognée contre l'évier.
(4yz) La carafe a été cassée. Elle a été cognée contre l'évier.
(5yz) ?¶ La carafe a été cassée par Luc. Elle a été cognée contre l'évier.

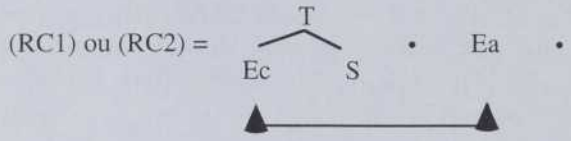
Les discours (2yz), (3yz) et (4yz), dont la première phrase ne mentionne pas H, ont une sémantique de relation causale directe, mais les discours (1yz) et (5yz), dont la première phrase mentionne H, reçoivent difficilement une interprétation de relation causale directe. Ce contraste entre (2yz), (3yz) et (4yz) d'une part, et (1yz) et (5yz) d'autre part est analogue au contraste suivant :

(15b) Marie a été victime d'un crime. Elle a été victime d'un viol.
(11b)? Luc a commis un crime contre Marie. Elle a été victime d'un viol.

Dans (15b), ni la première phrase ni la seconde ne comportent d'agent mais il est sous-entendu que c'est le même dans les deux phrases et ce discours s'interprète comme une relation de paraphrase particularisante. Situation analogue pour (2yz), (3yz) et (4yz). En revanche, la première phrase de (11b) spécifie un agent contrairement à la seconde et ce discours reçoit difficilement une interprétation de paraphrase particularisante. Situation analogue pour (1yz) et (5z).

En résumé, il semble que les discours parataxiques RÉSULTAT < CAUSE mettent en jeu une relation de coréférence événementielle qui est celle observée dans une relation de paraphrase particularisante. On pourrait donc avancer l'hypothèse (H3) :

(H3) *Un discours de structure (RC1) ou (RC2) avec l'analyse sémantique*



n'a une interprétation de relation causale directe que lorsque Ea = Parti (Ec).

Il reste que la relation de coréférence dite de particularisation demande à être formalisée au niveau de la distribution des arguments dans les deux phrases et de la relation d'hyponymie entre prédicats (voir Section 2.2).

3 Structures de discours non parataxiques

Jusqu'à présent, nous n'avons examiné que des discours parataxiques. Nous allons examiner brièvement des discours non parataxiques que l'on peut mettre en parallèle avec les discours (CRi). Par souci de simplification, nous laisserons désormais de côté la structure (CR5) où le verbe causatif est au passif avec agent. Néanmoins, cette structure s'intègre sans aucun problème dans les paradigmes que nous allons présenter.

3.1 Discours avec une relative

Parallèlement à chaque discours (CR2), (CR3) et (CR4), on peut construire un discours avec un pronom relatif sujet, si tant est que les conditions syntaxiques soient réunies, ce qui implique pour nos exemples de supprimer la séquence *contre l'évier* :

(2) α Luc a cogné la carafe. Elle s'est cassée.
β Luc a cogné la carafe, qui s'est cassée.

(3) α Luc a cogné la carafe. Elle est cassée.
β ? Luc a cogné la carafe, qui est cassée.

(4) α Luc a cogné la carafe. Elle a été cassée.
β ? Luc a cogné la carafe, qui a été cassée.

Le discours (3β) est douteux ; son acceptabilité est améliorée si l'on insère dans la relative un adverbe comme *maintenant* : *Luc a cogné la carafe, qui est cassée maintenant*. Le discours (4β) est aussi douteux et éventuellement ambigu : interprétation de relation causale directe et interprétation où la carafe avait été cassée avant que Luc ne la cogne. Cette ambiguïté disparaît pour d'autres exemples : *Luc a tiré sur Marie, qui a été tuée.* Quoi qu'il en soit, ces discours à relative ont exactement les mêmes propriétés que les discours (CR2), (CR3) et (CR4) comme en témoignent les faits suivants :

- impossibilité d'insérer dans la relative un adverbial relatif au comportement agentif de Luc ou un complément circonstanciel de lieu ou de date :

¶ Luc a cogné la carafe, qui s'est cassée (volontairement + chez Paul + à minuit + etc.).

¶ Luc a cogné la carafe, qui est cassée (volontairement + chez Paul + à minuit + etc.).

¶ Luc a cogné la carafe, qui a été cassée (volontairement + chez Paul + à minuit + etc.).

- impossibilité de construire la relative avec *assassiner* pour le cas (4β), ce qui montre qu'elle n'exprime pas un but :

¶ Luc a tiré sur Marie, qui a été assassinée.

On peut donc poser que les discours (β) avec une relative, dans la mesure où on les considère comme acceptables, sont de simples variantes des discours parataxiques (α) : le passage d'un point suivi d'un (pronom) sujet à un pronom relatif sujet ne serait donc qu'une simple modification de surface dont les effets seraient purement stylistiques.

Indiquons que parallèlement aux discours (CR1) on peut construire un discours avec un pronom relatif objet :

(1) α Luc a cogné la carafe. Il l'a cassée.
β ? Luc a cogné la carafe qu'il a cassée.

Le discours (1β) ne nous paraît pas très heureux. Quoiqu'il en soit, le lecteur vérifiera que (1β) présente les mêmes propriétés que (1α). Enfin, la situation est analogue pour la paire :

La carafe a été cognée contre l'évier par Luc. Il l'a cassée.
La carafe a été cognée contre l'évier par Luc qui l'a cassée.

On peut donc avancer l'hypothèse que dans les discours CAUSE < RÉSULTAT un pronom relatif peut éventuellement lier les deux phrases en n'induisant qu'un changement stylistique par rapport à la parataxe.

3.2 Discours avec un participe présent

Parallèlement à un discours (CR1), on peut construire un discours avec un participe présent :

(1) γ Luc a cogné la carafe contre l'évier. Il l'a cassée.
δ Luc a cogné la carafe contre l'évier, la cassant.

Les discours (1γ) et (1δ) ont exactement les mêmes propriétés comme en témoignent les faits suivants :

- impossibilité d'insérer dans le participe présent un modifieur orienté vers l'agent ou un complément circonstanciel de lieu ou de date :

¶ Luc a cogné la carafe contre l'évier, la cassant (volontairement + chez Paul + à minuit + etc.).

- impossibilité de construire le participe présent avec *assassiner* ou *se suicider*, ce qui montre qu'il n'exprime pas un but :

¶ Luc a poussé Marie par la fenêtre, l'assassinant.
¶ Luc s'est jeté par la fenêtre, se suicidant.

On peut donc poser que les discours (δ) sont de simples variantes des discours (γ) : le passage d'un point suivi d'un (pronom) sujet à un participe présent avec effacement du sujet ne serait donc qu'une simple modification de surface dont les effets seraient purement stylistiques.

3.3 Discours avec un connecteur

Comme nous l'avons signalé dans la Section 1.3, il est possible d'insérer un connecteur dans un discours parataxique afin de préciser le lien entre les deux phrases :

Luc a cogné la carafe contre l'évier. (De ce fait + par là-même) (il l'a cassée + elle s'est cassée + elle est cassée + elle a été cassée).

Nous avons étudié en détail dans Danlos (1988) les différents connecteurs que l'on peut insérer dans un discours exprimant une relation causale directe. Rappelons simplement ici qu'un connecteur donné n'est pas forcément adéquat pour les quatre structures de discours (CR1), (CR2), (CR3) et (CR4). Ainsi, rappelons que *ce faisant* n'est insérable que dans (CR1) pour des raisons syntaxiques :

Luc a cogné la carafe contre l'évier. Ce faisant, il l'a cassée.
* Luc a cogné la carafe contre l'évier. Ce faisant, elle (s'est cassée + est cassée + a été cassée).

De même, la conjonction *et* est parfaite dans le discours (CR2), douteuse dans le discours (CR1) qui devient ambigu (relation causale ou succession de deux événements), et malheureuse dans les discours (CR3) et (CR4) :

Luc a cogné la carafe contre l'évier et elle s'est cassée.
?¶ Luc a cogné la carafe contre l'évier et il l'a cassée.
? Luc a cogné la carafe contre l'évier et elle (est cassée + a été cassée).

Quoi qu'il en soit, l'insertion d'un connecteur ne change en rien le comportement des discours. Par exemple, un discours avec le verbe causatif *assassiner* est tout autant inacceptable avec ou sans *ce faisant* dans la structure (CR1) :

¶ Luc a poussé Marie par la fenêtre. Il l'a assassinée.
¶ Luc a poussé Marie par la fenêtre. Ce faisant, il l'a assassinée.

On comparera aussi les inacceptabilités suivantes :

¶ Luc a cogné la carafe contre l'évier. Elle s'est cassée (chez Paul + à minuit).
¶ Luc a cogné la carafe contre l'évier et elle s'est cassée (chez Paul + à minuit).

On peut donc poser que les discours avec un connecteur sont des variantes (stylistiques) des discours à parataxe.

3.4 Bilan

Les discours respectant l'ordre CAUSE < RÉSULTAT non parataxiques (avec une relative, un participe présent ou un connecteur) peuvent être considérés comme des variantes

stylistiques des discours parataxiques. Cette conclusion rejoint celle avancée dans la Section 1.6 et que nous rappelons : le comportement d'un discours (CRi) dépend de la valeur aspectuelle de sa seconde phrase, mais il ne dépend pas de la construction syntaxique choisie pour exprimer cette valeur aspectuelle. Nous pouvons maintenant avancer la conclusion plus générale qui suit : le comportement d'un discours exprimant une relation causale directe en respectant l'ordre CAUSE < RÉSULTAT dépend de la valeur aspectuelle de sa seconde phrase, mais il ne dépend ni de la construction syntaxique choisie pour exprimer cette valeur aspectuelle, ni des éventuels mots de liaison entre la phrase exprimant la cause et celle exprimant le résultat. Ajoutons qu'un tel discours n'induit jamais une notion de but. Autrement dit, les hypothèses que nous avons mises en avant pour expliquer le comportement des discours parataxiques de structure (CR1) et (CR2) s'appliquent aussi aux discours non parataxiques de structure parallèle à (CR1) ou (CR2).

Conclusion

L'explication que nous avons proposée pour les faits empiriques aberrants concernant les discours (CR1) et (CR2) repose sur une relation coréférentielle dite de généralisation et sur des arguments de sémantique lexicale, i.e. sémantique de *se suicider* et *assassiner*, structure événementielle de Pustejovsky (1991). Il est donc clair que l'analyse de discours peut profiter pleinement d'analyse de sémantique lexicale. Pour conclure, nous souhaitons montrer que la sémantique lexicale peut à son tour profiter avantageusement d'études fines et méthodiques du discours.

A) D'abord un simple argument de réciprocité : notre analyse de phénomènes discursifs, qui repose sur la structure événementielle de Pustejovsky et qui permet d'expliquer de façon élégante des comportements aberrants, conforte l'hypothèse de la structure événementielle qui a été établie sur l'étude de phrases isolées.

B) Il est connu que de nombreux verbes causatifs sont ambigus par rapport à l'intentionnalité de l'agent, cette ambiguïté pouvant être levée grâce à un adverbe comme *volontairement* :

(v1) Luc a cassé la carafe volontairement
(v2) Luc a tué Marie volontairement
(v3) Luc s'est tué volontairement

Mais que veulent dire exactement ces phrases ? On peut y voir (Dowty 1991) les deux notions suivantes :

- Luc *[*a cassé la carafe / a tué Marie / s'est tué*]* en agissant volontairement
- Luc voulait *[* que la carafe soit cassée / la mort de Marie / sa propre mort*]*

La question qui se pose est la suivante : peut-on lier ces deux notions par des implications comme :

(i1) si Luc a cassé la carafe en agissant volontairement, alors il voulait qu'elle soit cassée
(i2) si Luc a tué Marie en agissant volontairement, alors il voulait sa mort
(i3) si Luc s'est tué en agissant volontairement, alors il voulait sa propre mort

Les contraintes lexicales exposées dans la Section 1.4

(1o) Beregovoy s'est tiré une balle dans la tête. Il s'est tué.
(1p) ¶ Beregovoy s'est tiré une balle dans la tête. Il s'est suicidé.

(1q) Luc a poussé (volontairement) Marie par la fenêtre. Il l'a tuée.
(1r) ¶ Luc a poussé (volontairement) Marie par la fenêtre. Il l'a assassinée.

prouvent que ces implications sont fausses. En effet, si l'implication (i3) était vraie, alors les discours (1o) et (1p) véhiculeraient le même ensemble d'informations, à savoir : Beregovoy s'est tué en agissant volontairement (information venant de la première phrase) et il voulait sa propre mort (information venant de l'implication (i3) pour (1o) et de la sémantique de *se suicider* pour (1p)). Par conséquent, rien ne permettrait d'expliquer que (1o) est acceptable mais que (1p) ne l'est pas. Donc l'implication (i3) est fausse. Même raisonnement pour l'implication (i2) au vu du contraste entre (1q) et (1r).

Rappelons que des paires comme *se suicider / se tuer* ou *assassiner / tuer* sont assez rares, nous ne pouvons donc pas faire le même raisonnement pour *casser*. Néanmoins, ayant démontré que (i3) et (i2) sont fausses, nous nous permettons d'affirmer qu'il en est de même pour (i1) et pour toute implication de ce type. Autrement dit, des phénomènes discursifs montrent de façon rigoureuse que des phrases comme (v1)-(v3) sont ambiguës par rapport au but de l'agent : on sait qu'il a agi volontairement mais on ne sait *a priori* pas s'il voulait ou non le résultat.

C) Comme nous venons de le rappeler, le verbe *se suicider* implique que le résultat était voulu par l'agent contrairement à *se tuer* qui reste ambigu à ce niveau là. Milner (communication personnelle) poursuit ce raisonnement en notant que *se tuer* est à son tour moins ambigu que *mourir* : *se tuer*, c'est mourir par l'effet d'une décision volontaire tandis que *mourir* n'implique pas forcément une décision volontaire. Il appuie cette affirmation sur des contrastes comme :

* Trois millions de Français se tuent chaque année (par + avec) le cancer
Trois millions de Français meurent chaque année du cancer

Trois millions de Français se tuent chaque année par overdose
Trois millions de Français meurent chaque année d'overdose

Ces exemples opposent la mort suite à une maladie contractée par l'agent, ce qui ne relève pas de sa décision (seul *mourir* est alors acceptable) et la mort suite à un acte "à risque" effectué sciemment par l'agent (*se tuer* et *mourir* sont alors tous deux acceptables). La remarque de Milner peut en fait être élargie à d'autres formes pronominales réfléchies grâce à l'étude du discours :

(t1) Luc s'est jeté par la fenêtre. Il s'est blessé.
(t2) Luc est tombé par la fenêtre. Il s'est blessé.
(t3) ¶ Luc a été poussé dans une crevasse. Il s'est blessé.
(t4) ¶ Luc a reçu un pot de bégonia sur la tête. Il s'est blessé.

Les exemples (t1) et (t2) sont acceptables et Luc est responsable de ce qui lui arrive dans la première phrase, que celle-ci décrive un acte volontaire de sa part comme en (t1) ou non comme en

(t2). Par contre, Luc n'est en aucune manière responsable de ce qui lui arrive en (t3) et (t4) et ces discours sont inacceptables. Le contraste entre (t1)-(t2) d'une part et (t3)-(t4) d'autre part montre que la notion de responsabilité intervient dans l'acceptabilité des formes pronominales réfléchies et nous proposons dans Danlos (à paraître) de formaliser cette notion des responsabilité grâce aux proto-rôles thématiques de Dowty (1991). Cette notion ainsi que celle d'intentionnalité discutée dans la remarque B) sont délicates à cerner dans des phrases isolées sur lesquelles le jugement du locuteur peut varier selon le contexte qu'il imagine. Elles semblent par contre stables dans des discours comme ceux que nous étudions.

D) Les verbes psychologiques : rappelons que nous avons d'emblée écarté les relations causales dont le résultat est un changement d'état psychique ou moral pour nous concentrer sur celles dont le résultat est un changement d'état physique. Une des premières raisons est qu'il est impossible de faire des distinctions subtiles comme "relation causale directe" versus "relation causale indirecte" pour des changements d'état psychique puisque disons rapidement que n'importe quoi peut provoquer n'importe quel changement psychique chez n'importe qui :

Luc a quitté Marie. Il a choqué Zoé.
Luc n'a pas mis de cravate. Il a choqué Zoé.

D'autre part, les verbes causatifs psychologiques ont des propriétés que n'ont pas les verbes causatifs "physiques". Par exemple *choquer* dans sa construction transitive admet un sujet non humain contrairement à *casser* :

Cet accident a choqué Marie
* Cet accident a cassé la carafe

Il aurait donc fallu étudier, en plus des structures (CRi), des structures comme :

Luc a quitté Marie. Ceci a choqué Zoé.

Ceci aurait donc surchargé une étude déjà assez complexe.

Les relations causales dont le résultat est un changement psychique font donc partie de notre future recherche et nous avons l'intime conviction que les phénomènes discursifs

permettront de mettre en lumière des propriétés des verbes psychologiques qui, pour l'instant, n'ont été étudiés que dans des phrases isolées (littérature abondante).

E) Enfin, rappelons les contrastes observés avec les constructions neutres selon l'ordre des phrases :

Luc a cogné la carafe contre l'évier. Elle s'est cassée.
? La carafe s'est cassée. Luc l'a cognée contre l'évier.

Marie a lancé de l'eau sur le feu. Il s'est éteint.
? Le feu s'est éteint. Marie a lancé de l'eau dessus.

Ces contrastes pourraient confirmer ou infirmer les hypothèses faites sur la neutralité.

Au total, les quelques remarques que nous venons de faire tendent à montrer que le discours peut apporter un cadre rigoureux pour mettre en avant des phénomènes de sémantique lexicale. A l'inverse, le corps de cet article a essayé de montrer que l'étude du discours (et donc du TALN) ne saurait se passer d'études de sémantique lexicale.

Bibliographie

Aït-Kaci, H., 1986, LOGIN : A Logic Programming Language with Built-in Inheritance, *The Journal of Logic Programming*, Vol 3, n° 3.

Aït-Kaci, H., Lincoln, P., 1989, LIFE : A Naturel Language for Natural Language, *TA Informations, vol 30, N° 1-2.*

Boons, J.P., Guillet, A., Leclère, Ch, 1976, *La structure des phrases simples en français : constructions intransitives*, Droz, Genève.

Danlos, L., 1985, *Génération automatique de textes en langues naturelles*, Masson, Paris.

Danlos, L., 1987, *The Linguistic Basis of Text Generation,* Cambridge University Press, Cambridge.

Danlos, L., 1988, Connecteurs et relations causales, *Langue Française*, n° 77, Larousse, Paris.

Danlos, L., 1994, *G-TAG : Un formalisme pour la génération de textes inspiré des grammaires d'arbres adjoints (TAG),* Rapport interne TALANA, Université Paris 7.

Danlos, L., 1995a, Direct Causal Relations: Linguistic Data, Interlingual Representation and Multilingual Generation, *Proceedings of IJCAI95 Workshop on Multilingual Generation*, Montréal.

Danlos, L., 1995b, *Unacceptable* Discourses: How to avoid them in Computer Text Generation and to Use them in Computer Generated Litterature?, *Computer Literature : Hypertext, Interactive Fiction and Literary Theory*, éd. M. Lenoble, J. Murray, & A. Vuilemin éd., MIT Press.

Danlos, L., à paraître, *Formes pronominales réfléchies et rôles thématiques.*

Dowty, D., 1979, *Word meaning and Montague Grammar*, Reidel, Dordrecht.

Dowty, D, 1991, Thematic Proto-Roles and Argument Selection, *Language*, vol 67, n°3.

Fodor, J., 1970, Three reasons for Not Deriving *kill* from *cause to die*, *Linguistic Inquiry* 1, pp. 429-438.

Hovy, E., Maier, E., 1995, Parsimonious or Profligate : How many and Which Discourse Structure Relations?, *Discourse Processes.*

Jackendoff, R., 1983, *Semantics and Cognition,* MIT Press, Cambridge, MA.

Jackendoff, R, 1990, *Semantic Structures,* MIT Press, Cambridge, MA.

Knott, A., Dale, R., 1994, Using Lingusitic Phenomena to motivate a set of Rhethorical Relations, *Discourse Processes*, 18-1.

McCawley, J., 1968, Lexical Insertion in a Transformational Grammar without Deep Structure, B. Barden, C-J. Bailey and A. Davidson éd., *Papers from the fourth Regional Meeting of the Chicago Linguistic Society,* Chicago.

McCawley, J., 1978, Converstional Implicature and the Lexicon, P. Cole éd., *Syntax and Semantics,* vol. 9, pp. 245-259.

Milner, J.C., 1982, *Ordre et raisons de langue*, Le Seuil, Paris.

Molinier, C., 1990, Une classification des adverbes en *-ment*, *Langue Française,* n° 88, Larousse.

Pak, M., à paraître, *Les relations causales directes en français et en coréen.*

Pustejowsky, J., 1991, The syntax of event structure, B. Levin and S. Pinker éd., *Lexical and Conceptual Semantics,* Elsevier Science Publishers, Amsterdam.

Ruwet, N., 1972, *Théorie syntaxique et syntaxe du français*, Le Seuil, Paris.

Schank, R., 1975, *Conceptual Information Processing*, North Holland, Amsterdam.

Vendler, Z., 1967, *Linguistics and Philosophy,* Cornell University Press, Ithaca.

van Voorst, J., 1995, The Semantic Structure of Causative Constructions, *Studies in Language*.

L'ACCORD DANS UNE GRAMMAIRE COMPUTATIONNELLE DU FRANÇAIS

Louisette Emirkanian
Lyne Da Sylva
Lorne H. Bouchard

Résumé

Le traitement de l'accord du participe passé en français sert à illustrer le fossé qui peut exister entre théorie linguistique et implantation. Une première formalisation du phénomène de l'accord semble remettre en cause le Principe d'Accord et de Contrôle, l'un des trois principes universels d'instanciation de traits proposés par la théorie de la Grammaire syntagmatique généralisée (GSG). Nous montrons comment une analyse du fonctionnement de cette solution nous a amenés à proposer une deuxième analyse qui exploite les autres mécanismes d'instanciation de traits disponibles en GSG.

1 Introduction

Le but du programme de recherche, dont nous allons ici présenter un aspect, est le développement de la théorie de la Grammaire syntagmatique généralisée (GSG, Gazdar et al. 1985) et son exploitation dans le traitement automatique des

langues naturelles[1]. Ce programme est composé de trois volets interdépendants : un volet développement théorique, un volet développement descriptif et un volet exploitation dans le traitement automatique des langues naturelles, celui du français en particulier.

Cet article d'une part vise à illustrer le fossé qui peut exister entre théorie et implantation, d'autre part il a pour objectif de montrer comment l'implantation nous a amenés à proposer une solution permettant de traiter un problème difficile à résoudre dans le cadre particulier de cette théorie.

Nous parlerons d'abord de la problématique de l'implantation d'un modèle théorique. Puis, nous traiterons de certains phénomènes d'accord en faisant le parallèle entre le traitement en GSG et l'implantation réalisée avec l'outil *Grammar Development Environment* (GDE, Boguraev 1988 et Grover et al. 1989). Nous proposerons ensuite une première analyse de l'accord du participe passé en GSG. Cette analyse sera rejetée à la lumière des données. Nous verrons que les principes universels d'instanciation de traits, le Principe d'Accord et de Contrôle en particulier, semblent inadéquats pour décrire l'accord du participe passé. Nous exploiterons alors les autres mécanismes d'instanciation de traits disponibles en GSG.

2 Problématique

Dans la construction de systèmes automatisés de traitement de la langue, deux approches sont possibles, l'une empirique et l'autre linguistique où l'on s'intéresse à la caractérisation et à la modélisation des propriétés des langues naturelles. Nous avons opté pour la deuxième approche. Cependant, même si le but premier du linguiste est d'expliquer les propriétés des langues naturelles, il est primordial de décrire avant de pouvoir expliquer, surtout si l'on se place du point de vue de la linguistique informatique. Nous ne justifierons pas ici le choix du formalisme. Précisons simplement qu'en GSG, justement, une

[1] Cette recherche est financée par le CRSHC (40-93-0607) et le FCAR (95 ER 1198).

part importante est accordée à l'adéquation descriptive avant même qu'il soit question d'adéquation explicative.

Pour implanter notre grammaire, nous avons utilisé l'outil Grammar Development Environment (GDE), un banc d'essai pour les grammaires des langues naturelles définies dans un formalisme proche de la GSG.

La grammaire computationnelle que nous développons[2] vise l'adéquation descriptive et explicative. On se trouve cependant devant un problème. Une grammaire atteint l'adéquation explicative si elle décrit la structure des phrases par un petit ensemble de règles très contraintes. Or, là, on constate que lorsqu'on implante une grammaire on perd beaucoup de la généralité et de la simplicité du modèle, au niveau de l'expression des contraintes. Cependant, malgré ces inconvénients, la construction d'un modèle opératoire est un outil méthodologique précieux qui nous permet d'obtenir une assurance accrue de la validité du modèle proposé.

2.1 La Grammaire syntagmatique généralisée

La GSG présentée dans Gazdar et al. (1985) est une théorie monostratale qui utilise un seul niveau de représentation, soit celui correspondant aux chaînes de surface. Elle met en facteur deux types d'informations véhiculées simultanément par les règles de réécriture de la grammaire hors contexte usuelle : la Dominance Immédiate (règle de DI) et la Préséance Linéaire (règle de PL). Ces règles, dans lesquelles les catégories sont définies comme des ensembles de traits-valeurs, sont sous-spécifiées. Sur les projections de ces règles s'appliquent les principes qui instancient des traits. Ils sont de deux types. D'un côté, les Restrictions de Cooccurrence de Traits (RCT) et les Spécifications de Traits par Défaut (STD) imposent des

[2] À partir d'une grammaire noyau (Bouchard et al. 1992), nous avons enrichi les différentes catégories syntagmatiques et parallèlement à cela, nous nous sommes penchés sur des problèmes particuliers, tels que l'ambiguïté liée au rattachement des syntagmes prépositionnels, les coordinations de non constituants, la sous-catégorisation et la cliticisation, la spécification dans le syntagme nominal, etc.

contraintes sur la bonne formation des catégories, c'est-à-dire qu'elles permettent de dire si les catégories dans un sous-arbre local sont une projection légale d'une règle de DI. La RCT en (1) précise que la présence du trait VFORME (forme du verbe) implique celle des traits V +, N -, c'est-à-dire la présence d'un verbe.

(1) [VFORME] ⊃ [V +, N -]

La spécification de traits par défaut en (2) précise qu'à moins qu'il en soit indiqué autrement une catégorie n'est pas marquée pour le trait CONJ.

(2) ¬ [CONJ]

De l'autre côté, les principes universels d'instanciation de traits, le Principe de Traits de Tête (PTT), le Principe de Traits de Pied (PTP) et le Principe d'Accord et de Contrôle (PAC) permettent d'effectuer la propagation des traits entre les différentes catégories d'un même arbre local. Par exemple, le PTT dit que, dans tout arbre local, les traits de tête de la mère sont identiques aux traits de tête de la fille tête. Il s'agit des traits de tête libres.

2.2 L'outil Grammar Development Environment

Les principes de la GSG se manifestent en GDE sous une autre forme. Les RCT sont absentes et une partie de leur fonctionnalité est réalisée par des déclarations de catégories qui imposent de dire quels traits peuvent apparaître sur une catégorie donnée. La déclaration de catégorie en (3) dit que tous les éléments nominaux doivent avoir l'ensemble Traits_tête_Nominale que l'on trouve dans les ensembles de déclarations de traits comme en (4).

(3) LCATEGORY NL => Traits_tête_Nominale.

(4) SET Traits_tête_Nominale = {FEM, PLU, EGO, PTC, NP, PRO, COMPT, NOM, ACC, DAT, LOC, OBL}

Les STD de la GSG (GKPS 1985) sont absentes bien qu'un certain type de spécification par défaut soit possible. Par ailleurs, avec l'outil GDE, on doit programmer explicitement les principes universels d'instanciation en termes de règles de

propagation de traits. Celles-ci gèrent la propagation de traits spécifiques à l'intérieur de règles spécifiques (ou du moins, respectant un patron défini). C'est probablement à ce niveau que l'on perd le plus des aspects généralisation et explication.

3 Théorie et implantation : exemples d'accord

Nous ferons ici systématiquement le parallèle entre ce que nous avons implanté en GDE et le traitement GSG. Nous établirons donc un lien entre la théorie et l'implantation. Nous allons d'abord examiner un exemple très simple : celui de l'accord sujet-verbe. Cela nous permettra d'introduire le Principe d'Accord et de Contrôle.

3.1 L'accord sujet-verbe

En GSG

Le PAC rend compte du partage des traits d'accord, au niveau de l'accord sujet-verbe mais aussi à celui de l'accord entre un élément extrait et la position de laquelle il est extrait, dans le cas des topicalisées, des relatives, des interrogatives, par exemple. Le PAC s'appuie sur les types sémantiques associés aux catégories et ce sont ces types sémantiques qui vont permettre de déterminer si l'on est en présence d'un foncteur (catégorie contrôlée) ou d'un argument (catégorie contrôleur), le foncteur s'accordant avec son argument (Keenan 1974). Le PAC tel que défini dans Gazdar et al. (1985) comprend trois aspects : le schéma de contrôle, la définition du trait de contrôle et le PAC lui-même.

Examinons d'abord le premier aspect, soit le schéma de contrôle. Voici la définition que nous en donnent Gazdar et al. (1985, p.88) et que nous reproduisons en (5).

(5) Si ϕ est une projection de r, où $r = C_0 \rightarrow C_1, \ldots, C_n$, alors une catégorie $\phi(C_i)$ contrôle $\phi(C_j)$ en ϕ, $1 \leq i, j \leq n$, si et seulement si

(i) $TYP(\chi(\phi(C_j))) = \langle TYP(\chi(\phi(C_i))), TYP(\chi(\phi(C_0)))\rangle$, ou

(ii) $TYP(\chi(\phi(C_j))) = TYP(SV)$ et l'un des types associés à la tête de r est $\langle TYP(SV), \langle TYP(\chi(\phi(C_i))), TYP(SV)\rangle\rangle$

La première partie du schéma de contrôle dit que, dans une règle de DI, le type sémantique du contrôlé est tel qu'il cherche un constituant du type sémantique du contrôleur pour former la mère de la règle. La deuxième partie de la définition porte sur le contrôle des infinitives par les compléments pour des verbes tels que *persuader*[3]. Considérons la règle de DI en (6)[4].

(6) $P \rightarrow N2, V2$

La catégorie N2 contrôle la catégorie V2 puisque le type sémantique de V2 est <TYP (N2), TYP (P)>, c'est-à-dire que le type sémantique d'un V2 est tel qu'il cherche un constituant du type sémantique de celui du contrôleur, le N2 sujet (sa sœur), pour former une phrase (sa mère). L'instanciation de certains traits (le trait SLASH, par exemple) pouvant modifier le type sémantique d'une catégorie, les traits soumis au PAC sont ceux donnés par les spécifications χ des catégories (7) (Gazdar et al. 1985, p.88), c'est-à-dire les traits de tête, moins ceux qui sont aussi traits de pied, augmentés des traits de pied hérités, soient ceux qui sont dans les règles de DI.

(7) $\chi(\phi(C_i)) = \phi(C_i) \mid ((\text{TÊTE} - \text{PIED}) \cup (\text{DOM}(C_i \mid \text{PIED})))$

Considérons ensuite le trait de contrôle dont voici la définition en (8) (Gazdar et al. 1985, p.89).

(8) Soient C_i une catégorie dans une règle r, avec $C_i(\text{BARRE}) \neq 0$ et ϕ une projection de r, alors un trait t est le trait de contrôle de $\phi(C_i)$ si et seulement si

(i) $t = \text{SLASH}$ et $t \in \text{DOM}(C_i)$, ou

(ii) $\text{SLASH} \notin \text{DOM}(C_i)$ et $t = \text{AGR}$

Le trait de contrôle, soit celui qui est pertinent pour l'accord, est AGReement, trait à valeur catégorielle, à moins que la règle ne comporte le trait SLASH (nous reviendrons sur ce trait plus loin). Ainsi, dans la projection de la règle (6), le trait de contrôle sera bien AGR puisqu'aucun SLASH n'est hérité, c'est-à-dire ne se trouve dans la règle.

---

[3] Les travaux de Baschung (1988 et 1991) sur le français portent plus particulièrement sur ce point.

[4] N2 est une abréviation utilisée pour noter la catégorie [N +, V -, BARRE 2] et V2, la catégorie [N -, V +, BARRE 2].

Considérons enfin le PAC lui-même tel que défini par Gazdar et al. (1985, p.89).

(9) Soit Φ_r l'ensemble des projections de r, où $r = C_0 \rightarrow C_1, ..., C_n$, alors $\phi \in \Phi_r$ satisfait le PAC sur r si et seulement si

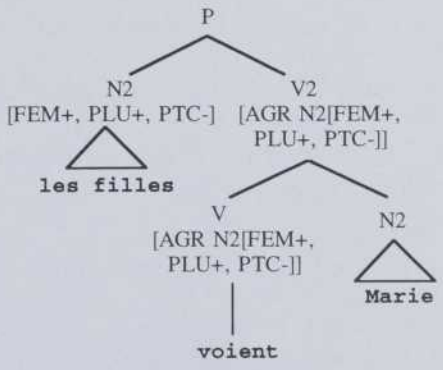
(i) si $\phi(C_j)$ contrôle $\phi(C_i)$, alors $\phi(C_i)(t_i) = \chi(\phi(C_j)) \cup \phi(C_j) \mid \{t_i\}$, où t est le trait de contrôle de $\phi(C_i)$.

(ii) si $\phi(C_i)$ est une catégorie prédicative sans contrôleur, alors $\phi(C_i)(t_i) = \phi(C_0)(t_0)$ où t_i et t_0 sont respectivement les traits de contrôle de $\phi(C_i)$ et $\phi(C_0)$.

La première partie de la définition stipule que le trait de contrôle de la catégorie contrôlée (C_i) devra être unifié avec la catégorie contrôleur (C_j). La deuxième partie de la définition précise que si l'on est en présence d'une catégorie prédicative sans contrôleur, la valeur du trait de contrôle de cette catégorie sera celle du trait de contrôle de sa mère. En (6), le V2, le foncteur, portera le trait de contrôle AGR dont la valeur sera la catégorie avec laquelle il s'accorde (le N2) comme nous le montre l'arbre en (10b)[5] pour la phrase (10a).

(10) a Les filles voient Marie.

b



[5] Nous n'avons noté ici, comme nous le ferons pour les projections suivantes, que les traits pertinents pour la démonstration. Précisons que [PTC -] signale la troisième personne. Les traits [EGO +, PTC+] renvoient à la première personne et [EGO -, PTC +], à la deuxième.

C'est le verbe au lexique qui porte le trait AGR, qui est à la fois un trait de contrôle et un trait de tête. Le mot *voient* est noté au lexique comme indiqué en (11) ; il n'est pas marqué pour le genre.

(11) {N -, V +, BARRE 0, SUBCAT 1, AGR N2[PLU +, PTC-] }

Le trait AGR est propagé par le PTT jusqu'à la mère V2. On remarque que dans l'arbre ci-dessus le trait AGR sur le V2 a pour valeur la catégorie N2[FEM +, PLU +, PTC -]. Le trait [FEM +] provient de l'unification des traits de tête de la catégorie argument et de la valeur du trait de contrôle. C'est donc l'interaction du PTT et du PAC qui rend compte de l'accord sujet-verbe en GSG.

En GDE

L'implantation en GDE est semblable dans la mesure où le contenu de AGR doit s'unifier avec les traits d'accord du sujet. Cependant, d'une part nous devons en GDE programmer explicitement les principes en termes de règles de propagation, d'autre part, pour assurer la propagation d'un trait donné, il doit être présent explicitement dans la règle de DI. De plus, tous les traits doivent être présents sur les catégories. Par exemple, au lexique, *voient* portera également le trait FEM avec une valeur variable @ qui peut être unifiée avec + ou avec –.

La règle de DI de la phrase étant donnée en (12), la règle de propagation (13)[6] assure cet accord.

(12) IDRULE Phrase/suj :
V2+SUJ --> N2, H2[SUJ -, AGR N2].

(13) PROPRULE Accord/Sujet_SV :
V2+SUJ --> N2, H2[SUJ -, AGR N2].
F(1) = F(2[AGR]), F in TraitsAccord.

---

[6] À noter que, dans les règles de propagation, (1) et (2) renvoient aux filles et (0) à la mère. De plus, la notation F(2[AGR]) réfère à la valeur du trait catégoriel AGR de la fille 2.

Précisons que l'ensemble TraitsAccord regroupe les traits suivants : FEM, PLU, EGO, PTC. Comme on peut le constater, on doit noter dans la règle de DI la présence du trait à valeur catégorielle AGR pour qu'on puisse assurer sa propagation. Des règles de propagation supplémentaires telles que celle mentionnée en (14) font percoler le trait AGR dans la structure verbale et l'instancient sur le premier élément verbal, ce qui simule l'effet du PTT de la GSG sur le trait AGR.

(14) PROPRULE Propagation/AGR/V2 :
V2[SUJ -] --> V[H +], U.
F(0[AGR]) = F(1[AGR]) , F in TraitsAccord+Catégoriel.

Comme on peut le constater, jusque là, pour un exemple aussi simple, les différences ne sont pas trop grandes. Nous allons analyser un cas plus complexe : celui de l'accord du participe passé, en particulier dans une dépendance non bornée. Cependant, avant de nous pencher sur l'accord du participe passé proprement dit, nous allons examiner ce qui se passe au haut de la dépendance, c'est-à-dire au niveau de la correspondance entre l'élément extrait et la phrase "trouée" d'une part, et entre l'antécédent et la relative d'autre part. Par exemple, nous regardons, dans (15), le lien entre *fille* et *que Léo a vue*, et entre *que* et *Léo a vue*.

(15) la fille que Léo a vue ...

3.2 L'accord dans une dépendance non bornée

L'accord du participe passé employé avec *avoir* est sensible à l'antéposition de l'objet direct[7] ; il a donc été implanté en conjonction avec le trait SLASH. L'accord au haut de la dépendance peut être décrit en trois étapes : l'instanciation et la propagation du trait SLASH, l'unification du SLASH et du syntagme relatif, enfin l'unification entre les traits de la relative et ceux de l'antécédent.

---

[7] Dans le cadre de la théorie Gouvernement Liage, l'accord du participe passé a été abondamment traité, en particulier pour le français par Kayne (1985), Lefebvre (1986) et Bouchard (1987). Bien sûr, les mécanismes invoqués sont tout à fait différents de ceux de la GSG.

3.2.1 Instanciation et propagation du trait SLASH

En GSG

Considérons d'abord le trait SLASH et la façon dont on traite les dépendances non bornées dans la GSG. À partir d'une règle de DI contenant un certain complément, une métarègle produit une nouvelle règle de DI où ce complément est absent. Dans la GSG, cette métarègle d'extraction des compléments (16) est très générale. Elle permet d'extraire tout X2 (Gazdar et al. 1985, p.143).

(16) MR (Slash Termination Metarule 1 - STM1) :
X --> W, X2
⇓
X --> W, X2[+NUL]

Une RCT (17) va placer le trait SLASH sur une catégorie [+NUL], à condition que celle-ci ne soit pas une catégorie lexicale.

(17) RCT : [+NUL] ⊃ [SLASH]

Un élément lexical, ε, pourra être inséré sous le N2[+NUL]. Il porte les traits α [+NUL, SLASH α].

Le trait SLASH est un trait de tête et un trait de pied. Il est donc soumis au PTT et au PTP[8]. En tant que trait de pied, il percolera jusqu'au niveau P[SLASH X2] et l'unification des traits sera ainsi assurée pour ce trait à valeur catégorielle. Il ne percole plus à ce niveau puisqu'il est introduit dans une règle de DI, donc hérité et non instancié. Cette règle de DI très générale rend compte de tous les cas d'extraction ; elle est notée en (18).

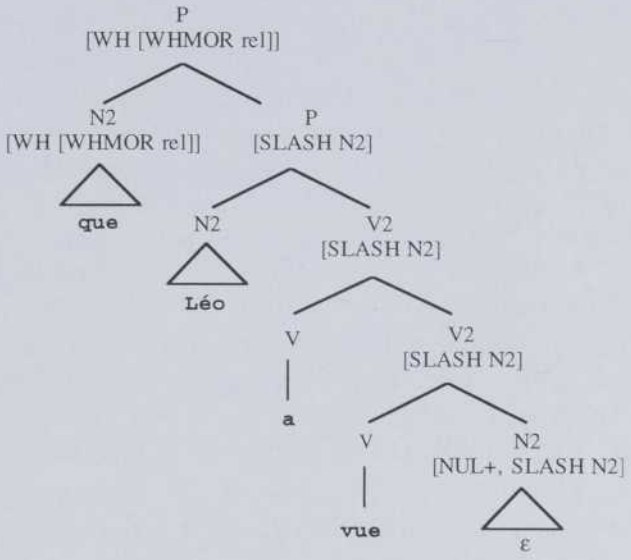
(18) P → X2, P[SLASH X2]

La projection en (19) rend compte de l'instanciation et de la propagation du trait SLASH dans la relative.

---

[8] Voici une version allégée de la définition que donnent Gazdar et al. (1985, p.82) du PTP : les traits de pied instanciés sur la catégorie mère d'un arbre local doivent être identiques à l'unification des traits de pied instanciés sur les filles.

(19)



Alors qu'en GSG le PTP à lui seul permet d'introduire le trait et de veiller à l'unification, en GDE, nous sommes obligés de passer par deux étapes, et ce à chacun des niveaux.

En GDE

En effet, dans la GDE, le PTP n'étant pas implanté, l'introduction des traits de pied, tels que SLASH, doit être spécifiée pour chaque structure et leur percolation doit être faite par des règles de propagation. Pour l'extraction d'un complément, on applique une batterie de métarègles sur les règles de complémentation ce qui produit de nouvelles règles de DI où le complément pertinent est absent ; de plus, la mère de la règle de DI résultante porte le trait SLASH. La métarègle est donnée en (20)[9].

[9] Le tilde (~) utilisé devant un trait spécifie un patron de règle où ce trait est absent. Ainsi, ~SLASH indique que SLASH n'est pas présent sur la catégorie.

(20) METARULE Extraction/slash/N2 :
V2[~SLASH] --> H, N2[~SLASH], U.
==>
V2[SLASH N2] --> H, N2[NUL +,SLASH N2], U.

Dans cette métarègle, la correspondance au niveau du site même de l'extraction, entre la valeur du SLASH et la catégorie N2 est faite de façon explicite N2[SLASH N2]. En GSG, en revanche, cette correspondance est faite par l'entrée lexicale α [+NUL, SLASH α] (voir ci-dessus) qui permet de mettre en correspondance le contenu des deux variables de type libre α.

Les règles d'extraction de compléments ne représentent qu'une composante du mécanisme d'extraposition. Il faut maintenant propager le trait SLASH. Au bas de la structure, il est introduit par la métarègle (20). Il suffit alors d'appliquer la règle de propagation en (21).

(21) PROPRULE Propagation/slash/extraction:
V2[SLASH X2] --> H, X2[SLASH X2], U.
F(0[SLASH]) = F(2[SLASH]), F in TraitsSLASH.

L'ensemble TraitsSLASH rassemble tous les traits pertinents de l'extraction : les traits catégoriels (N, V, BARRE), les traits de cas (NOM, ACC, DAT, LOC, OBL), la forme de la préposition (PFORME) et enfin les traits d'accord en genre, nombre et personne (FEM, PLU, EGO, PTC).

Il faut ensuite, pour simuler l'effet du PTP (c'est-à-dire l'ajout du trait SLASH dans toute la structure), dupliquer toutes les règles de DI où le SLASH peut être instancié. Il faut d'une part introduire le SLASH dans les règles et d'autre part faire en sorte qu'il soit soumis aux règles de propagation.

La métarègle (22) insère le SLASH dans la structure du syntagme verbal.

(22) METARULE Introduction/slash/V2:
V2[~SLASH] --> V2[H +, ~SLASH], U.
==>
V2[SLASH X2] --> V2[H +, SLASH X2], U.

Puis, la règle de propagation (23) fera percoler la valeur du SLASH jusqu'au niveau supérieur.

(23) PROPRULE Propagation/slash/v2/v2 :
V2[SLASH X2] --> V2[H +, SLASH X2], U.
F(0[SLASH]) = F(1[SLASH]), F in TraitsSLASH.

Nous obtenons en GDE la même projection que celle en (19) pour l'analyse GSG. Il faut maintenant décrire l'étape d'unification au niveau supérieur.

3.2.2 Unification du SLASH et du syntagme relatif

En GSG

Nous l'avons vu, en GSG, une règle très générale (18) permet de rendre compte de tous les cas d'extraction de la phrase. SLASH y est hérité et non instancié puisqu'il est bien présent dans la règle. Il est donc le trait de contrôle. La phrase SLASHée est le foncteur et le N2 relatif, l'argument. Ainsi le PAC peut s'appliquer et la valeur du SLASH est unifiée avec la catégorie contrôleur (N2). Dans le lexique, le relatif *que* est noté comme indiqué en (24).

(24) *que* N[PRO +, FEM @1, PLU @2, CAS acc,
WH [WHMOR rel, PRO +, FEM @1, PLU @2, CAS acc]]

Précisons qu'en (24) les valeurs des traits de genre (@1) et de nombre (@2) sont des variables libres. L'unification avec les traits de l'antécédent permettra de préciser les valeurs de ces traits.

En GDE

La règle de DI (25) rend compte de la formation de la proposition relative dans notre grammaire GDE. Cette règle est spécifique à l'extraction d'un syntagme nominal. Des règles différentes permettent de rendre compte de l'extraction d'autres catégories.

(25) IDRULE RelativeN2 :
V2+SUJ[WH [N +, V -, WHMOR rel]] -->
N2[WH [N +, V -, WHMOR rel]], V2+SUJ[SLASH N2].

L'unification entre les traits du N2 contrôleur, le relatif, et le SLASH est assurée par la règle de propagation (26) qui reprend

la notion, présente en GSG, d'unification par le PAC des traits de tête seulement. L'alias Rel est une abréviation pour WH [N +, V -, WHMOR rel].

(26) PROPRULE Encaissement/slash :
V2+SUJ[Rel] --> N2[Rel], V2+SUJ[SLASH N2].
F(1) = F(2[SLASH]), F in Traits_tête_Nominale.

Il faut maintenant récupérer les traits d'accord de l'antécédent.

3.2.3 Unification entre les traits de la relative et ceux de l'antécédent

En GSG

La règle de DI (27) rend compte de l'introduction des relatives.

(27) N1 → H, P [WH [N +, V -, WHMOR rel]]

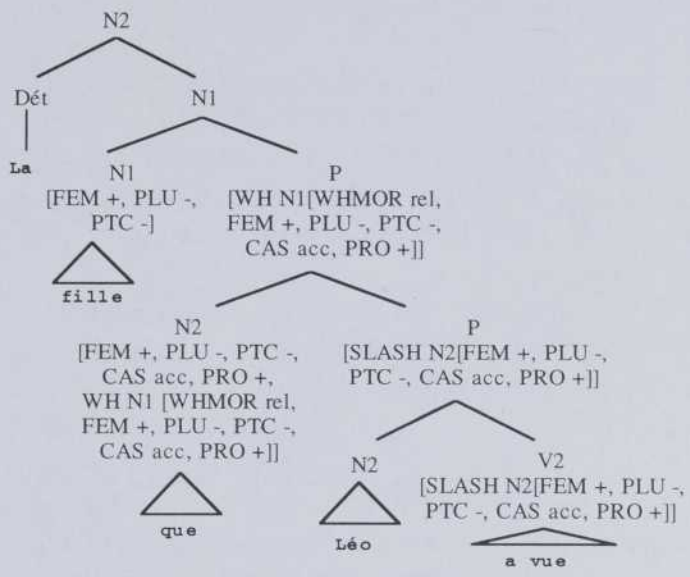
WH est un trait de pied qui identifie entre autres les relatifs. On est ici dans une situation de contrôle définie par le type sémantique de la proposition relative donné en (28).

(28) <TYP N1, TYP N1>

Donc, le N1 est l'argument et la proposition relative, le foncteur. Il faut que la catégorie N1 soit unifiée à la valeur du trait de contrôle de P. Les traits de contrôle disponibles dans Gazdar et al. (1985) sont SLASH (s'il est hérité) ou AGR. SLASH est ici absent. Or, AGR ne peut être le trait de contrôle puisqu'il y aurait conflit avec le trait AGR de la phrase SLASHée dont la valeur est la catégorie du sujet de la proposition relative, *Léo*. Nous proposons donc de faire de WH un trait de contrôle, au même titre que SLASH et AGR. Nous rejoignons ainsi une proposition similaire de Horrocks (1987) qui ajoute REflexif aux traits de contrôle. WH, tout comme SLASH, sera trait de contrôle s'il est hérité. C'est bien le cas ici. Les traits du contrôleur soumis au PAC (N1, FEM + et PLU - et autres traits de tête) seront donc instanciés dans WH.

Le PTP effectuera l'unification entre le N2 relatif et les valeurs de WH de la relative. Ainsi, nous obtenons la projection en (29) où toutes les unifications sont notées.

(29)



Soulignons qu'on doit postuler que les pronoms relatifs ont une entrée lexicale de la forme α[WH α] (voir l'entrée lexicale de *que* en (24)). En effet, le PAC instancie dans le SLASH les traits de tête du N2 relatif, dont FEM et PLU. Or, ceux-ci reçoivent leur valeur de l'antécédent *fille* et sont véhiculés uniquement dans le trait WH. L'entrée lexicale est donc nécessaire pour faire le lien entre les deux.

En GDE

La règle d'introduction de la relative en GDE est telle que notée en (30). L'accord entre le N1 antécédent et la relative est assuré par la règle de propagation en (31).

(30) IDRULE N1/relative : N1 --> H1, V2+SUJ[Rel].

(31) PROPRULE Accord/Rel_SN :
N1 --> H1 , V2+SUJ[Rel].
F(1) = F(2[WH]), F in Traits_tête_Nominale.

La règle de propagation (32) assure la percolation du trait WH du syntagme relatif à la proposition relative conformément au PTP.

(32) PROPRULE Encaissement/accord/rel :
V2+SUJ[Rel] --> X2[Rel], V2+SUJ[SLASH X2].
F(0[WH]) = F(1[WH]), F in TraitsAccord+Catégoriel.

Il faut aussi simuler l'effet de l'entrée lexicale des pronoms relatifs α [WH α] par la règle de propagation en (33).

(33) PROPRULE Propagation/rel_pro :
V2+SUJ[Rel] --> N2[Rel], V2+SUJ[SLASH N2].
F(1) = F(1[WH]), F in Traits_Nominaux.

La projection GDE sera identique à celle que nous avons pour la GSG en (29).

Jusqu'à présent, nous pouvons dire que l'instanciation et la propagation du trait SLASH, tout comme l'unification, sont bien gérées par les principes universels d'instanciation, et que le PAC, en particulier, permet de rendre compte de l'accord au haut d'une dépendance non bornée. Par ailleurs, nous avons vu, en comparant systématiquement les analyses GSG et GDE, que la généralité de ces principes était quelque peu perdue lors de l'implantation, par la quantité de métarègles ou encore de règles de propagation nécessaires pour les simuler. Nous allons voir que la situation est différente lorsqu'on examine le bas de la dépendance, là justement où se fait l'accord du participe passé.

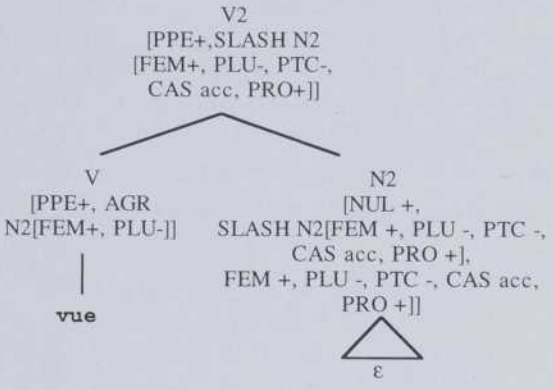
4 L'accord du participe passé

La solution de l'accord du participe passé en GDE devient triviale grâce au formalisme plus souple de l'outil informatique. En effet, il suffit d'unifier deux traits à valeur catégorielle différents, le SLASH N2 de la mère V2 et le trait AGR du participe passé. Le PAC ne peut pas réaliser cette unification puisque nous ne sommes pas en présence d'une configuration de contrôle adéquate. L'analyse GSG nous a donc posé plus de problèmes, comme nous allons le voir maintenant. La solution GDE a été instructive, dans la mesure où elle a suggéré en partie la solution proposée.

4.1 Remise en question du PAC et de la structure du syntagme verbal

Nous pourrions émettre l'hypothèse qu'en (34), dans l'arbre local dominant *vue* et ε, nous sommes dans une situation de contrôle, le participe passé étant le foncteur et le N2 [+NUL], l'argument.

(34)



Cela permettrait en particulier de rendre compte de l'accord du verbe avec son objet dans certaines langues. Nous devrions alors modifier la définition du schéma de contrôle (5) en introduisant la possibilité d'appliquer le PAC, pour le français, entre un verbe [BARRE 0], à condition qu'il soit au participe passé, et un N2 à condition qu'il soit [+NUL] et accusatif. Nous devrions aussi modifier la définition du trait de contrôle (8) en enlevant la restriction BARRE ≠ 0. Cette solution nous paraît toutefois lourde de conséquences et elle est intuitivement peu satisfaisante. En effet, l'accord se fait avec l'objet lorsque celui-ci est antéposé, donc il se fait avec l'élément extrait et non avec le trou laissé par son extraction.

Par ailleurs, cette analyse ne permet pas de rendre compte de tous les cas d'accord du participe passé. En particulier, elle ne couvre pas l'accord dans certaines formes composées, et dans les formes surcomposées. Nos travaux nous ont amenés à examiner de plus près la structure du syntagme verbal en français et à proposer, tout comme Abeillé et Godard (1994), une structure plate pour les temps composés plutôt qu'une structure en

cascade telle que celle proposée jusqu'à présent et calquée sur celle de l'anglais (Gazdar et al. 1982) ; voir à ce sujet Emirkanian et Da Sylva (1995)[10]. Avec une telle structure, la solution que nous venons de proposer et qui implique une modification au PAC, n'est guère plus satisfaisante, et elle ne rend toujours pas compte de certains accords.

Il semble donc que les principes universels d'instanciation, le PAC en particulier, soient inadéquats pour rendre compte de l'accord du participe passé en français. Notons par ailleurs que divers auteurs ont déjà souligné certaines limites du PAC (Jacobson 1987, Hukari et Levine 1986). Il nous reste à explorer les autres mécanismes d'instanciation de traits : les Restrictions de Cooccurrence de Traits et les Spécifications de Traits par Défaut. La solution que nous proposons maintenant exploite ces mécanismes ainsi que la structure plate pour les temps composés.

4.2 Analyse proposée

Pour obtenir la structure plate, il faut introduire l'auxiliaire dans les règles de DI des compléments des verbes. Cela est réalisé par une métarègle (35) qui ajoute une tête auxiliaire à la structure.

(35) MR Insertion des auxiliaires
SV [PASSIF–]--> V[AUX -]
⇓
SV --> H[AUX +, AGRPPé α, SUBCAT β, TYPAUX γ,
PRONL δ, ESS ω], V[AUX -, PPé +, AGR α, SUBCAT β,
AUXREQ γ, PRONL δ, ESS ω].

Il est nécessaire d'apporter certaines précisions à cette métarègle. D'abord, elle introduit une tête dans une règle qui en possède déjà une. Cela ne pose en fait pas de problème, puisque l'identification de la tête d'une règle est un mécanisme métagrammatical : identifier l'auxiliaire comme la tête de la règle

[10] Nous conservons la structure en cascade pour les passifs et les autres prédicatifs. Dans ce cas, la copule *être* a pour soeur une catégorie X2 portant le trait [PRD +].

suffit pour retirer au verbe lexical cette propriété. Ensuite, l'auxiliaire doit servir de véhicule aux traits d'accord du participe passé. L'information pertinente pour spécifier l'accord inclut la présence éventuelle d'un SLASH SN mais aussi la transitivité du verbe et son caractère pronominal (ou non). Il est donc nécessaire de propager à l'auxiliaire ces traits provenant du verbe lexical. Nous réalisons cette propagation en utilisant le partage de variables tel que spécifié dans la métarègle même, ainsi que des RCT et des STD. Le verbe lexical ([AUX -]) devra être au participe passé. La correspondance est assurée entre auxiliaire requis par un verbe (AUXREQ) et le type d'auxiliaire présent (TYPAUX). Les traits SUBCAT, PRONL (pronominal) et ESS[11] sont partagés, comme cela est nécessaire. Le transfert de la sous-catégorisation du verbe permet de signaler entre autres la présence d'un SN objet et en cas d'antéposition de déclencher l'accord du participe passé. Nous introduisons le trait AGRPPé[12] sur l'auxiliaire pour contenir les traits d'accord de son participe passé (indépendamment de ses propres traits d'accord avec le sujet, contenus dans AGR). Ce trait est restreint aux auxiliaires par la RCT suivante.

(36) RCT : [AGRPPé] ≡ [AUX +]

AGRPPé est lié au trait AGR du participe passé par la métarègle. C'est un trait de tête qui percolera au niveau du SV. Il pourra être unifié aux traits d'accord du SN dans le SLASH, le cas échéant.

4.2.1 Accord lors d'une extraction

Avec cette structure pour les temps composés, nous pouvons implanter les règles d'accord du participe passé. Pour le cas de l'extraction d'un SN, c'est le SLASH qui contient ces traits d'accord. Le trait AGRPPé doit alors être unifié aux traits

---

[11] Le trait ESS ("essentiel") est utilisé pour identifier les essentiellement pronominaux.

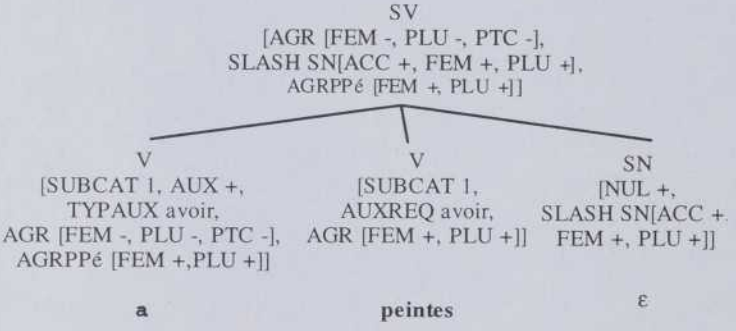
[12] En fait, nous présentons ici une version simplifiée de la solution. La solution complète nécessite un autre trait, AGRO, comme véhicule intermédiaire des traits d'accord dans des structures plus complexes.

d'accord présents dans le SLASH. Sinon, il doit être au masculin singulier. Le trait AGRPPé, nous l'avons dit, est un trait de tête. Instancié sur l'auxiliaire, il percolera au niveau du SV. Il suffit alors d'introduire une RCT qui unifie les valeurs des traits d'accord du SLASH SN[ACC +] avec celles de AGRPPé[13].

(37) RCT : SV[SLASH SN[ACC +, FEM α, PLU β]] ⊃ [AGRPPé [FEM α, PLU β]]

Il y aura donc unification avec le AGRPPé de l'auxiliaire qui est lui-même lié au AGR du participe passé par la métarègle d'insertion des auxiliaires[14]. Par exemple, pour l'énoncé *les tables que Paul a peintes*, la représentation pour le SV *a peintes* est la suivante.

(38)



4.2.2 Accord avec le sujet

Pour les cas d'accord avec le sujet, il faut unifier la valeur de AGRPPé à celle de AGR (qui contient les traits d'accord du sujet). Ceci vaut par exemple lorsque l'auxiliaire est *être*, sauf

---

[13] Cette règle est en fait l'abréviation de quatre règles différentes qui ont le schéma général suivant :
[SLASH [FEM+]] ⊃ [AGR [FEM +]]
[SLASH [FEM -]] ⊃ [AGR [FEM -]]
etc.

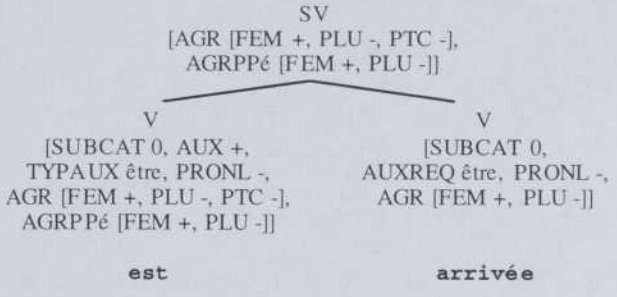
[14] Pour les clitiques objets accusatifs, c'est également une RCT qui déclenchera l'accord. La valeur des traits d'accord dans CLO, un trait de pied du verbe lié à la présence d'un clitique objet, dictera l'accord du participe passé, et sera propagée au trait AGRPPé. Notre traitement de la cliticisation s'inspire de celui de Miller (1991).

pour les essentiellement pronominaux transitifs (ex. *s'arroger*, *s'imaginer*, etc.) et les pronominaux accidentels ; pour ces deux derniers, l'accord est dicté par l'objet s'il est antéposé, sinon le participe est "invariable". L'accord pour l'auxiliaire *être* avec un non pronominal est fait par la RCT suivante[15].

(39) RCT : V[AUX +, TYPAUX être, ~TRDIR -, AGR α] ⊃
[AGRPPé α]

Ainsi, on aura la représentation suivante pour le SV *est arrivée* de *Mireille est arrivée*.

(40)



4.2.3 Participe invariable

Par défaut, le trait AGRPPé portera les valeurs [FEM -, PLU -].

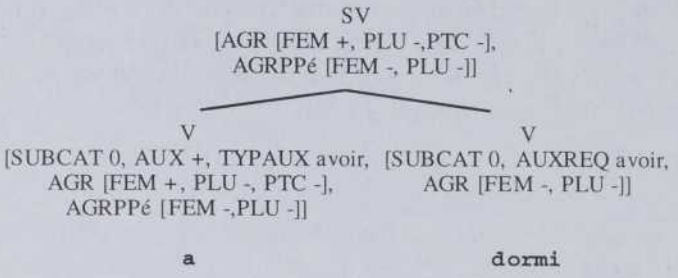
(41) STD : [AGRPPé] ⊃ [AGRPPé [FEM -, PLU -]]

Cette valeur par défaut sera instanciée dans les cas où la valeur de AGRPPé n'aura pas été instanciée par les RCT en (37) et (39). Il s'agit des verbes transitifs directs sans SN extrait, pronominaux ou non (auxquels (37) ne s'applique pas en vertu du trait SLASH) et des participes passés employés avec *avoir* sans SN dans le SLASH (auxquels ni (37) ni (39) ne s'appliquent à cause de SLASH ou de TYPAUX). Regardons à titre d'exemple un verbe qui n'est pas transitif direct, conjugué avec *avoir* comme dans *Mireille a dormi*.

---

[15] Où ~TRDIR est un alias pour ~SUBCAT(1,3), par exemple, où 1 et 3 sont les indices de sous-catégorisation des verbes transitifs directs.

(42)



4.2.4 Accord avec les passifs et autres prédicatifs

Le passif, nous l'avons dit, est un exemple de complément prédicatif de la copule *être* (voir note 10), au même titre que les syntagmes adjectivaux (SA). La règle de DI responsable de les introduire est donnée en (43) (soit 15 l'indice de sous-catégorisation).

(43) SV --> H[SUBCAT 15], SX[PRD +]

La métarègle du passif dérive un passif à partir d'un verbe transitif.

(44) MR Passif
SV --> W, SN
⇓
SV[PASSIF +] --> W, (SP[FORMEP par])

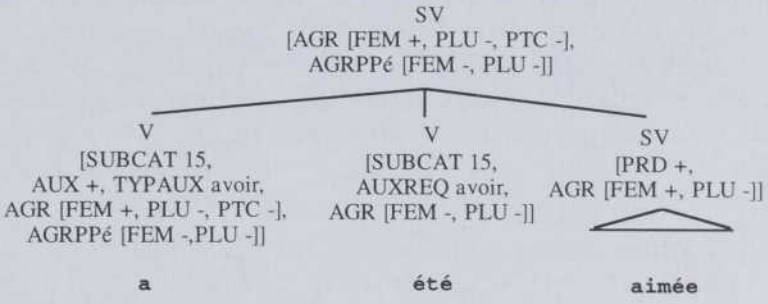
Une RCT impose les traits [PRD+, PPé+] sur les SV portant le trait [PASSIF +].

(45) RCT SV[PASSIF +] ⊃ [PRD +, PPé +]

L'accord des SX prédicatifs avec leur sujet est effectué par le PAC, en vertu de la clause des "catégories prédicatives sans contrôleur". La métarègle d'insertion des auxiliaires pourra s'appliquer, pour former le temps passé pour le passif comme dans *Mireille a été aimée*. En effet, la copule porte les traits [AUX -, AUXREQ avoir] et l'auxiliaire *avoir* peut être inséré. Le participe passé de la copule, *été*, ne manifeste pas d'accord ; il est accompagné de l'auxiliaire *avoir* et il n'y a aucun SN dans le

SLASH. Le trait AGRPPé recevra alors la valeur par défaut [FEM -, PLU -]. La structure est illustrée en (46).

(46)



Parmi les prédicatifs, on compte aussi certains verbes conjugués avec *être* qui peuvent prendre un sens résultatif : *elle est arrivée*, *elle est tombée*. On remarque entre autres que le participe passé avec ses arguments peut être cliticisé, ce qui est compatible avec une structure où ce participe passé est un complément barre 2 (formant un constituant) et non pas avec celle des temps composés. On a le contraste en (47).

(47)	Elle est aimée.	==>	Elle l'est.
	Elle est arrivée	==>	Elle l'est.
	Elle est arrivée hier à cinq heures	==>	*Elle l'est.

Ces prédicatifs à sens résultatif peuvent être insérés dans la règle de DI en (43), si on leur attribue le trait [PRD +].

Nous arrivons ainsi, avec cet ensemble de métarègles, RCT et STD, à rendre compte des différents cas d'accord du participe passé[16]. Le PAC ne peut, quant à lui, décrire que l'accord dans le cas des structures prédicatives, ce qui remet en cause son pouvoir explicatif.

[16] Voir Emirkanian et Da Sylva (1995) pour une description plus complète des données.

5 Conclusion

Le sujet que nous venons d'aborder concerne les trois volets de notre projet de recherche : théorie, description et exploitation. Nous avons montré comment nous avons pu rendre compte de l'accord du participe passé dans le cadre de la Grammaire syntagmatique généralisée. Cette analyse nous a amenés, pour l'accord du participe passé, à rejeter le Principe d'Accord et de Contrôle, l'un des trois principes universels d'instanciation de traits proposés par cette théorie.

La présentation a pu également mettre en relief le fossé qui existe entre la théorie et la pratique. La réalisation d'un modèle opératoire exige non seulement un souci du détail, mais également une grande rigueur. On doit accorder une attention minutieuse aux détails formels. Cette rigueur et cette importance accordée à la couverture de la grammaire sont cruciaux dans le traitement automatique des langues naturelles. En effet ce n'est qu'à ce prix que les solutions théorique et implantée pourront être mises en correspondance étroite.

Bibliographie

Abeillé, A. et D. Godard, 1994, The Complementation of Tense Auxiliaries in French, *Actes 13e WCCFL*, San Diego, CSLI.

Baschung, K., 1988, Contrôle et relations de paraphrase et d'ambiguïté dans les enchassées verbales, *Lexique*, 6, pp. 83-95.

Baschung, K., 1991, *Grammaire d'unification à traits et contrôle des infinitives en français*, Clermont-Ferrand, Adosa.

Boguraev, B., 1988, A natural Language Toolkit: Reconciling Theory with Practice, in U. Reyle et C. Rohrer (éd.), *Natural Language Parsing and Linguistic Theory*, Dordrecht, D. Reidel, pp. 95-130.

Bouchard, D., 1987, A Few Remarks on Past Participle Agreement, *Linguistics and Philosophy*, vol. 10, pp. 449-474.

Bouchard, L. H., L. Emirkanian, D. Estival, C. Fay-Varnier, C. Fouqueré, G. Prigent et P. Zweigenbaum, 1992, First Results of a French Linguistic Development Environment, *Quatorzième Conférence internationale en linguistique informatique, Proc. COLING'92*, pp. 1177-1181.

Emirkanian, L. et L. DaSylva, 1995, *Quelques phénomènes d'accord dans les grammaires syntagmatiques*, manuscrit, Département de linguistique, UQAM, Montréal.

Gazdar, G., E. Klein, G. Pullum et I. Sag, 1985, *Generalized Phrase Structure Grammar*, Harvard University Press, Cambridge MA.

Gazdar, G., G. Pullum et I. Sag, 1982, Auxiliaries and Related Phenomena in a Restrictive Theory of Grammar, *Language*, 3, pp. 591-638.

Grover, C., T. Briscoe, J. Carroll et B. Boguraev, 1989, *The Alvey Natural Language Tools Grammar* (seconde édition), Tech. Report no 162, Computer Laboratory, University of Cambridge.

Horrocks, G., 1987, *Generative Grammar*, Longman, London.

Hukari, T. E. et R. D. Levine, 1986, Generalized Phrase Structure Grammar : A Review Article, *Linguistic Analysis*, 16:3-4.

Jacobson, Pauline, 1987, Generalized Phrase Structure Grammar, *Linguistics and Philosophy*, 10:3, pp. 389-426.

Kayne, R., S., 1985, L'accord du participe passé en français et en italien, *Modèles linguistiques*, 7, 1, pp. 73-89.

Keenan, E. L., 1974, The functional principle: generalizing the notion of "subject of", in M. La Galy, R. Fox et A. Bruck (éd.), *Papers from the 10th Regional Meeting of the Chicago Linguistics Society*, pp. 298-309.

Lefebvre, Cl., 1986, L'accord du participe passé en français: accord = cas, *Revue Québécoise de Linguistique*, vol.15, no 2, pp. 121-134.

Miller, P., 1991, *Clitics and Constituent Structure*, Thèse de doctorat, Université d'Utrecht [publiée (1992), *Clitics and Constituents in Phrase Structure Grammar*, Garland, New York].

FORMALISME UNIFIÉ ET VALIDATION DE GRAMMAIRES

Christophe Fouqueré

Résumé

Parallèlement aux recherches portant sur la spécification de formalismes adaptés aux langues naturelles, des travaux importants concernent le développement de grammaires. La réutilisabilité des données est alors devenue un facteur essentiel. Cette réutilisabilité est à considérer sous deux aspects complémentaires : d'une part, il doit être possible d'effectuer des études différentielles sur ces formalismes en éclairant les propriétés essentielles communes tant d'un point de vue informatique que linguistique ; d'autre part, les grammaires doivent être validées quant à leur forme et au langage qu'elles génèrent. Nous présenterons un formalisme unifié permettant d'aborder ces deux aspects.

Les formalismes linguistiques que nous nous proposons d'aborder sont tous à base d'unification (GPSG, LFG, HPSG) et modélisent la connaissance linguistique par des structures de traits. Les contraintes qui décrivent de manière déclarative les propriétés assignées à ces entités peuvent ainsi être données par des équations ou des formules sur des termes. Nous montrerons comment le formalisme unifié permet d'en rendre compte. Celui-ci conserve les structures fondamentales sans niveler les descriptions comme ce peut être le cas dans d'autres formalismes génériques. Nous indiquerons les propriétés computationnelles qui nous paraissent essentielles.

Par ailleurs, les principes auxquels doivent satisfaire les descriptions linguistiques reposent sur trois points fondamentaux : nécessité des symboles grammaticaux, suffisance de la description par rapport au langage (pas de sur- ou sous-génération) et traitement 'tractable' (un traitement en analyse ou en génération de complexité polynomiale est souhaitable). La validation d'une grammaire a alors un double aspect : d'un point de vue linguistique, le langage engendré par la grammaire doit correspondre au (sous-) langage naturel envisagé, d'un point de vue computationnel, il s'agit de vérifier la nécessité et la suffisance des objets manipulés par la grammaire. Nous indiquerons sommairement une méthode de résolution de contraintes adaptée à cette situation. Nous terminerons en proposant une modélisation du formalisme unifié en logique linéaire. Celle-ci permet en effet d'allier réécriture et calcul logique, et rend compte de certaines fonctionnalités indispensables en linguistique computationnelle.

1 Introduction

Parallèlement aux recherches portant sur la spécification de formalismes adaptés aux langues naturelles et sur la caractérisation des phénomènes linguistiques, des travaux importants concernent les environnements de développement de grammaires et la mise au point d'algorithmes d'analyse et de génération (Shieber 1992, Estival 1993). La réutilisabilité des données, lexique mais aussi grammaires et algorithmes, est alors devenue un facteur essentiel. Cette réutilisabilité est à considérer sous deux aspects complémentaires. D'une part, il doit être possible d'effectuer des études différentielles sur les formalismes linguistiques en éclairant les propriétés essentielles communes tant d'un point de vue informatique que linguistique ; en effet, les principes à la base de ces formalismes sont censés rendre compte de phénomènes langagiers ; toutefois, ces principes ne sont que rarement explicites. Il est dès lors naturel de se donner les moyens de les comparer, plus exactement d'en étudier les mécanismes propres et de savoir dans quelle mesure ces mécanismes sont "exportables" vers d'autres formalismes. D'autre part, les grammaires doivent être validées quant à leur forme et au langage qu'elles analysent / génèrent. La performance d'une grammaire ne se mesure en effet pas seulement à sa capacité d'analyse et de génération, mais aussi à la concision de cette représentation. Là encore, les

formalismes évoluant constamment, il y a nécessité de disposer d'outils informatiques adaptés à cette tâche. Nous avons cherché à développer dans cet esprit un *environnement de génie linguistique* (Bouchard et al. 1991, 1992a, 1992b), et plus particulièrement un *formalisme unifié*, permettant d'aborder ces deux aspects.

Une part croissante de la recherche en linguistique computationnelle s'est orientée vers le développement de langages de description de type attribut-valeur. Ainsi la plupart des formalismes modélisent la connaissance linguistique par des structures de traits, qu'il s'agisse des formalismes grammaticaux actuels (GPSG Grammaire Syntagmatique Généralisée (Gazdar et al. 1985), LFG Grammaire Lexicale-Fonctionnelle (Bresnan 1982), HPSG Grammaire Syntagmatique à Tête (Pollard et Sag 1994)), ou des formalismes computationnels servant à l'implémentation des données grammaticales (PATR-II (Shieber 1992), ELU (Estival 1990)). Le lecteur pourra trouver divers points de vue complémentaires sur les structures attribut-valeur dans Johnson (1988), Shieber (1985, 1992), Smolka (1991). Les contraintes qui décrivent de manière déclarative les propriétés assignées à ces entités (structures attribut-valeur) peuvent ainsi être données par des équations (dans le cas de LFG), ou des formules sur des termes (dans le cas des descriptions conceptuelles, de GPSG, ou de HPSG). Certaines facilités[1] de représentation ne sont pas disponibles partout ; toutefois, et sous certaines réserves mineures, certaines de ces facilités, non disponibles, peuvent y être accessibles : les méta-règles de GPSG sont ainsi à mettre en vis à vis des règles d'insertion lexicale utilisées en LFG et HPSG. Nous montrerons succinctement comment le formalisme unifié peut rendre compte et de ses facilités et de ses relations. Nous avons choisi en effet de conserver les structures fondamentales sans niveler les descriptions comme ce peut être le cas dans d'autres formalismes génériques.

[1] On entend par facilité un moyen formel : méta-règle, connecteur du langage de contraintes ...

Par ailleurs, les principes auxquels doivent satisfaire les descriptions linguistiques reposent sur trois points fondamentaux : nécessité des symboles grammaticaux, suffisance de la description par rapport au langage (pas de sur- ou sous-génération) et traitement 'tractable' (un traitement en analyse ou en génération de complexité polynomiale est souhaitable). La validation d'une grammaire a alors un double aspect : d'un point de vue linguistique, le langage engendré par la grammaire doit correspondre au (sous-) langage naturel envisagé, d'un point de vue computationnel, il s'agit de vérifier la nécessité et la suffisance des objets manipulés par la grammaire. Ces propriétés sont les conditions nécessaires pour que la théorie grammaticale proposée soit explicative. L'explicativité n'est toutefois pas un objectif essentiel dans les systèmes appliqués. Il n'empêche que la concision de l'information reste encore un facteur important. Remarquons que la réflexion sur ce sujet est encore balbutiante : même s'il est vrai que des études théoriques approfondies ont été menées sur les classes de grammaires définies par Chomsky, ce n'est plus toujours le cas pour certains nouveaux formalismes qui s'en écartent. Un des objectifs a donc été de définir formellement ce que nous entendons à ce sujet. Nous avons développé une méthode de résolution de contraintes (Belabbas et Fouqueré 1991, Belabbas et al. 1994) adaptée au formalisme unifié que nous proposons.

Le projet *Environnement de Génie Linguistique* a été développé conjointement par cinq équipes[2]. Son objectif principal était de se donner les moyens informatique, linguistique et théorique de comparer plusieurs formalismes et analyseurs grammaticaux. Nous avions volontairement restreint notre champ de recherches aux formalismes d'unification qui nous semblaient susceptibles de fournir des réponses tant linguistiques qu'informatiques au traitement de la langue : ainsi, les Grammaires Lexicales-Fonctionnelles (Bresnan 1982), les Grammaires Syntagmatiques Généralisées (Gazdar et al. 1985)

[2] CRIN - Nancy, DIAM - INSERM, CNET - Lannion, ISSCO - Genève, GIREIL - UQAM - Montréal.

sont au coeur de nos préoccupations. Cinq composantes constituaient l'essentiel de ce projet :

- une composante d'analyse,
- une grammaire du français,
- un corpus de phrases-tests,
- un environnement graphique,
- une composante de validation.

Cette dernière composante devait être commune à divers formalismes (tous d'unification). Cette approche nous permettait donc d'en faire un module portable. De plus, il devient possible dans ce contexte d'effectuer une étude différentielle poussée des formalismes. Pour cela, nous avons défini un formalisme commun que nous présentons succinctement ci-dessous, et des algorithmes de transfert entre formalismes linguistiques (LFG, ou GPSG, ou ...) et ce formalisme commun, suivant en cela la démarche adoptée par PATR (Shieber 1992). Nous nous en démarquons toutefois en incluant complètement les spécificités de chaque formalisme (ce qui permet un meilleur contrôle).

Cette étude des formalismes grammaticaux nous paraît essentielle. Il convient d'ajouter que celle-ci est complétée par d'autres études liées aux phénomènes linguistiques dont nous esquissons les points forts ci-dessous. Notons aussi que ces recherches prolongent l'étude critique des systèmes d'analyse du français écrit (Fay-Varnier et al. 1991). Il devenait en effet nécessaire d'approfondir l'étude linguistique et computationnelle du français écrit. Ainsi, alors que la syntaxe de l'anglais est un sujet abondamment traité dans la littérature, la syntaxe du français restait, d'un point de vue formel, peu développée. Les résultats de notre analyse indiquaient clairement le besoin de spécifier de nouveaux outils d'aide au développement de tels systèmes, et d'approfondir les prérequis linguistiques et informatiques.

2 Formalisme unifié

Après avoir décrit et commenté sommairement les trois formalismes nous servant de base d'exemples, nous spécifierons

le formalisme unifié que nous avons proposé en indiquant en quoi celui-ci se différencie d'autres formalismes génériques. La section suivante identifiera les critères essentiels à une validation de grammaires.

2.1 Exemples de formalismes d'unification : LFG, GPSG, HPSG

La présentation des formalismes restera très superficielle, nous renvoyons le lecteur intéressé aux références déjà mentionnées.

GPSG et LFG induisent des analyses à double niveau. En GPSG, une analyse sémantique, dont nous ne parlerons pas, est effectuée concurremment à une analyse syntagmatique dont le résultat est un arbre. Le système de règles d'où est issu un tel arbre est fondamentalement hors-contexte. Toutefois, les éléments manipulés par ces règles sont des structures attribut-valeur et le système de règles est 'décomposé' (d'un certain point de vue, on pourra dire factorisé) à l'aide de notions comme les méta-règles (générateur de règles), les contraintes (devant être satisfaites sur toute structure attribut-valeur pour les contraintes de catégorie, sur toute règle pour les contraintes de règle) et les règles de préséance (édictant les principes de positionnement temporel des catégories). À noter que chaque règle doit admettre en partie droite un élément particulier, appelé la tête. Des principes spécifiques régissent le transfert d'informations de cette tête vers le père de la règle. Le niveau syntagmatique de LFG fonctionne à partir d'une grammaire hors-contexte 'standard'. À ce niveau, s'ajoute un niveau fonctionnel. Il s'agit d'associer ainsi à chaque noeud de l'arbre d'analyse, non pas une structure morpho-syntaxique comme en GPSG, mais une structure fonctionnelle. Cette construction est établie à l'aide de contraintes associées aux règles. Pollard et Sag (1994), en développant HPSG, ont tenté de concilier les principes pertinents associés à GPSG et LFG mais aussi à la théorie Gouvernement et Liage prônée par Chomsky (1981). La grammaire est complètement lexicalisée : chaque élément lexical contient les informations suffisantes à l'analyse afférente. L'introduction des règles d'insertion lexicale permet, dans une certaine mesure, de rendre compte des méta-règles présentes

en GPSG ; on les trouve déjà présentes en LFG. Notons aussi que les structures attribut-valeur utilisées combinent des informations morpho-syntaxiques et sémantiques ; en particulier, les principes de sous-catégorisation sont exprimés encore une fois à l'aide de structures attribut-valeur : on y retrouve donc la structure fonctionnelle mise en oeuvre en LFG.

2.2 Formalisme unifié

Afin de rendre compte de ces divers principes et fonctionnalités, il nous semblait utile de conserver dans un formalisme unifié la possibilité d'exprimer ces facilités, quitte à obtenir un formalisme redondant. C'est justement cette redondance, facteur d'expressivité, qui fait défaut dans les formalismes génériques comme PATR ou ELU. Notons que la redondance du cadre formel n'implique pas une redondance dans l'écriture de la grammaire. Des modules de traduction ont été développés permettant de 'convertir' une grammaire écrite en LFG ou GPSG en ce formalisme unifié. Nous indiquerons dans cette section les caractéristiques essentielles de notre formalisme ; des exemples extraits des divers formalismes linguistiques éclaireront notre point de vue.

Une **grammaire**, telle qu'elle apparaît dans notre formalisme unifié, est constituée de trois parties :

- une *partie générative* : règles (de réécriture ou de dominance immédiate), schémas de règles ... qui définissent les éléments d'analyse,
- un *ensemble de contraintes* : règles de préséance, règles de spécification de valeurs par défaut, règles de cooccurrence ... qui doivent s'appliquer sur chaque élément d'analyse,
- un *ensemble de principes* qui régissent les limites formelles que doit satisfaire la grammaire.

Les structures manipulées par la grammaire sont de deux types : *catégorie* et *arbre local*. Rappelons qu'un arbre local est un arbre de hauteur 1. Ces arbres locaux seront induits par les règles de dominance (voir infra), les noeuds (ou éléments d'une règle de dominance) seront des catégories que l'on peut

référencer par $\$_0$ pour la racine, $\$_i$ pour le i^{ème} fils, \$ indiquant le noeud lui-même.

Plus formellement, une catégorie C est un triplet (I, F, A) où
I est l'*identificateur catégoriel*,
F est la *formule combinée* associée à la catégorie,
A est une structure attribut-valeur.

Le fait de considérer une catégorie comme un triplet est linguistiquement motivé puisque chacune des parties a une motivation linguistique spécifique (type versus structure, ou contrainte versus structure, ou f-structure versus c-structure selon les formalismes linguistiques utilisés). Cet éclatement est aussi informatiquement motivé puisque la grammaire obtenue en ne considérant que la partie structure attribut-valeur des catégories intervenant dans une grammaire initiale est hors-contexte. La satisfaisabilité sur cette description partielle est donc une condition nécessaire à la satisfaisabilité de la grammaire initiale[3], et il s'agit d'un problème de complexité polynomiale dans ce cas alors que ce problème est non-décidable dans le cas général.

Une structure attribut-valeur est un ensemble fini de doublets *attribut - valeur* noté $[t_1\ v_1,\ldots,t_n\ v_n]$,[4] où t_1, ..., t_n sont des attributs et $v_1,\ldots,\ v_n$ les valeurs respectives de ces attributs. Un attribut est une constante appartenant à un ensemble fini d'attributs, une valeur est une constante, une structure attribut-valeur ou un ensemble fini de structures attribut-valeur. Nous ne détaillerons pas ici le processus d'unification entre deux structures attribut-valeur, celui-ci est standard (Shieber 1985) : l'unifié de deux structures attribut-valeur est une structure attribut-valeur ayant les doublets attribut-valeur apparaissant dans exactement une des deux structures, et les doublets apparaissant dans les deux structures quand les

---

[3] Nous dirons qu'une grammaire est satisfaisable si le langage engendré par cette grammaire n'est pas vide.

[4] ou encore noté $\begin{bmatrix} t_1 & v_1 \\ \ldots & \ldots \\ t_n & v_n \end{bmatrix}$

valeurs sont les mêmes constantes, ou le doublet attribut-unifié des deux valeurs. L'unification échoue quand un attribut admet des valeurs différentes dans les deux structures. Notons toutefois que deux valeurs spéciales complètent ce langage : $\top$ (le 'vrai') est unifiable avec toute structure sauf $\bot$, $\bot$ (le 'faux') n'est unifiable qu'avec $\bot$. Enfin, une structure attribut-valeur peut être référencée explicitement (par description de la structure elle-même) ou implicitement par référence à une sous-structure de la structure en cours de description. Cette référence implicite est appelée de manière traditionnelle chemin dans la structure.

Exemples :

- $\begin{bmatrix} N+ \\ V- \end{bmatrix}$ est une structure nominale minimalement spécifiée en GPSG,

- $\begin{bmatrix} SUJ & [Item \ \ 'Jean'] \\ OBJ & [Item \ \ 'Marie'] \\ XCOMP & \begin{bmatrix} SUJ & OBJ\bullet \\ Item & 'marcher' \end{bmatrix} \\ Item & 'fait' \end{bmatrix}$ est une structure manipulée en LFG.

Notons que (OBJ •) et (XCOMP • SUJ •) référencent les mêmes valeurs.

L'*identificateur catégoriel* est une constante appartenant à un ensemble fini. Cet identificateur catégoriel permet, par exemple, de représenter de manière directe le symbole de catégorie utilisé dans les règles de LFG. Nous envisageons d'étendre son utilisation à la représentation des types catégoriels dans les grammaires combinatoires (Desclés 1990). $\top$ est un symbole pouvant s'identifier avec n'importe quel identificateur catégoriel.

Les *formules combinées* que l'on considère sont définies comme suit : une formule combinée est la donnée d'un triplet d'ensembles de formules ($\mathcal{S}$, $\mathcal{C}$, $\mathcal{D}$). Les formules de ces ensembles appartiennent toutes au même langage. Elles sont toutefois interprétées différemment. Une telle formule est de l'un des cas

suivants, où p est un chemin, ϕ et ψ sont elles-mêmes de telles formules :

• $p = q$ où q est un chemin,	les chemins p et q partagent la même valeur,
• $p = c$ où c est une constante,	le chemin p a comme valeur la constante c,
• $p \in \{q_i / i \in [1,n]\}$ où q_i est un chemin ou une constante,	le chemin p partage ou a une des valeurs définies par l'ensemble,
• $p\downarrow$	il existe une valeur pour le chemin p,
• $\phi \otimes \psi$	ϕ et ψ doivent être satisfaites,
• $\phi \oplus \psi$	ϕ ou ψ doit être satisfaite,
• $\phi \multimap \psi$	si ϕ est satisfaite, alors ψ l'est aussi,
• $\neg\phi$	ϕ n'est pas satisfaisable,
• $\top$	le 'vrai',[5]
• $\bot$	le 'faux'.

On interprète les ensembles de formules $\mathcal{S}$, $\mathcal{C}$ et $\mathcal{D}$ comme étant respectivement l'ensemble des formules 'strictes', des formules 'de contrainte', des formules 'par défaut'. Chaque formule stricte doit être satisfaite ; une formule ϕ par défaut a l'interprétation suivante : si rien ne contredit le fait que ϕ puisse être satisfaite, alors ϕ doit être satisfaite. En d'autres termes, ϕ doit être satisfaite par défaut. Une formule de contrainte doit être vérifiée a posteriori. Celle-ci ne peut donc servir à spécifier les structures attribut-valeur satisfaisant la formule combinée ; toutefois, l'ensemble des structures étant spécifiées (par les formules strictes et par défaut), chaque telle structure doit satisfaire chaque formule de contrainte. Celles-ci sont principalement utilisées en LFG et notées par un indice c accolé à l'opérateur (p.e. '$=_c$'). Etant donné que ces "opérateurs indicées" ne peuvent être combinés aux opérateurs classiques, il nous a semblé plus judicieux de les distinguer dans un ensemble spécial. Nous allons indiquer au travers d'un exemple simple les différences essentielles. On considère les trois

[5] Le même symbole $\top$ est utilisable dans les trois composantes d'une catégorie. Sa signification est identique dans chaque position.

formules combinées suivantes : F_1 est définie par ({x = 1,x = y}, ∅, ∅) , F_2 par ({x = 1},∅,{x = y}) , F_3 par ({x = y}, {x = 1}, ∅) où x et y sont supposés être des attributs (chemins simples). F_1 admet une solution [x 1,y 1] satisfaisant les deux contraintes strictes. F_2 n'admet pas de solution puisque les ensembles de formules strictes et par défaut ne suffisent pas à satisfaire l'équation x = y. F_3 admet une solution [x ①,y ①] où ① indique le partage des valeurs ; en effet, l'équation x =1 est par défaut, x devant être égal à y, soit $y \neq 1$, x est alors aussi différent de 1 donc l'équation est contredite, soit y = 1 auquel cas x est aussi égal à 1 et l'équation par défaut est d'ores et déjà satisfaite.

À une formule combinée peut être associé un ensemble d'ensembles de structures attribut-valeur. Nous ne détaillerons pas cette propriété ici. Mentionnons seulement le fait qu'une formule peut être mise sous forme normale disjonctive, i.e. une disjonction de conjonctions d'équations. Une conjonction d'équations, si elle est satisfaisable, a pour modèle une structure attribut-valeur :

$\$_i \bullet t_1 = v_1 \otimes \ldots \otimes \$_i \bullet t_n = v_n$ a pour modèle $[t_1\ v_1,\ldots,t_n\ v_n]^{\$_i}$

Toute structure attribut-valeur moins générale[6] que la structure attribut-valeur qui est modèle d'une conjonction d'équations satisfait donc cette conjonction d'équations. Une formule combinée est donc satisfaite par un ensemble de structures attribut-valeur. Le symbole $ sera omis en indice supérieur des structures attribut-valeur.

Exemples (linguistique computationnelle) :

- $\begin{bmatrix} N+ \\ V- \end{bmatrix}$ correspond à la formule ($ • N = +) ⊗ ($ • V = –)
- (dans ℘) ($ • SUBCAT ↓) ⊸ ($ • H = +)

  Cette formule appartient à l'ensemble des contraintes d'une grammaire GPSG. Elle signifie que toute catégorie sous-catégorisée est une tête.

---

6 Une structure attribut-valeur s_1 est plus générale qu'une structure attribut-valeur s_2 si l'unifié de s_1 et s_2 est exactement s_2.

- (dans $\mathcal{D}$) ($ • V = + ⊗ $ • N = –) ⊸ ($ • VFORM = V ⊗ $ • PASSIF = –)

  Une catégorie verbale est par défaut à la forme active.

L' 'unification' de deux catégories (I,F,A) et (I',F',A') a pour résultat (I",F",A") si I = I' et si A et A' sont unifiables, et alors I" = I = I', F" = F ⊗ F', et A" est l'unifié de A et A'. Notons que ce résultat peut ne pas être satisfaisable si F ⊗ F' ne l'est pas.

Une règle à dominance immédiate augmentée (plus simplement AID-règle ou règle par la suite quand il n'y a pas d'ambiguïté) caractérise un ensemble d'arbres locaux. Pour permettre à l'utilisateur, linguiste par exemple, d'inclure directement des relations de préséance entre fils, de factoriser des règles, la syntaxe des règles à dominance immédiate telle qu'elle est définie en GPSG a été augmentée en incluant la possibilité de répéter des groupes de catégories, et de déclarer les contraintes d'utilisation conjointe de ces catégories.

Une *règle de préséance linéaire* (encore appelée LP-règle) A < B indique une relation de préséance entre les deux catégories A et B, cette relation de préséance doit être satisfaite sur tout couple de catégories fils d'un arbre local. C'est pourquoi ces règles font partie de l'ensemble de contraintes de la grammaire (en formalisme unifié).

Exemple :

- le fait qu'un groupe verbal précède un groupe nominal non sujet se traduit par

(⊤,⊤,[N -,V +,BAR 2,PRED -]) < (⊤,⊤,[N +,V -,BAR 2,PRED +,SUJ -])

Une *méta-règle* est une règle de construction d'AID-règle à partir d'un schéma d'AID-règle. Un schéma d'AID-règle est une AID-règle avec utilisation possible de variables de groupe (le symbole spécial W est unifiable avec un groupe). Notons que le symbole spécial V en GPSG référant une catégorie quelconque s'identifie avec la catégorie (⊤, ⊤, ⊤).

Exemple :

(méta-règle du passif de GPSG dans notre approche)

$$(\top, \top, [V +, N -, BAR\ 2]) \longrightarrow \{\otimes, 1, V, (\top, \top, [N +, V -, BAR\ 2])\}$$

$$\Rightarrow$$

$$(\top, \top, [V +, N -, BAR\ 2, PAS +]) \longrightarrow \{\otimes, 1, W, \{\oplus, \top, (\top, \top, [N +, V -, BAR\ 2, PFORM\ par])\}\}$$

$\{\oplus, \top, (\top, \top, [N +, V -, BAR\ 2, PFORM\ par])\}$ permet d'indiquer la présence optionnelle du complément en 'par'.

3 Validation de grammaires

La *validation* de grammaires est un élément essentiel de l'étude linguistique et computationnelle d'une langue naturelle. Elle permet de justifier a posteriori certains choix effectués tant pour la structure du langage de représentation que pour la modélisation elle-même. Il nous semble que cette validation est composée de quatre parties :

- étude de la capacité d'analyse de la grammaire proposée : depuis une dizaine d'années, la mise au point de corpus de phrases-tests permet justement de connaître les 'qualités linguistiques' du système. Ces phrases-tests doivent être annotées, c'est à dire que la structure de la phrase ainsi que les dépendances fonctionnelles ou sémantiques entre syntagmes doivent être indiquées ; il s'agit là d'un processus délicat tant cette annotation peut dépendre de manière importante de choix métalinguistiques,
- étude générative de la grammaire : pour cela, un générateur automatique *paramétrable* doit être mis en place. Son rôle consiste à valider les structures permises et à tester la cyclicité potentielle de la grammaire. Il ne s'agit bien entendu pas d'interdire la cyclicité mais bien d'indiquer au linguiste les structures sujettes à cyclicité,
- étude analytique de la grammaire : il s'agit de l'étude de la *cohérence* des données par rapport aux spécifications linguistiques et formelles,
- étude différentielle de versions en cours de développement de la grammaire. À une difficulté linguistique correspond fréquemment une collection de données linguistiques

(règles, principes, ...), le développement d'une grammaire rendant compte d'un ensemble de difficultés linguistiques peut alors s'avérer un véritable casse-tête ; l'explicitation des contraintes (relevant de la modélisation) liées à chaque difficulté peut en cela aider le linguiste.

Seul le troisième point sera développé dans cette section, nous renvoyons les lecteurs intéressés à (Bouchard et al. 1991) pour des propositions concernant les autres points. Le formalisme unifié présenté dans la section précédente sert de support formel à cette section ; rappelons que son existence permet de ne développer qu'un seul module de validation, outre l'apport dans l'étude différentielle des formalismes. Dans cette section, nous expliciterons chacun des aspects de la cohérence d'une grammaire et nous indiquerons l'état actuel des connaissances sur ces divers sujets. Enfin, nous indiquerons comment des techniques de réorganisation d'ensembles de données linguistiques peuvent être utilisées dans ce cadre (nous rapprocherons cette problématique de celle de la résolution d'ensembles de contraintes).

Il nous faut d'abord indiquer les aspects qui nous ont intéressés en ce qui concerne la cohérence d'une grammaire donnée dans notre formalisme grammatical, c'est-à-dire en ce qui concerne le fait que les éléments utilisés dans la grammaire soient nécessaires, suffisants et aboutissent à un traitement tractable :

- *satisfaisabilité* des formules et contraintes déterminées par la grammaire,
- *cyclicité*, en particulier des ensembles de règles entraînant une cyclicité,
- *superfluité* potentielle de symboles, de règles ...,
- *accessibilité* et *co-accessibilité* des symboles et règles.

Nous rappelons ici schématiquement les méthodes que nous avons proposées (voir aussi Belabbas et al. 1994). Une remarque préalable s'impose. Le principe de vérification de la cohérence dans les formalismes du langage naturel, tel qu'il a été exposé précédemment, est un problème complexe. Cette complexité est engendrée principalement d'une part par la

nature indécidable de certains problèmes fondamentaux (égalité de deux langages par exemple), d'autre part par la NP-Complétude de certains autres problèmes tels que la vérification de la non contradiction de formules comportant une disjonction de valeurs (Kasper et Rounds 1986).

Nous indiquerons brièvement les définitions et quelques propriétés concernant superfluité et cyclicité. Comme on peut aisément s'en rendre compte, ces deux aspects sont fortement liés (dans leur traitement) au problème de l'accessibilité que nous présentons plus avant. Nous ne nous attarderons donc pas sur les détails de ces traitements particuliers.

Un symbole est dit *cyclique* si et seulement si il existe un arbre d'analyse partiel de la grammaire, et un chemin cyclique dans cet arbre passant par ce symbole. Une grammaire est *cyclique* si et seulement si elle admet au moins un symbole cyclique dans un arbre d'analyse valide. On peut en fait étendre cette définition aux cas de cyclicité sur l'ensemble des règles de préséance linéaire.

Les remarques suivantes s'imposent. Un symbole est cyclique si et seulement si il est accessible à partir de lui-même, si et seulement si il est co-accessible à partir de lui-même. La cyclicité sur un symbole relève donc de l'accessibilité lorsque ce symbole est pris comme axiome de la grammaire. Si une grammaire est cyclique, alors elle admet au moins un point de cyclicité accessible. Notons que la réciproque n'est pas vraie. Toutefois, dans le cas d'une grammaire hors-contexte, si une grammaire a des points de cyclicité accessibles, alors elle est cyclique. Nous pourrons donc ainsi obtenir des conditions nécessaires de cyclicité en ne tenant compte que de la grammaire limitée aux spécifications sur les structures attribut-valeur. Puisque le partage de valeurs n'est pas autorisé sur les règles de préséance linéaire, le problème de la cyclicité des règles de préséance linéaire est décidable.

Un symbole est dit *fortement superflu* dans une grammaire si et seulement si ce symbole ne modifie pas la capacité générative forte de la grammaire, c'est à dire si et seulement si l'ensemble des arbres valides reste inchangé.

Un symbole est dit *faiblement superflu* dans une grammaire si et seulement si ce symbole ne modifie pas la capacité générative faible de la grammaire, c'est à dire si et seulement si l'ensemble des phrases générées reste inchangé.

Naturellement, si un symbole est fortement superflu alors il est faiblement superflu. Remarquons de plus que si un symbole est non accessible et non co-accessible alors il est fortement superflu, ce qui donne une condition suffisante de superfluité. Enfin le problème est décidable là encore lorsque la grammaire est limitée aux spécifications sur les structures attribut-valeur.

Dans le formalisme unifié proposé pour cette étude, les composantes contextuelle et non contextuelle des catégories sont en effet représentées de façon distincte (partie formule combinée et partie structure attribut-valeur). Nous nous sommes servis de cette distinction afin de proposer des méthodes computationnelles efficaces : la grammaire limitée aux spécifications sur les structures attribut-valeur est une grammaire hors-contexte dont les symboles sont des structures attribut-valeur, dans le cas où les méta-règles sont "lexicales", c'est à dire qu'elles ne portent que sur des règles dont la tête est lexicale (l'attribut BAR a pour valeur 0 dans le cadre GPSG).

L'accessibilité constitue donc l'opération essentielle à effectuer sur une grammaire. Elle indique si tout objet (contrainte, règle ou catégorie) de la grammaire peut être vraiment atteint (accessible) en phase de génération. De manière standard, nous dirons qu'un objet (catégorie, règle, contrainte) est *accessible* si et seulement si il est utilisé au moins une fois en phase de génération. De même, un objet (catégorie, règle, contrainte) est dit *coaccessible* si et seulement si il est utilisé au moins une fois en phase d'analyse. Pour qu'un objet soit 'utile' il convient donc qu'il soit à la fois accessible et coaccessible.

Dans le module que nous avons développé (Belabbas et al. 1994), la vérification de l'accessibilité à un objet de la grammaire s'effectue en deux passes : l'accessibilité hors-contexte, i.e. sur les spécifications de structures attribut-valeur de la grammaire, essentiellement de complexité polynomiale, fournit des conditions nécessaires, l'accessibilité contextuelle, i.e. sur la grammaire prise dans sa totalité, requiert une méthodologie

issue des CSP (Constraint-Satisfaction Problem, problèmes de satisfaction de contraintes) afin de gérer au mieux l'explosion combinatoire. L'*accessibilité hors contexte* consiste à effectuer un parcours d'arbre d'analyse direct (sans retour arrière) en ne tenant compte d'aucune spécification de défaut, ni d'aucune spécification dont les valeurs qu'elle fait intervenir ne sont pas encore connues (c'est typiquement le cas pour les formules de contrainte). Nous associons un label à chacune des règles utilisées, label donné par la catégorie ayant servi à son déclenchement. Ce label est géré de la façon suivante :

- initialement, la valeur du label, pour une règle donnée, est égale à sa partie gauche,
- à chaque tentative de déclenchement d'une règle, on vérifie si le déclencheur (le père de la règle) apporte bien une information supplémentaire (une nouvelle valeur à un attribut ou un nouvel attribut) qui pourrait éventuellement être transmise aux constituants de la partie droite de la règle et ainsi engendrer de nouvelles catégories admissibles (nouvelles catégories accessibles) ou rendre accessibles de nouvelles contraintes linguistiques. Dans ce cas seulement, le déclenchement de la règle s'effectue en comptabilisant l'information supplémentaire apportée par le déclencheur. Étant donné que le nombre d'attributs ainsi que le nombre de valeurs associées à un attribut sont finis, la terminaison du parcours est assurée. En d'autres termes, cette technique nous permet de déterminer d'abord l'accessibilité des symboles en les considérant comme des constantes (sans prendre en compte l'aspect linguistique, ou encore, sans appliquer les contraintes linguistiques), ce qui permet de délimiter le domaine des valeurs possibles, car l'application des contraintes linguistiques ne ferait que spécifier certains attributs (n'apparaissant pas encore dans la structure, et par conséquent considérés comme pouvant avoir toutes les valeurs possibles pour cet attribut) ou en éliminer complètement certaines structures.

Il reste maintenant à étudier le problème de l'application des contraintes qui font intervenir des valeurs référençant des

éléments extérieurs au sous-arbre local sur lequel elles se trouvent. Nous avons proposé comme solution de construire dans un premier temps le graphe engendré par toutes ces contraintes. Ce graphe est tel que les noeuds représentent les attributs, et les arcs indiquent l'existence d'une contrainte reliant les attributs se trouvant à leurs extrémités. Dans un second temps, ce graphe est restructuré sous une forme arborescente en exploitant les algorithmes de restructuration des graphes de contraintes, les noeuds de cet arbre sont alors des ensembles de contraintes appelés cliques.

Nous avons montré que les stratégies de résolution de contraintes s'adaptent au traitement de la validation des grammaires du langage naturel. En effet, la structure des ensembles de contraintes obtenus en phase de génération est une arborescence de sous-ensembles de contraintes ; l'arbre de génération peut donc être utilisé directement comme étant la structuration de l'ensemble de contraintes. Il est à noter aussi que le problème de satisfaction de contraintes ainsi défini est dynamique au sens où le nombre de variables (i.e. catégories) et de contraintes augmente lors de la vérification de la validation tant qu'une inconsistance n'est pas détectée ; si au contraire une inconsistance est détectée, un ensemble de variables et de contraintes est éliminé du système. La réalisabilité de cette adaptation est garantie par le fait que la méthode de restructuration d'un graphe de contraintes ne fait aucune restriction sur la nature des contraintes utilisées.

Nous indiquons ci-dessous la méthode que nous avons utilisée permettant d'appréhender le problème de la satisfaisabilité d'une formule combinée comme un problème de satisfaction de contraintes (CSP). Un CSP est défini par un ensemble de variables X_1, X_2, ..., X_n et un ensemble de contraintes portant sur ces variables. La résolution d'un CSP consiste à trouver un ensemble des valeurs de variables satisfaisant les contraintes.

Une formule combinée étant donnée, la modélisation en un CSP est décrite par :

- une *variable* du CSP est une structure attribut-valeur $X=[t_1\ v_1, t_2\ v_2, ..., t_n\ v_n]$ où $t_1, t_2, ..., t_n$ sont appelés champs[7] de X dans la terminologie CSP, et n sa dimension.
- le domaine de chaque variable est défini par le produit cartésien des domaines de ses champs.
- les contraintes peuvent porter directement sur les variables du CSP, ou sur les champs des variables. Les contraintes sont dans notre cas les formules qui n'ont pas pu être satisfaites lors du parcours hors contexte.

Remarquons que la dimension de chaque variable du CSP est dynamique, car le nombre de champs de chaque variable peut augmenter au fur et à mesure de la résolution du CSP.

La méthode repose sur une classification des opérateurs en opérateurs actifs (=, $\in$, $\neg$ dans $\mathcal{S}$), semi-actifs ($\multimap$ dans $\mathcal{S}$) et passifs (= et $\multimap$ dans $\mathcal{D}$, = et $\multimap$ dans $\mathcal{C}$). Un traitement local prend en compte des contraintes actives et certaines contraintes semi-actives et une propagation globale est contrôlée par une gestion des variables inter-cliques. La phase de résolution énumérative correspond à la prise en compte des contraintes passives et du reste des contraintes semi-actives suivant un principe énumératif. Afin de réduire encore l'espace de recherche en cours de résolution, nous utilisons une heuristique d'ordonnancement des variables qui consiste à instancier d'abord celles qui apparaissent dans les contraintes semi-actives. Il faut noter qu'à chaque énumération, lors de la phase de résolution, il est possible de réutiliser le pré-traitement précédent jusqu'à ce qu'il ne reste plus de contraintes semi-actives dans le système.

4 Non-monotonie et formalisation logique

En sus des principes précédemment mentionnés auxquels doit se soumettre toute description du langage écrit, la monotonie de l'analyse syntaxique est longtemps apparue comme fondamentale. Toutefois, celle-ci est remise en cause pour

[7] *champ* est le terme consacré en CSP, *attribut* est le terme utilisé pour les logiques attribut-valeur.

plusieurs raisons. La description, très fine, des lexiques actuels requiert une structuration des entrées qui s'apparente aux taxonomies à héritage avec exceptions (Russell et al. 1992). Des processus non-monotones sont aussi utilisés dans le cas de correction de textes (Tsujii et al 1988) et d'apprentissage de langues (Chanier et Pengelly 1990). Parmi d'autres motivations linguistiques, rappelons que Dunin-Keplicz (1988) a montré que les axiomes syntaxiques gérant la résolution de l'anaphore s'exprimaient aisément dans une logique non-monotone. De même (St Dizier 1988), certains phénomènes linguistiques faisant intervenir les quantificateurs peuvent être interprétés en logique des défauts. Enfin, Zernik et Brown (1988) ont noté qu'un des désavantages des langages de description linguistique était la mise à plat des représentations grammaticales. La stratification de la connaissance favorise la prise en compte de préférences. Nous avons montré par ailleurs comment intégrer à un formalisme grammatical des mécanismes non-monotones (Fouqueré 1989, 1991a, 1991b). En particulier, l'extension de la notion linguistique de méta-règle permet de répondre à ces besoins (Fouqueré 1991b). Une méta-règle est un schéma de réécriture de règles, l'extension consiste alors à conditionner cette réécriture par la non-présence d'éléments d'analyse.

Nous ne nous étendrons pas ici sur les applications de cette approche et renvoyons le lecteur aux articles mentionnés. Nous préférons indiquer comment intégrer cette dimension en exhibant une formalisation logique, en l'occurrence dans le cadre de la logique linéaire. Nous esquissons ci-dessous les propriétés de la Logique Linéaire pertinentes à notre propos. Celle-ci mériterait certainement une présentation plus approfondie. Nous ne pouvons que renvoyer le lecteur à deux exposés clairs et fort instructifs : Girard (1987) et Troelstra (1992). Le lecteur y trouvera les développements complets concernant tant une sémantique, qu'une axiomatique.

Une modélisation logique est (souvent) fastidieuse et technique, c'est pourquoi nous ne chercherons pas à être exhaustif mais bien à montrer que l'approche logique linéaire est prometteuse, et ce pour deux raisons principales. D'une part, la logique linéaire permet de réconcilier calcul et réécriture ; nous

allons ainsi montrer que le traitement des règles est plus 'naturel' en logique linéaire qu'en logique classique. D'autre part, certaines fonctionnalités peuvent effectivement être modélisées en logique linéaire, alors même que celles-ci restent problématiques dès lors que l'on cherche à accroître la puissance de la logique classique : les logiques non-monotones censées rendre compte des défauts, si utilisés en description morpho-lexicale et présents en GPSG, induisent de fâcheuses propriétés de complexité.

La logique linéaire ne fait qu'un appel restreint à l'affaiblissement et à la contraction. Rappelons que l'affaiblissement correspond à la capacité de pouvoir ajouter des hypothèses, alors que la contraction permet d'oublier des conclusions. Concernant le traitement de la syntaxe, ces restrictions vont nous permettre d'avoir directement la relation *catégorie = formule*, ce que ne permet pas l'interprétation (traditionnelle) des grammaires en logique classique (interprétation utilisée par Pereira et Warren (1980) pour convertir une grammaire DCG en PROLOG). Elle allie donc la souplesse de la réécriture à la force de la logique. En contrepartie (mais c'est aussi un avantage), les connecteurs binaires 'et' et 'ou' sont dédoublés, en conséquence les constantes aussi. Pour des raisons qui deviennent évidentes dès lors que l'on s'intéresse aux isomorphismes sous-jacents, les connecteurs sont partitionnés en trois groupes. Le groupe multiplicatif comporte deux connecteurs binaires, "fois" (ou "times", connecteur 'et' de type multiplicatif) noté $\otimes$ et "par" ('ou' multiplicatif) noté $\parr$, deux constantes 1 ('vrai' multiplicatif) et $\bot$ ('faux' multiplicatif), et l'implication (linéaire) noté $\multimap$, le groupe additif comporte aussi deux connecteurs binaires, "avec" (ou "with", 'ou' additif) noté & et "plus" ('et' additif) noté $\oplus$, deux constantes $\top$ ('vrai' additif) et 0 ('faux' additif). Il existe aussi une négation, que nous noterons $\neg$[8]. Affaiblissement et contraction ne sont possibles que sur les formules du type !A ou ?A : "bien sûr" (ou "of course") noté !, et "pourquoi pas" (ou "why not") noté ?. La logique linéaire est une logique de gestion de ressources : ainsi, l'implication *substitue* à un ensemble de

[8] Cette négation est notée $\bot$ par Girard.

données une certaine conclusion. Ce mécanisme fondamental est bien entendu exactement le mécanisme de réécriture utilisé en théorie des langages. À côté de ce mécanisme, les divers opérateurs 'et', 'ou', 'non', confèrent à cette logique un pouvoir expressif (et calculatoire) important que nous utiliserons ci-dessous.

Nous allons montrer, à travers un exemple, comment se comporte la modélisation d'une grammaire en logique linéaire. Nous compléterons notre propos au fur et à mesure en incluant les diverses notions présentes dans une grammaire en formalisme unifié.

Soit la grammaire G (hors-contexte à symboles simples à dominance immédiate), où S, A, B sont les symboles non-terminaux, S étant l'axiome de la grammaire, et a1, a2, b les symboles terminaux :

$S \rightarrow A,B$ $\quad$ $B \rightarrow b$

$A \rightarrow a_1, A$ $\quad$ $A \rightarrow a_2$

Les mots a_2b, ba_1a_2, ... appartiennent au langage généré par cette grammaire G. La modélisation de cette grammaire G en logique linéaire est donnée par la formule Γ_G suivante :

$!(S \multimap A \otimes B) \otimes !(A \multimap (a_1 \otimes A) \,\&\, a_2) \otimes !(B \multimap b)$

que l'on peut paraphraser par : chaque 'règle' est utilisable autant de fois que l'on veut (!(...)), la première indique que l'axiome S peut se dériver (i.e. implique / se réécrit) en A et B (et on ne peut oublier une des conclusions), la deuxième indique que A se réécrit soit en a_1 et A, soit en a_2 (soit ... soit ... est dû au connecteur &) (on peut donc avoir l'une des deux dérivations, mais pas les deux, pour chaque application de la règle), la troisième indique que B se réécrit en b.

On peut ainsi prouver :

Γ_G ,S	$\vdash \Gamma_G \otimes (S \multimap A \otimes B) \otimes S$	multiplicité due au connecteur !
	$\vdash \Gamma_G \otimes A \otimes B$	$\multimap$ est fondamentalement une réécriture
	$\vdash \Gamma_G \otimes (A \multimap (a_1 \otimes A)$ et $a_2) \otimes A \otimes B$	(1^{er} pas)
	$\vdash \Gamma_G \otimes ((a_1 \otimes A)$ et $a_2) \otimes B$	($2^{ème}$ pas)
	$\vdash \Gamma_G \otimes a_2 \otimes B$	on choisit une des deux parties du &
	$\vdash \Gamma_G \otimes a_2 \otimes (B \multimap b) \otimes B$	(1^{er} pas)
	$\vdash \Gamma_G \otimes a_2 \otimes b$	($2^{ème}$ pas)
	$\vdash a_2 \otimes b$	multiplicité zéro du connecteur !

Le 'mot' a_2b est ainsi dérivé.

De même, on peut dériver a_1a_2b :

Γ_G ,S	$\vdash \Gamma_G \otimes (a_1 \otimes A)$ et $a_2 \otimes B$	(comme précédemment)
	$\vdash \Gamma_G \otimes a_1 \otimes A \otimes B$	on choisit une des deux parties du &
	$\vdash \Gamma_G \otimes a_1 \otimes (A \multimap (a_1 \otimes A)$ et $a_2) \otimes A \otimes B$	(1^{er} pas)
	$\vdash \Gamma_G \otimes a_1 \otimes ((a_1 \otimes A)$ et $a_2) \otimes B$	($2^{ème}$ pas)
	$\vdash \Gamma_G \otimes a_1 \otimes a_2 \otimes B$	on choisit une des deux parties du &
	$\vdash a_1 \otimes a_2 \otimes b$	(comme précédemment)

Remarquons que l'ordre des éléments dérivés n'influe pas puisque le connecteur $\otimes$ est commutatif : les règles représentées sont des règles de dominance immédiate (i.e. commutative).

La première amélioration consiste à utiliser des structures attribut-valeur en lieu et place des symboles simples. Nous reprendrons, dans ce cas, l'approche adoptée classiquement : une structure attribut-valeur est considérée comme un terme dans une logique du premier ordre. Ainsi, la structure [t_1 v_1, t_2 v_2] est convertie en le terme liste(t_1(v_1), liste(t_2(v_2), fin)), où

liste(,) est une fonction binaire et fin une nouvelle constante. On supposera que struct(.) est un prédicat dont l'argument est un tel terme. La logique linéaire du premier ordre admet des quantificateurs ayant le même comportement que pour la logique classique, cette amélioration ne pose donc aucun problème particulier. L'existence de variables permet là encore le partage de valeurs. Pour des questions de lisibilité toutefois, nous conserverons la notation en structure.

Une deuxième amélioration consiste à savoir traiter correctement les formules combinées présentes dans les catégories. Commençons par les formules 'strictes'. Celles-ci doivent être satisfaites sur la structure attribut-valeur associée. Remarquons d'abord qu'une telle formule stricte est convertie directement en logique linéaire (la présence d'une valeur pour un attribut est dénotée par un quantificateur existentiel, l'égalité s'introduit comme en logique classique). L'ensemble formule stricte - structure attribut-valeur est alors représenté par la conjonction (à l'aide de $\otimes$) des deux formules linéaires correspondantes. Les formules par défaut nécessitent un traitement plus élaboré que la logique linéaire permet néanmoins. Fouqueré et Vauzeilles (1994) montrent en effet que certains types d'inférences non-monotones sont modélisables en logique linéaire, les formules par défaut s'inscrivent dans ce cadre : l'opérateur & permet en effet de choisir entre l'application de la règle et son oubli (implication vers 1). Enfin, les formules de contrainte doivent être vérifiées a posteriori : encore une fois, puisqu'une hypothèse ne peut être ajoutée, si la formule est satisfaite, c'est que les conditions de sa satisfaction existaient. Une formule de contrainte F sera donc traduite par $(F \multimap F)$ & $(\neg F \multimap \bot)$ (si la formule est satisfaite, on la laisse inchangée, sinon on obtient le faux).

Il nous reste maintenant à étudier le cas des méta-règles et des règles de préséance linéaire. Il importe de constater que nous sommes obligés de borner la longueur de W intervenant dans les schémas de règles (W dénote une séquence de catégories). Nous considérerons donc que l'appel des méta-règles ne peut être récursif (ce qui est le cas avec la restriction de lexicalité définie en GPSG). Supposons une méta-règle de la forme

$S \rightarrow A,B,W \Rightarrow S \rightarrow C,W$ et prenons 2 comme borne pour la longueur de W. Nous convertissons cette méta-règle en la formule :

$\forall x \forall y\ (S \multimap A \otimes B \otimes struct(x) \otimes struct(y)) \multimap (S \multimap C \otimes struct(x) \otimes struct(y))$

Enfin, les règles de préséance linéaire sont en l'état actuel considérées comme des contraintes sur des indices de positionnement, ces indices sont des entiers intervenant comme argument du prédicat struct(.) ; ce traitement est classique, il conviendrait d'étudier pour ce cas l'intérêt que présenterait une modélisation en logique linéaire non commutative (par rapport au connecteur $\otimes$) nous permettant de lever cet inconvénient.

5 Conclusion

Nous nous sommes proposé dans cet article de définir les conditions de validation d'une grammaire. Nous avons ainsi pu exposer un formalisme unifié nous servant de cadre de travail. Il nous semble que la systématisation de la démarche (tenir compte de l'ensemble des spécificités des formalismes) permet de mieux appréhender les forces et faiblesses de ces systèmes. Les premiers résultats montrent que l'étude de la cohérence passe par l'étude de la satisfaisabilité et de l'accessibilité. Ces deux points admettent des réponses directes si l'on ne s'attache qu'aux sous-structures définissant une grammaire hors-contexte. Nous avons par ailleurs indiqué une méthode permettant de traiter de manière efficace le cas général quand cela est possible, en liant ce problème au cadre de la satisfaction de contraintes.

Enfin, la modélisation en logique linéaire nous paraît importante. Celle-ci ouvre en effet des perspectives afin de mieux comprendre les prérequis logiques des notions utilisées depuis quelques années en linguistique computationnelle. Nous avons montré de plus que cette modélisation permettait de rapprocher l'optique réécriture de l'optique logique.

Bibliographie

Belabbas A., H. Bennaceur et C. Fouqueré, 1994, Résolution de CSP par classification des contraintes : application aux logiques attribut-valeur, *9ème Congrès Reconnaissance des Formes et Intelligence Artificielle*, pp. 409-420, Paris.

Belabbas A. et C. Fouqueré, 1991, Cohérence dans les grammaires du Langage Naturel, *Actes des Journées du LIPN*, pp. 211-222, Université Paris-Nord.

Bouchard L., L. Emirkanian, D. Estival, C. Fay-Varnier, C. Fouqueré, G. Prigent et P. Zweigenbaum, 1991, *Environnement de Génie Linguistique*. Rapport de recherche, LIPN.

Bouchard L., L. Emirkanian, D. Estival, C. Fay-Varnier, C. Fouqueré, G. Prigent et P. Zweigenbaum, 1992, First Results of a French Linguistic Development Environment,*COLING'92*, pp. 1177-1181.

Bouchard L., L. Emirkanian, D. Estival, C. Fay-Varnier, C. Fouqueré, G. Prigent et P. Zweigenbaum, 1992, EGL : a French Linguistic Development Environment, *Avignon'92*, pp. 227-238.

Bresnan J., 1982,*The Mental Representation of Grammatical Relations*, Cambridge, Ma, MIT Press.

Chanier T. et M. Pengelly, 1990, Conceptual Modelling in Error Analysis in Computer Assisted Language Learning Systems, Washington NATO Workshop : *Intelligent Tutoring Systems and Second Language Learning*.

Chomsky N., 1981, *Lectures on Government and Binding*, Dordrecht, Foris.

Desclés J.-P., 1990, *Langages applicatifs, langues naturelles et cognition*, Paris, Hermès.

Dunin-Keplicz B., 1988, Partial Reconstruction of Coreferential Structure, *ECAI'88*, pp. 732-737.

Estival D., 1990, *ELU User Manual*, Technical Report, ISSCO.

Estival D., 1993, Une grammaire pour l'analyse et la génération, *Traitement automatique des langues*, vol. 34, n° 1, pp. 83-99.

Fay-Varnier C., C. Fouqueré, G. Prigent et P. Zweigenbaum, 1991, Modèles syntaxiques des systèmes d'analyse du français, *Technique et Science Informatiques*, vol. 10, n° 6, pp. 403-425.

Fouqueré C., 1989, Is Nonmonotonic Grammar A Solution to Natural Language Processing? *Eurospeech'89*, Paris, pp. 394-397.

Fouqueré C., 1991a, Evidence for Preferential Analysis, *ACH/ALLC'91*, Phoenix, pp. 153-158.

Fouqueré C., 1991b, Preferred analysis? A nonmonotonic parsing algorithm, *ICO'91*, Montréal, pp. 149-153.

Fouqueré C. et J. Vauzeilles, 1994, Linear Logic and Exceptions, *Journal of Logic and Computation*, vol 4, n° 6, pp. 859-875.

Gazdar G., E. Klein, K. Pullum et I. Sag, 1985, *Generalised Phrase Structure Grammar*, Cambridge, Ma, Harvard University Press.

Girard J.-Y., 1987, Linear Logic, *Theoretical Computer Science*, vol. 50, pp. 1-102.

Johnson M., 1988, *Attribute-Value Logic and the Theory of Grammar*. CSLI Lecture Notes, vol. 16, Stanford, Ca.

Kasper R. et W. Rounds, 1986, A logical semantics for feature structures, *ACL Proceedings*, 25th Annual Meeting, pp. 235-242.

Pereira F. et D. Warren, 1980, Definite clause grammars for natural language analysis : a survey of the formalism and a comparison with Augmented Transition Networks, *Artificial Intelligence*, vol. 13, pp. 231-278.

Pollard C. et I. Sag, 1994, *Head-driven Phrase Structure Grammar*, Chicago, The University of Chicago Press.

Russell G., A. Ballim, J. Carroll et S. Warwick-Armstrong, 1992, A Practical Approach to Multiple Default Inheritance for Unification-Based Lexicons, *Computational Linguistics*, vol. 18, n° 3, pp. 311-337.

Sells P., 1987, *Lectures on Contemporary Syntactic Theories*, CSLI Lecture Notes, vol. 3, Stanford, Ca.

Shieber S., 1985, *An Introduction to Unification-based Approaches to Grammar*, CSLI Lecture Notes, vol. 4, Stanford, Ca.

Shieber S., 1992, *Constraint-based Grammar Formalisms*, Cambridge, Ma, MIT Press.

Smolka G., 1991, Computational and Logical Foundations of Constraint Grammars, *3rd European Summer School in Language, Logic and Information*, Saarbrücken.

St Dizier P., 1988, Contextual Discontinuous Grammars, *Natural Language Understanding and Logic Programming*, V. Dahl et P. St Dizier (éd.).

Troelstra A.S., 1992, *Lectures on Linear Logic*, CSLI Lectures Notes, Vol. 29, Stanford, Ca.

Tsujii J., Y. Muto, Y. Ikeda et M. Nagao, 1988, How to Get Preferred Readings in Natural Language Analysis, *COLING'88*, pp. 801-805.

Zernik U. et A. Brown, 1988, Default Reasoning in Natural Language Processing, *COLING'88*, pp. 806-810.

ANALYSE LINGUISTIQUE PARALLÈLE

Lyne Da Sylva, Denis Bouchard, Lorne Bouchard,
Henrietta Cedergren, Anne-Marie Di Sciullo, André Dugas,
Louisette Emirkanian, Betsy Klipple, François Léveillé,
Hélène Perreault, Céline Robitaille, Jan van Voorst

Résumé

L'article décrit les résultats d'un projet de recherche intitulé "Analyse linguistique parallèle" effectué par des chercheurs des départements de linguistique et d'informatique de l'UQAM en collaboration avec le centre ATO•CI de l'UQAM ainsi que la compagnie ALEX Informatique. Le projet visait à élaborer un système de traitement de la langue naturelle dans un environnement informatique parallèle. Nous nous sommes penchés sur la résolution de l'ambiguïté associée au rattachement des syntagmes prépositionnels, une problématique difficile pour l'analyse linguistique par ordinateur. Nous avons développé un prototype qui cherche à optimiser l'exploitation du parallélisme en reconceptualisant le problème de façon parallèle. La résolution du problème d'ambiguïté dans notre prototype procède par l'interaction d'experts, selon une conception multi-agents ; les experts correspondent aux niveaux d'analyse linguistique suivants : la syntaxe, la sémantique, le lexique, la morphologie et la prosodie. Le formalisme utilisé pour le prototype s'inscrit dans le cadre des grammaires d'unification, qui encodent les règles linguistiques en termes de structures d'attribut-valeur. Ces règles sont traduites en règles de production, compilées en un réseau de noeuds représentant un graphe de décision qui est reproduit sur chaque processeur agissant en parallèle. Le système est implanté

sur un ordinateur à l'architecture parallèle VOLVOX de la compagnie ALEX.

1 Introduction

Le projet de recherche intitulé "Analyse linguistique parallèle"[1] s'inscrit dans le paradigme du parallélisme, avec application dans le domaine de l'analyse linguistique : il a pour objectif l'élaboration d'un prototype parallèle pour analyser la langue (en l'occurrence le français) et plus particulièrement pour résoudre certaines des ambiguïtés dans l'analyse des phrases. Le projet cherche en fait à reconceptualiser ce problème de résolution d'ambiguïté dans l'environnement d'une chaîne de traitement linguistique parallèle.

Un énoncé est ambigu si on peut lui attribuer plus d'une interprétation ; l'ambiguïté est omniprésente dans les langues naturelles, et représente un problème considérable dans les systèmes de traitement informatique. Ce problème a été largement documenté, entre autres dans Church et Patil (1982). Nos objectifs sont, d'une part, la conception de la résolution du problème de l'ambiguïté par des techniques d'analyse parallèle, et d'autre part l'identification, la représentation et la communication des informations linguistiques des divers niveaux d'analyse, à savoir la morphologie, la syntaxe, la sémantique, le lexique et la prosodie, pour résoudre ces ambiguïtés. La mise en oeuvre de ces résultats est effectuée par l'implantation d'un prototype d'analyse linguistique parallèle sur une machine VOLVOX munie de 64 transputers T800 ayant 4 Mo de mémoire vive chacun.

2 Problématique

Nous présentons d'abord l'aspect linguistique du problème puis nous exposons les contraintes imposées par l'approche parallèle.

---

[1] Ce projet a été rendu possible grâce à une subvention de la compagnie ALEX-Informatique (Montréal) par l'entremise du Centre ATO•CI de l'Université du Québec à Montréal.

2.1 Problématique linguistique

Notre prototype a été élaboré dans le but précis d'analyser un type d'ambiguïté qui relève principalement de la syntaxe qui parfois génère plus d'une structure pour une même phrase, selon le rattachement des syntagmes prépositionnels. Notre grammaire analyse correctement, par exemple, les phrases en (1), dont les structures sont données à la figure 1.

(1) [[Paul]$_{SN}$ [distribue [les bonbons [au miel]$_{SP}$]$_{SN}$]$_{SV}$]$_{P}$
[[Paul]$_{SN}$ [distribue [les bonbons]$_{SN}$ [aux enfants]$_{SP}$]$_{SV}$]$_{P}$

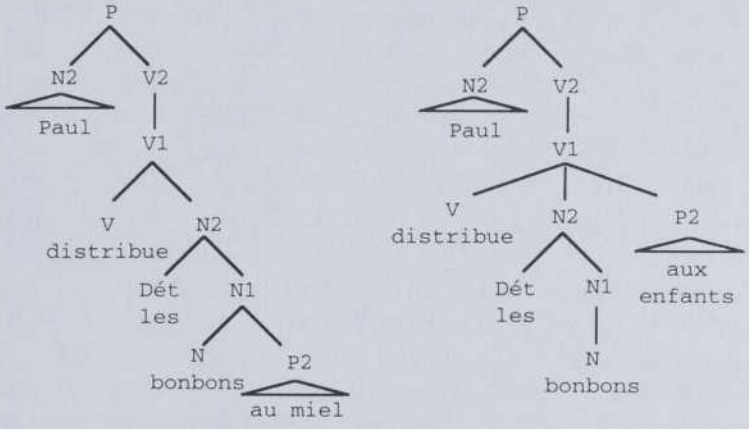


Figure 1 : Structure des phrases exemples

Dans le premier cas, le syntagme prépositionnel constitue le complément du nom[2]. Dans le second cas, le syntagme prépositionnel est un argument du verbe. Toutefois, les règles syntaxiques génèrent dans un premier temps ces deux structures pour chacun des exemples. Ainsi *au miel* sera également analysé comme un argument possible du verbe et *aux enfants* comme un complément du nom, d'où l'ambiguïté à résoudre.

Dans notre système parallèle de conception modulaire, les différents experts en place, correspondant à divers niveaux

[2] Nous utilisons ici "complément" pour décrire à la fois les arguments et les adjoints.

d'analyse linguistique, sont en mesure de donner de l'information pertinente au moment opportun grâce à leur interaction. C'est pourquoi une analyse juste peut être produite pour chacune des phrases et que, par le fait même, les analyses inadéquates sont éliminées. La phrase suivante en constitue un exemple.

(2) Paul distribue des bonbons aux enfants

La syntaxe produit dans un premier temps les deux analyses que nous avons vues précédemment, soit une où *aux enfants* est le complément du nom *bonbons*, et l'autre où *aux enfants* est l'argument du verbe. L'analyse morphologique identifie le suffixe du pluriel des mots *bonbons* et *enfants*. Par ailleurs la sémantique produit la structure argumentale du verbe à partir de certaines propriétés sémantiques listées dans les descriptions lexicales. Le verbe *distribuer* sous-catégorise, entre autres, un objet direct et un objet indirect. En plus, le verbe exige que l'objet indirect dénote une entité humaine, un fait qui va de pair avec l'information de l'entrée *enfant* qui est fournie par le lexique. Le lexique indique que le mot *bonbon* peut être suivi d'un syntagme prépositionnel en *à* qui dénote de la nourriture, comme dans *bonbons au miel*. Le syntagme *aux enfants*, n'étant pas identifié comme de la nourriture, ne peut pas jouer le rôle de complément de *bonbons*. Ainsi l'analyse produite par le système identifie correctement *aux enfants* comme complément du verbe ici. Notons toutefois que cette analyse se limite aux informations fournies par les modules choisis et il s'ensuit que notre modèle ne peut pas fournir une analyse adéquate si la désambiguïsation d'une phrase dépend d'autres informations, par exemple pragmatiques.

Pour restreindre la portée du projet, nous avons formulé un corpus d'à peu près soixante phrases contenant des verbes ayant des syntagmes prépositionnels compléments optionnels en *de* et en *à*, comme *distribuer* et *acheter* par exemple. Des règles linguistiques ont été formulées pour traiter ce corpus restreint.

Les formalismes linguistiques expriment naturellement le parallélisme des phénomènes linguistiques et permettent une exploitation efficace du traitement parallèle. Dans ce projet,

nous avons exprimé les informations linguistiques de façon modulaire afin d'optimiser la parallélisation du prototype. Le traitement que nous avons développé est parallèle de deux façons. Premièrement, le système ne traite pas une phrase "mot par mot" ou "syntagme par syntagme", mais peut s'appliquer à tous les mots ou syntagmes en même temps. Les syntagmes regroupent les mots en plus grandes unités ou constituants et le système peut traiter un syntagme individuel dès que celui-ci est créé. Deuxièmement, les règles relevant des divers niveaux, ou modules d'analyse linguistique, sont modulaires : elles visent des phénomènes distincts, relativement indépendants, qui sont toutefois tous impliqués conjointement dans la description totale d'un fragment de langue. En outre, nous avons poussé plus loin ce dernier aspect du traitement : là où les analyses linguistiques postulent généralement un ordre d'application pour les niveaux (par exemple, toute la morphologie doit être exploitée avant de passer à l'analyse syntaxique), le système peut appliquer simultanément aux données des règles provenant de différents modules.

2.2 Problématique informatique

L'élaboration d'un système d'analyse linguistique parallèle doit tenir compte de certaines contraintes reliées au formalisme qui doit être compatible à la fois avec la linguistique et avec l'informatique.

La description d'un système linguistique comporte plusieurs types de règles qui correspondent à différents phénomènes linguistiques. Souvent, ces règles ne fournissent qu'un minimum d'informations pertinentes ; plus précisément les règles sont concises, ce qui n'empêche pas l'information pertinente d'y être, souvent de façon implicite. Du côté informatique, la situation est inverse : il est préférable d'avoir un formalisme uniforme contenant de façon explicite un maximum d'informations pertinentes. Ainsi d'une part le formalisme utilisé pour décrire le système de linguistique computationnelle doit posséder une sémantique permettant de tenir compte des réalités linguistiques et d'autre part, il doit être traduit dans un formalisme approprié pour l'ordinateur.

Dans le développement de tout système parallèle plusieurs difficultés sont rencontrées. D'abord, le déroulement d'un système séquentiel cache souvent des dépendances chronologiques entre plusieurs processus. Dans un contexte parallèle, le partage des tâches est crucial ainsi que leur synchronisation. Le système doit pouvoir définir un maximum de processus indépendants, minimiser les interactions entre eux et garantir un résultat identique quelle que soit la synchronisation entre les processus parallèles. En d'autres termes, il faut éviter les courses. Aussi, dans un système parallèle où les tâches sont distribuées parmi différents processeurs, il est souvent nécessaire de permettre aux processeurs de s'échanger des messages. Le système d'échange de messages doit éviter les situations où deux processus s'attendent mutuellement. En outre, plus les messages échangés sont nombreux et plus la performance du système est réduite, d'où la nécessité de limiter les interactions. Finalement, pour exploiter au maximum la puissance offerte par une machine parallèle, il est crucial de spécifier les informations de façon à aider le système à identifier le potentiel de parallélisme, et ce, dans notre cas, tout en conservant la logique et l'essentiel du système linguistique. Nous verrons comment le formalisme que nous avons développé et la façon dont nous avons choisi d'exprimer l'information linguistique nous ont permis d'atteindre en grande partie notre but, soit d'exploiter le potentiel de parallélisme du problème à l'étude.

3 Brève description du système

Dans un premier temps, nous allons esquisser le système dans ses grandes lignes ainsi que donner quelques précisions sur le formalisme utilisé. Le système, tel que documenté dans Bouchard et al. (1993), peut être décrit de deux points de vue, linguistique et informatique. Du point de vue linguistique, le système contient des ensembles de règles correspondant aux divers experts linguistiques. Chaque expert ou module contient le type de règles pertinent à celui-ci. Toutefois ils utilisent tous le formalisme des grammaires d'unification (Shieber 1986), ce qui facilite le partage d'information lorsque c'est nécessaire. Tous les objets linguistiques sont représentés à l'aide de

structures attribut-valeur (SAV). Ce sont des ensembles de paires attribut-valeur par lesquels sont encodées les diverses informations linguistiques. Un attribut correspond à une propriété linguistique et peut posséder différentes valeurs. Par exemple, l'attribut **fém** à valeur **+** ou **-** désigne le genre féminin et masculin, respectivement. Si l'on retient un trait **cat** (catégorie lexicale), **fém** et **plu** (nombre), un item lexical comme le déterminant *la* a la forme suivante :

(3) la : [cat:Dét, fém:+, déf:+, plu:-]

Les interfaces entre les modules sont assurées grâce à un vocabulaire partagé. Par exemple, le trait **fém** indique le trait de genre à la fois en morphologie et en syntaxe. Des abréviations sont utilisées pour des ensembles de paires attribut-valeurs récurrentes ; ainsi N0 représente [cat:N, barre:0].

Il y a certaines règles qui sont spécifiques à un module donné, par exemple les règles-PL en syntaxe (voir en 4.2). Par contre, deux types de règles sont utilisées par tous les modules : les règles de redondance et les règles de défaut. Les règles de redondance spécifient des cooccurrences de traits nécessairement vraies à l'intérieur d'une SAV tandis que les règles de défaut donnent une valeur par défaut à un trait non spécifié[3]. Par exemple :

(4) REDONDANCE si [cat:N, propre:+] alors [pers:3]
toujours vraie

(5) DÉFAUT si [cat:N] alors [pronom:-]
parmi les noms, les pronoms sont marqués [pronom:+] au lexique, les autres sont non spécifiés, et deviennent [pronom:-] par cette règle de défaut

La morphologie et la syntaxe utilisent quant à elles des règles de combinaison, pour former des mots à partir de morphèmes ou des syntagmes à partir de mots, respectivement. Ces règles sont décrites en 4.2 et en 4.5

---

[3] Le système contient en fait deux types différents de règles de défaut, les défauts *statiques* et les défauts *dynamiques*, que nous distinguons en 5.1.

Du point de vue informatique maintenant, le système se présente comme un ensemble de correspondances entre le formalisme linguistique et l'implantation sur la machine parallèle. Les règles linguistiques subissent une série de transformations automatisées pour être finalement incorporées dans l'environnement des acteurs qui agissent en parallèle. Elles sont d'abord transformées en règles de production.

Nous avons opté pour les règles de production entre autres raisons parce que l'utilisation d'un langage déclaratif est une première étape nécessaire pour pouvoir réaliser un traitement parallèle. Il faut également qu'il soit possible d'effectuer plusieurs calculs simultanément. Le paradigme des systèmes de règles de production est un pont idéal entre le formalisme linguistique et le problème de parallélisme par ordinateur. D'abord, il est possible de traduire nos règles linguistiques en règles de production. Ensuite, beaucoup de recherches ont été effectuées sur les systèmes de règles de production et on peut maintenant considérer que les solutions séquentielles sont quasi-optimales. Ces systèmes sont encore relativement lents, c'est pourquoi l'attention s'est dirigée vers l'utilisation du parallélisme. La majorité de ces efforts visaient une architecture à mémoire partagée. Une telle orientation est incompatible avec l'architecture à mémoire distribuée de l'ordinateur VOLVOX à notre disposition ; mais l'approche générale reste la même : il s'agit d'utiliser des idées développées dans un contexte séquentiel mais qui sont compatibles avec le modèle de parallélisme utilisé.

Notre système est divisé en deux parties : une phase de compilation de la grammaire linguistique et une phase d'analyse de texte proprement dit. La phase de compilation est complexe et sera esquissée plus loin. Quant à l'analyse, le texte soumis au système est une phrase écrite en code orthographique et qui contient aussi des marqueurs prosodiques. On aurait pu imaginer que l'input soit phonétique ou phonologique, mais ceci aurait représenté une quantité de travail qui aurait largement dépassé le cadre de ce projet[4]. L'analyse procède par

[4] Voir en effet Erman et al. (1980) sur le système HEARSAY-II.

construction d'unités linguistiques à partir des mots de la phrase ; des constituants de plus en plus grands sont identifiés et construits, selon les règles disponibles. Les objets linguistiques créés seront seulement ceux légitimés par les règles ; ainsi les analyses résultantes sont celles qui sont validées par les règles.

Après cet aperçu du système, nous décrivons plus en détail l'ensemble des règles linguistiques du prototype.

4 Les experts linguistique et la modularité

Cette section présente l'organisation et le fonctionnement des experts linguistiques. Bien que l'analyse linguistique soit souvent traitée comme séquentielle, cette séquentialité n'est pas essentielle. Dans notre système, nous avons cherché à réduire les dépendances séquentielles au minimum, en fractionnant les tâches de la grammaire. Le prototype est donc composé de cinq experts linguistiques, ou modules : le lexique, la syntaxe, la sémantique, la prosodie, et la morphologie. Chacun est un bloc de règles qui s'appliquent à des types d'informations spécifiques. Les règles de redondance et de défaut ne s'appliquent en général qu'à l'intérieur du module auquel elles appartiennent, mais certaines sont définies globalement et s'appliquent à toutes les règles. Cette modularité facilite une parallélisation du prototype linguistique. Chaque module opère sur une partie différente de la phrase ou de la structure indépendamment des autres. Nous décrivons tour à tour chacun des modules linguistiques.

4.1 Le lexique

Le lexique contient les entrées lexicales. Les verbes sont listés sous la forme de leur racine, c'est-à-dire sans la terminaison ; par exemple, le lexique contient *distribu,* qui est la racine de *distribuer.* La reconnaissance des flexions relève du module morphologique. À titre d'exemple, nous avons au lexique des entrées comme les suivantes (nous n'expliquerons que les traits pertinents à la discussion) :

(6) bonbon : [cat:N, fém:-, concret:+, génér:Nourr, scat:<SP[formep:à, cpl:N2[génér:Nourr, déf:+]]>]
(7) distribu : [cat:V, gr:6, génér:Transf, cohésion:+, orientation:Fin, répétition:+]
(8) enfant : [cat:N, hum:+, génér:Humain]

Le trait **scat**, qui indique la sous-catégorisation (ou les compléments) d'un item lexical, mérite une explication. La sous-catégorisation des verbes est dérivée par le module sémantique à partir de traits sémantiques dont *génér*, qui fournit le générique sémantique d'un mot (ou encore son hyperonyme). Ici, *distribu* est un verbe dont le générique est *Transf*, c'est-à-dire qu'il est un verbe de transfert. Cette propriété a des conséquences sur sa sous-catégorisation. Entre autres, comme il s'agit d'un transfert (de possession), le verbe aura un complément "bénéficiaire" ; voir la section 3.3 sur la sémantique. On reconnaît aussi des compléments à d'autres items lexicaux, comme les noms. Ceux-ci, idiosyncratiques pour la plupart, sont notés au lexique. C'est le cas du nom *bonbon* ci-dessus, qui a comme *génér*ique *Nourr*iture, et qui prend un SP complément en *à* également de type générique *Nourr*.

Le module du lexique comporte aussi des règles de défaut et de redondance. Par exemple, la règle de défaut en (9) donne des valeurs par défaut pour divers traits dont le nombre (singulier par défaut). Elle s'applique à l'entrée lexicale pour *enfant*, et complète sa liste attribut-valeur, ce qui est illustré en (10). Elle ne change pas le trait **[hum:+]**, puisque les règles de défaut ne modifient pas les traits déjà spécifiés.

(9) Si [cat:N] alors [propre:-, plu:-, coll:-, masse:-, hum:-, lié:-, scat<>]
(10) enfant : [cat:N, hum:+, génér:Humain]
==>
enfant : [cat:N, hum:+, génér:Humain, propre:-, plu:-, coll:-, masse:-, lié:-, scat<>]

Ces règles de défaut (et les règles de redondance) permettent de simplifier les entrées lexicales, qui ne doivent alors contenir que les traits qui ne sont pas prévisibles par les règles de redondance et de défaut.

4.2 La syntaxe

Le module syntaxique s'inspire de la théorie GPSG (Generalized Phrase Structure Grammar, Gazdar, Klein, Pullum et Sag 1985) avec quelques simplifications. À la base se trouvent les règles de dominance immédiate (règles-DI), qui décrivent les constituants. Sur ces règles-DI s'appliquent les règles de préséance linéaire (règles-PL) qui régissent l'ordre des constituants-soeurs dans un arbre local[5]. Nous avons modifié (par rapport à Gazdar et al. 1985) les règles-DI qui ont trait à la sous-catégorisation.

Comme exemple, voici deux règles qui construisent le V1 (V-barre) ; le premier inclut un syntagme prépositionnel, et le deuxième n'en projette pas (NB : X2 est le sujet).

(11)V1 --> V0[tête:+, scat: <X2[rôle:Agent], N2[rôle:Patient],P2>], N2, P2;
<V0 scat N2> = N2,
<V0 scat P2> = P2,
<V1 scat> = <X2>.

(12)V1 --> V0[scat: <X2[rôle:Agent], N2[rôle:Patient],P2[opt:+]>], N2;
<V0 scat N2> = N2,
<V1 scat> = <X2>.

Par ces règles est exprimée la structure du syntagme verbal : pour les deux règles par exemple, un verbe suivi d'un syntagme nominal et, pour la première règle, d'un syntagme prépositionnel. Les règles de réécriture sont suivies d'un ensemble d'équations d'unification qui comparent et assignent la valeur de certains traits pertinents (ici, le trait **scat**). Il est aussi possible d'utiliser des ensembles de traits (voir les règles de combinaison morphologique) ; ceux-ci permettent une certaine concision dans l'écriture des règles, en permettant d'exprimer par une seule équation l'unification de ces ensembles.

---

[5] Nous ne nous servons pas de métarègles ; elles ne sont pas nécessaires, étant donné la nature des phénomènes étudiés dans le prototype.

Ces règles permettent les deux structures pour le SV que l'on trouve dans nos phrases exemples (illustrées à la figure 1). Les restrictions nécessaires pour les désambiguïser, rappelons-le, sont imposées par le lexique, et les modules prosodique et sémantique.

La gestion des traits dans les structures syntaxiques est faite à l'aide des trois ensembles de règles suivants : règles de redondance, règles de défaut et règles de propagation. Ces règles s'apparentent à certaines règles de Gazdar et al. (1985) ; les règles de redondance et les règles de défaut implantent les restrictions de cooccurrence de traits (*Feature Co-occurrence Restrictions*, *FCR*) et les spécifications de traits par défaut (*Feature Specification Defaults*, *FSD*) respectivement. Les règles de propagation sont utilisées pour simuler les principes d'instanciation universels comme le principe de traits de tête[6]. Il est formulé ainsi :

(13) si X --> Y[tête : +] alors <X TraitsTête> = <Y TraitsTête>

4.3 La sémantique

Le module sémantique s'occupe de deux types de phénomènes qui ont trait aux relations sémantiques entre les syntagmes : la sous-catégorisation et les restrictions de sélection. Dans notre analyse, la sous-catégorisation d'un verbe est générée à partir de règles de redondance dont l'application dépend de la présence d'un certain trait sémantique dans l'entrée lexicale du verbe. Par exemple, la règle de sous-catégorisation suivante ajoute un objet indirect à la liste SCAT d'un verbe de transfert comme *distribuer* :

(14) si [génér:Transf] alors [scat:<SP[formep:à, hum:+]>]

La règle transforme l'item lexical de la façon suivante :

---

[6] En fait, c'est le seul principe d'instanciation de GPSG qui nous sert ; l'accord est assuré par des équations d'unification spécifiques tel qu'indiqué ci-dessus, et le principe de traits de pied n'a pas été implanté, compte tenu du corpus choisi.

(15) distribu : [cat:V, gr:6, génér:Transf, cohésion:+, orientation:Fin, répétition:+]
==>
distribu : [cat:V, gr:6, génér:Transf, cohésion:+, orientation:Fin, répétition:+, scat:<SP[formep:à, hum:+]>]

Le sujet, l'objet direct et les autres compléments sont ajoutés par d'autres règles.

Dans le cas des restrictions de sélection, il s'agit d'une mise en correspondance entre plusieurs traits du verbe et certains traits du syntagme auquel ces restrictions s'appliquent. Dans la règle en (14), le trait **scat** du verbe exige que le SN inclus dans l'objet indirect dénote une entité humaine. Le système d'analyse vérifie, par unification, si le syntagme en question contient réellement ce trait ; sinon, l'analyse est rejetée. Comme nous l'avons déjà mentionné, cette restriction élimine l'ambiguïté de la phrase *Paul distribue les bonbons au miel* car *miel* porte le trait **hum:-** et ne peut donc pas être le complément de *distribue* ; il ne peut être que celui de *bonbons*.

Les noms et les morphèmes jouent également un rôle dans la détermination des relations sémantiques entre les syntagmes. Par exemple, le mot *bonbon* sous-catégorise un SP de type nourriture **[génér:Nourr]**, un trait qui apparaît dans la SAV de *miel* (17), mais pas dans celle du mot *enfant* (18) comme en témoignent les entrées lexicales suivantes[7] :

(16) bonbon: [cat:N, fém:-, concret:+, génér:Nourr, scat:<SP[formep:à, cpl:N2[génér:Nourr, déf:+]]>]
(17) miel: [cat:N, fém:-, masse:+, génér:Nourr, scat:<SP[formep:de cpl:N0[génér:Vég]]>]
(18) enfant: [cat:N, hum:+, génér:Humain, propre:-, plu:-, coll:-, masse:-, lié:-, scat<>]

[7] Le trait **hum :+** est nécessaire, en plus du trait **génér:Humain** car plusieurs valeurs de **génér**, différentes de **Humain**, désignent néanmoins des humains. Si notre système avait été associé à un "réseau" sémantique hiérarchique, où les sous-classes des **Humain**s héritaient de cette propriété, il aurait été possible de se débarrasser du trait **hum :+**.

La sémantique joue ainsi un rôle actif plutôt que seulement interprétatif. Les décisions que prennent les autres modules peuvent dépendre de la présence d'un trait sémantique particulier.

4.4 La prosodie

Le module prosodique contribue à la résolution des ambiguïtés syntaxiques reliées au rattachement des syntagmes prépositionnels[8]. L'expert prosodique identifie les particularités des regroupements prosodiques qui permettent de lever certaines ambiguïtés de ce type. Les regroupements prosodiques représentent en quelque sorte la syntaxe de l'oral, mais ils ne sont pas nécessairement l'équivalent des constituants syntaxiques, comme on peut voir en (19), où RMi = regroupement prosodique mineur.

(19) $[\text{Paul}]_{SN}$ [distribue les bonbons]$_{SV}$ *constituants syntaxiques*
$\{\text{Paul distribue}\}_{RMi}$ $\{\text{les bonbons}\}_{RMi}$ *regroupements prosodiques*

Pour nos phrases exemples, la prosodie désambiguïse la structure, puisqu'on y trouve des regroupements prosodiques différents.

(20) $\{\text{Paul distribue}\}_{RMi}$ $\{\text{des bonbons}\}_{RMi}$ $\{\text{aux enfants}\}_{RMi}$
$\{\text{Paul distribue}\}_{RMi}$ $\{\text{des bonbons au miel}\}_{RMi}$

Les règles prosodiques ont la forme suivante :

(21) si RMi --> V det N
alors V det N est inclus dans X2, où X2 est le premier noeud dominant immédiatement les trois catégories

(22) si RMi_z --> det N ET RMi_y --> Prep N
alors RMi_z et RMi_y ne sont pas inclus dans X2, où X2 est le premier noeud dominant immédiatement les deux regroupements

---

[8] Notons que nous avons identifié les contraintes permettant à la prosodie de désambiguïser les structures, mais ces règles n'ont pas été écrites dans le formalisme des SAV et n'ont pas été intégrées au système.

Pour les deux exemples précédents, ces règles nous indiquent que *bonbons* et *aux enfants* se retrouvent dans deux regroupements prosodiques distincts et par conséquent, que *aux enfants* ne peut être le complément du nom *bonbons*. Par contre, la prosodie permettra le rattachement de *au miel* au nom *bonbons*.

4.5 La morphologie

L'expert morphologique traite la formation des mots. On traite deux sortes de morphèmes : les racines et les suffixes. La racine peut être un mot en soi (une forme libre ou non-liée), ou elle peut être une forme qui nécessite un suffixe (une forme liée).

(23)	enfant racine libre	+	s suffixe	==>	enfants
(24)	distribu racine liée	+	tion suffixe	==>	distribution

Lorsqu'un mot est formé de deux morphèmes ou plus, seule la racine est donnée au lexique et le module morphologique s'applique afin de 1) découper le mot en morphèmes séparés, 2) identifier les suffixes à l'aide de la liste de morphèmes inclus dans le module et 3) construire la structure arborescente du mot.

Par exemple, on construit le mot pluriel *bonbons* en utilisant la définition du morphème *-s* (25)[9] et la règle (26) de combinaison de morphologique pour les flexions (pluriel, genre, temps, mode, etc.) :

(25) ~s:[suff:+, plu:+, SCAT:<[cat:N, plu:-]>].

(26) M[lié:-] --> X[tête] # Y[suff:+, dériv:-]
<<X>> = <Y scat>,

[9] On spécifie aussi, avec le suffixe, la modification éventuelle de la racine. Par exemple pour le suffixe pluriel des mots en *-al*, la définition du suffixe contient l'information que *-al* doit être remplacé par *-aux*. Dans le cas du suffixe *-s* toutefois, la racine reste inchangée.

<M INFL> = <Y INFL>.
(M, X et Y sont des variables)

La règle a la même forme que les règles de combinaison syntaxique, sauf l'opérateur de concaténation #. L'ensemble INFL utilisé par cette règle comprend les traits flexionnels et a la définition suivante :

(27) ENS INFL={temps, mode, pers, plu, fém, sep}

Ainsi, par la seule équation <M INFL> = <Y INFL>, tous ces traits du suffixe sont transmis à la mère (au mot). À partir de la racine *bonbon*, du suffixe *-s* et de la règle de combinaison, le système peut produire la structure suivante :

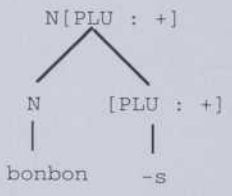


Figure 2 : Structure morphologique

4.6 Le rôle des modules dans l'analyse

Dans notre modèle, nous avons cherché à séparer les tâches afin que les modules, et même les règles qui constituent les modules, puissent agir indépendamment les uns des autres. Ainsi, les modules travaillent séparément à construire l'analyse d'une phrase ; par exemple, pour construire le syntagme verbal, le lexique donne le SAV du verbe, la sémantique en dérive la liste argumentale à partir des traits sémantiques, la syntaxe construit la structure arborescente et propage les traits pertinents, la prosodie vérifie que les structures syntaxiques respectent les restrictions des regroupements prosodiques et la morphologie s'occupe des flexions et de la structure des mots dérivés. Cette indépendance mène à optimiser la parallélisation du prototype linguistique, parce qu'elle permet aux règles des modules de s'appliquer dans n'importe quel ordre, pour autant

que les contraintes informatiques le permettent. En réalité, l'ordre d'application est restreint uniquement par l'existence d'un objet auquel une règle particulière peut s'appliquer, concept que nous développons dans la section suivante.

5 Le prototype

Nous décrivons maintenant le prototype, son architecture et le déroulement d'une analyse.

5.1 Types de règles

L'analyse est centrée autour de la notion d'objet linguistique. Au départ, les objets disponibles aux processeurs du système sont les mots de la phrase à analyser. À l'aide des règles, le système cherche à construire à partir de ces objets des objets de plus en plus grands, éventuellement un objet correspondant à une phrase. À ce moment l'analyse sera réussie. Lorsqu'un objet est créé, une copie est fournie à chacun des processeurs ; cela leur permet de travailler en parallèle sur les données.

En fait, il existe plusieurs types d'objets linguistiques dans le système :

- les mots (MOT), simples suites de lettres comme *bonbons* ou *distribue*.
- les morphèmes (MORPH), des parties de mots associées à une SAV, par exemple :

(28) -s : [suff:+, plu:+, SCAT:<[cat:N, plu:-]>]

- les entrées lexicales (LEX), suites de lettres associées à une SAV, par exemple :

(29) bonbon : [cat:N, fém:- , concret:+, génér:Nourr, scat:
<SP[formep:à, cpl:N2[génér:Nourr, déf:+]]>]

- les mots ou constituants (ITEM) sans les informations prédites par la sémantique, par exemple :

(30) V0[génér:Transf]

- les mots ou constituants (OBJ-SEM) avec les informations ajoutées par la sémantique :

(31)V0[génér:Transf, scat: <X2[rôle:Agent], N2[rôle:Patient], SP[formep:à, hum:+]>]

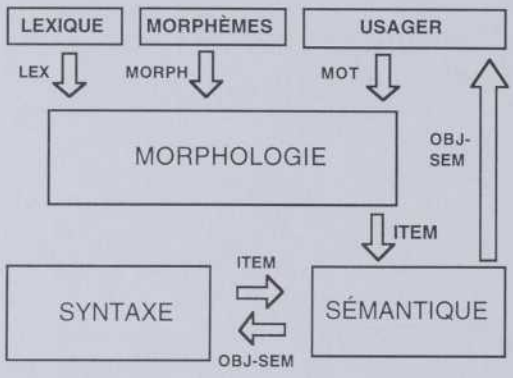


Figure 3 : Schéma des types d'objets

Chaque type d'objet est pertinent pour un type de règle : par exemple, la sémantique ne peut s'appliquer que sur des ITEMs ; la syntaxe, que sur des OBJ-SEMs, etc., comme il est illustré à la figure 3. Cela fait ressortir les dépendances intrinsèques au système linguistique.

Voici comment les divers types de règles manipulent et créent les objets :

- les entrées lexicales et les morphèmes servent à associer à une suite de lettres un ensemble de traits-valeurs (une SAV). Ils introduisent donc les objets linguistiques initiaux dans l'analyse.
- les règles de combinaison morphologiques ou syntaxiques combinent des objets, ce qui veut dire qu'elles créent de nouveaux objets à partir de plus petits constituants (morphèmes, mots ou syntagmes, selon le cas) ; dès lors, ces nouveaux objets sont disponibles au système.
- les règles de redondance spécifient les traits prévisibles à partir d'autres traits. Leur implémentation : en théorie,

elles devraient ajouter, au fur et à mesure que les traits sont identifiés, d'autres traits aux SAV. En pratique, toutes les règles de redondance sont appliquées avant de créer un nouvel objet.

- les règles de défaut spécifient des valeurs à attribuer aux traits qui n'ont pas encore reçu de valeur. Les deux types de règles de défaut de notre système sont mises à contribution à des moments différents, ou plutôt dans des contextes différents. Les règles de défauts statiques sont utilisées lors de la compilation de la grammaire, pour compléter les descriptions. Ces valeurs par défaut sont donc ajoutées systématiquement à toutes les entrées lexicales qui ne sont pas spécifiées autrement. Les règles de défauts dynamiques, elles, sont appliquées lors de l'analyse d'une phrase ; elles servent ainsi à fournir des valeurs par défaut pour des propriétés non spécifiées par les items lexicaux précis. Cela permet de capter des valeurs par défaut qui ne sont pas lexicales (c'est-à-dire qui ne peuvent être spécifiées sur des items lexicaux) mais bien syntaxiques (c'est-à-dire qui doivent être déterminées en fonction du contexte syntaxique). L'accord (verbal) est un exemple d'un défaut dynamique : il se fait à la troisième personne du singulier par défaut. Il faut attribuer ces valeurs par défaut en dernier recours. Dans notre système, ce sera au moment de la création d'un nouvel objet. Après avoir appliqué les règles de redondance, les SAV seront complétées par les règles de défaut.
- les règles de préséance linéaire sont appliquées à la compilation de la grammaire et ordonnent les constituants dans les règles syntaxiques. En d'autres termes, les règles utilisées lors de l'analyse sont des règles de réécriture dont les filles sont ordonnées, et non des règles-DI.
- les règles de propagation servent à propager des traits dans la structure syntaxique. Lorsqu'une règle syntaxique est utilisée, c'est-à-dire quand un nouvel objet est reconnu, certains de ses traits (par exemple ses traits de tête) sont calculés à partir de traits correspondant fournis par les sous-constituants de la règle (par exemple la tête).

5.2 Compilation de la grammaire

Notre système effectue des transformations successives sur les règles linguistiques qui, tout en conservant les informations linguistiques, effectuent des opérations de plus en plus éloignées du modèle de départ. La correspondance est effectuée par les étapes suivantes :

1. traduction en règles de production
2. compilation des règles de production en un réseau de noeuds
3. implantation parallèle d'un système d'acteurs

Un aperçu du système est donné à la figure 4 (où seuls trois modules sont représentés).

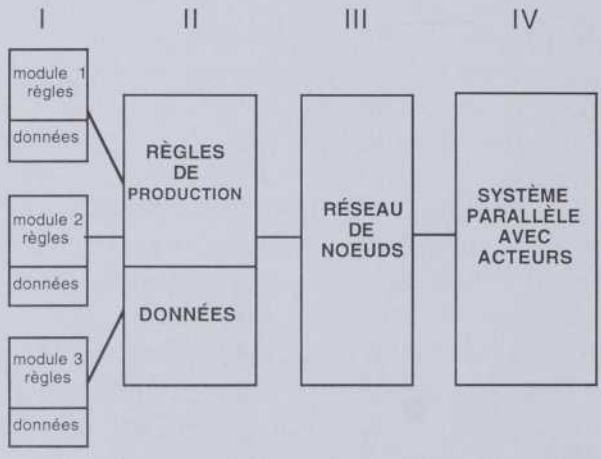


Figure 4 : Aperçu du système

5.2.1 Traduction en règles de production

La première transformation a pour but d'exprimer les règles linguistiques en règles de production. Un système de règles de production est un ensemble de règles décrivant des situations, et pour chacune de ces situations, une série d'actions à effectuer. Il n'y a pas d'ordre prédéfini entre les règles. Un interpréteur de règles regarde les données et détermine les

règles à exécuter. L'aspect descriptif de ce formalisme est proche du formalisme utilisé en linguistique et il a été possible de traduire le formalisme linguistique en règles de production. Du point de vue du parallélisme, ceci représentait l'étape la plus importante.

Dans notre prototype, les règles linguistiques sont réparties dans différentes sections de la "grammaire" : il y a une section LEXIQUE, une section SYNTAXE, une section MORPHOLOGIE, une section SEMANTIQUE, etc. Chaque section contient les règles qui lui sont propres, y compris toutes les règles de redondance, de défaut ou autres. Le processus de traduction des règles linguistiques tient compte de toutes ces spécifications locales.

Une règle de production est formée de deux parties : les conditions et les actions. On peut voir une règle-DI, par exemple celle en (32), comme un ensemble de conditions et une action. Les conditions correspondent à l'identification dans le système des constituants formant le corps de la règle et l'action à la création de l'objet correspondant à la mère. On a la même correspondance pour les règles de combinaison morphologique : un mot est reconnu (créé comme objet dans le système) si ses morphèmes (c'est-à-dire les conditions) sont identifiés.

```
(32)V1 -->  V0 [tête:+, scat:<X2[rôle:Agent],N2[rôle:Patient],P2>],
            N2[rôle:Patient],
            P2[formep:à];
                <V0 scat N2 TraitsScat> = <N2 TraitsScat>,
                < V0 scat P2 TraitsScat> = <P2 TraitsScat>,
                < V1 scat> = <<X2>>.
```

Voici une description informelle d'une règle de production correspondant à une des règles linguistiques : si le système détecte un ou plusieurs objets linguistiques contigus et que chacun respecte les contraintes de compatibilité décrites par une règle linguistique, alors on fabrique un nouvel objet et on y ajoute les informations décrites par les règles de redondance et de défaut. En particulier, les conditions d'une règle de production sont formées du corps de la règle de combinaison (les constituants), des équations d'unification (elles doivent être vérifiées pour que la règle s'applique) et de certains défauts (les

défauts statiques), ainsi que d'autres règles particulières aux modules. Elles ont été déterminées lors de la compilation de la grammaire. Les actions d'une règle de production consistent en la création de l'objet linguistique correspondant à la mère de la règle linguistique, qui doit être soumis à toutes les règles de redondance et les règles de défaut (dynamiques) pertinentes ; ces règles sont appliquées à ce moment, et dans cet ordre. Ainsi, l'objet est complet puisque toutes les règles de redondance pertinentes se sont appliquées, et les défauts sont appliqués "au moment opportun", c'est-à-dire en dernier lieu. Ensuite l'objet est introduit dans le système et est rendu disponible à tous les processeurs. Les mots du lexique et la liste des morphèmes, eux, ne sont pas des règles mais des éléments de mémoire. Ils sont disponibles à tous les processus à tout moment.

Par exemple, étant donné une règle quelconque, la traduction se fait comme l'indique le schéma à la figure 5, où les C_i correspondent à des conditions, et les M_i à des éléments de mémoire, c'est-à-dire les données de l'analyse en cours.

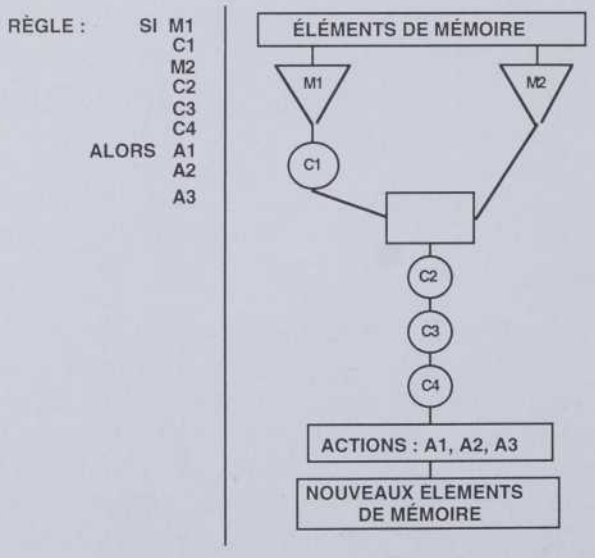


Figure 5 : Correspondance entre une règle de production et une partie du réseau

Lorsque chacune des règles linguistiques a été traduite en règle de production, celles-ci sont soumises à la prochaine transformation.

5.2.2 Compilation des règles de production en un réseau de noeuds

Les techniques mises au point pour programmer des interpréteurs de règles de production sont basées sur la construction d'un graphe où chaque noeud représente une précondition dans une ou plusieurs règles. Le schéma à la figure 5 n'illustre qu'une partie du réseau de noeuds qui doit être construit pour représenter toutes les règles de grammaire. L'algorithme que nous avons utilisé pour construire le réseau complet porte le nom de TREAT (Miranker 1987). Une fois que celui-ci est créé, on peut l'utiliser pour analyser des données, c'est-à-dire des phrases. Les données sont alors insérées aux points d'entrée du graphe, puis des jetons contenant des résultats partiels sont propagés entre les noeuds. Quand un jeton est reçu par un noeud terminal, une situation d'application d'une règle est reconnue.

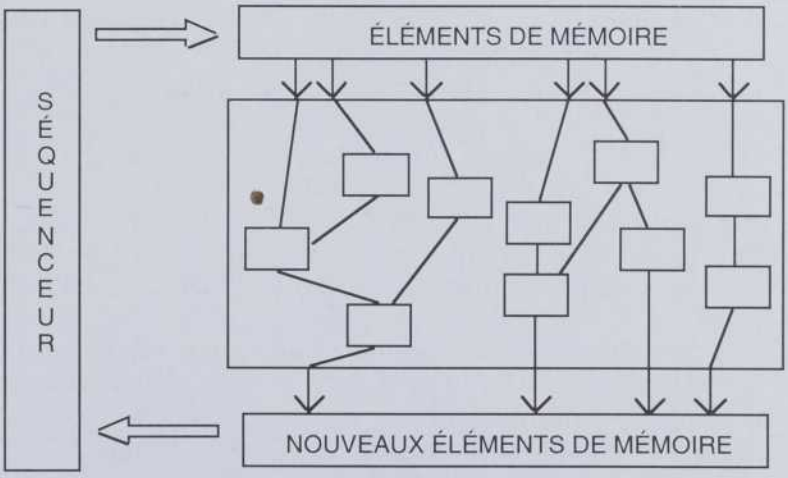


Figure 6 : Unité de travail

L'application de la règle, rappelons-le, correspond à la création d'un nouvel objet, selon les spécifications contenues dans la règle. Le schéma à la figure 6 donne une idée du système. Le séquenceur contrôle les actions : il lance les éléments de mémoire initiaux qui parcourent le graphe, et il récupère les nouveaux éléments pour les fournir à nouveau au réseau. L'analyse procède donc par une série de cycles gérés par le séquenceur.

La phase finale de transformation consiste à implanter le réseau de noeuds ainsi construit sur chacun des processeurs de la machine parallèle.

5.2.3 Implantation parallèle

Notre implantation d'un système d'acteurs fonctionnant sur ordinateur VOLVOX s'est faite selon le système CLAP (Desbiens et al. 1993). Chaque processeur représente un acteur et contient l'équivalent de la totalité des règles et des informations connues. L'interconnexion entre les acteurs se fait selon une organisation "maître-esclave", où l'acteur-maître est responsable de l'entrée-sortie, de l'envoi et de la réception des nouvelles informations et de l'assignation des tâches aux esclaves. Plus précisément, chaque nouvelle information, obtenue par quelqu'acteur que ce soit, est envoyée à tous les acteurs du système et le séquenceur associe un numéro de séquence à ces informations. Les acteurs ont la tâche d'accomplir des *partitions de travail*. Une partition de travail est définie de la façon suivante : tout ce qu'il est possible de dériver à partir d'une information et, de façon optionnelle, d'informations antérieures (ayant un numéro de séquence moindre). Pour chaque information ayant un numéro de séquence n, il existe une seule partition de travail et aucun élément de travail n'est commun à deux partitions. Aussitôt qu'un acteur a terminé sa partition de travail, il peut en commencer une autre. La synchronisation des acteurs entre eux ne provoque pas de perte de temps.

La tâche est terminée lorsqu'aucun acteur ne peut ajouter d'informations à celles qui existent déjà.

6 Conclusion

Le prototype de traitement linguistique parallèle que nous avons développé a servi à tester nos hypothèses quant à la possibilité de lever les ambiguïtés syntaxiques en mettant à contribution simultanément des experts linguistiques divers. Il nous a été nécessaire pour accomplir cela de revoir entièrement la tâche à travers les "lunettes" du parallélisme ; cela nous a amenés à identifier et exploiter le potentiel de parallélisme dans la description et la résolution du problème.

L'élaboration d'un prototype de traitement linguistique parallèle nous a permis aussi de prendre conscience d'une part de la modularité relative des règles linguistiques et d'autre part des contraintes inhérentes à la conception d'une solution parallèle à un problème.

En effet, s'il est vrai que certaines règles appartiennent véritablement à un niveau d'analyse linguistique plutôt qu'à un autre, d'autres sont à la frontière de deux niveaux. Ainsi les règles de combinaison syntaxique sont bien syntaxiques, mais celles traitant de la morpho-syntaxe pourraient souvent être exprimées dans l'un ou l'autre de ces niveaux. Cette constatation explique la division parfois arbitraire des expertises dans notre prototype ; certains phénomènes ont été traités en sémantique alors qu'ils auraient pu être ajoutés au lexique, et vice versa. En outre, une partie du travail est parfois dupliqué, comme l'attribution du trait **plu:+** à *des bonbons* par la morphologie, la sémantique et la syntaxe à la fois dans l'analyse de la phrase *Paul distribue des bonbons aux enfants*. Cela soulève la question de l'existence de frontières absolues entre les niveaux d'analyse linguistique ; nous n'avons pas eu à y répondre, car nous avons préféré opter pour la redondance : une règle justifiée par plus d'un module apparaît (souvent) dans chacun d'entre eux. D'ailleurs, cette redondance est sans doute inhérente au système linguistique.

Toutefois la représentation choisie a rendu aisé l'échange d'informations entre les niveaux d'analyse ; le contraire aurait été pour le moins surprenant, même si les traditions dans

chacun des domaines de la linguistique empruntent des formalismes très divers.

La mise en œuvre du prototype a exigé à maintes reprises de porter une attention particulière aux dépendances chronologiques ou conceptuelles présentes dans la solution. Les dépendances chronologiques ont pu être réduites au minimum, croyons-nous, grâce au découpage très fin qui a été fait des informations linguistiques : on fait appel à des règles très locales, indépendantes les unes des autres. Il faut dire, ou plutôt répéter, que les formalismes linguistiques sont déjà hautement modulaires et se prêtent bien à un traitement parallèle.

Nous concluons cet article en faisant deux remarques qui ont trait aux dépendances conceptuelles dans le prototype que nous avons élaboré. L'une porte sur la notion d'objet linguistique que nous avons développé et l'autre sur le parallélisme et la modularité dans le prototype.

Les implications du concept d'objet linguistique que nous avons été amenés à développer dans le prototype sont complexes et amènent des incompatibilités avec les méthodes linguistiques habituelles. C'était le cas pour les objets créés par les règles sémantiques et syntaxiques. Nous avons dû en effet adopter une solution distinguant deux types d'objets "syntaxiques" : certains contenant l'information fournie par la sémantique (OBJ-SEM), et d'autres sans cette information (ITEMs). Cette distinction était nécessaire pour éviter que certaines inférences syntaxiques soient tirées sur des objets en quelque sorte incomplets, c'est-à-dire dont la sous-catégorisation n'avait pas été fournie par les règles sémantiques. Par exemple, lorsqu'un verbe comme *distribue* est reconnu, il est impératif que la sémantique dérive d'abord sa sous-catégorisation correcte ; sinon, son trait **scat**, non instancié, pourrait être unifié avec le trait **scat** d'une règle syntaxique quelconque. La production de cet objet erroné entraînerait toute une suite d'opérations inutiles, potentiellement très coûteuses en temps et en mémoire, qui se solderaient par un échec. Pour court-circuiter ce problème à la source et coller davantage à la description linguistique (c'est-à-dire ne considérer comme com-

pléments de *distribue* que les objets compatibles avec sa sous-catégorisation), nous avons choisi de créer ces deux types d'objets. Or en linguistique ces derniers n'ont pas de correspondants. Cette solution est d'ailleurs incomplète envers le traitement sémantique : l'application d'une règle sémantique à un objet de type ITEM crée un objet de type OBJ-SEM, qui ne peut plus être traité par la sémantique. Pourtant l'information sémantique pertinente pour un objet linguistique peut être calculée à partir de plusieurs règles sémantiques ; il peut y avoir plusieurs compléments prédits par divers traits sémantiques. Ainsi le concept d'objet linguistique semble essentiel à l'implantation parallèle d'un système linguistique mais nécessite un examen plus approfondi.

Pour ce qui est des propriétés de notre prototype, en réalité, son parallélisme et sa modularité ne sont pas complets dans la vision naïve que nous en avions au départ.

D'abord, la modularité linguistique n'est pas maintenue intégralement dans l'architecture de notre système informatique. Chaque module linguistique est traduit en un ensemble de règles de production qui sont regroupées selon les besoins du réseau d'après des contraintes de gestion informatique. Un "bloc" ainsi créé peut contenir des règles relevant soit d'un seul module, soit de divers modules linguistiques. Les modules représentent donc une abstraction linguistique des règles finales du système.

Enfin, il reste de la séquentialité apparente :

1) les règles de redondance doivent être appliquées avant les règles de défaut.
2) la formation d'unités d'analyse plus simples doit précéder celle d'unités plus complexes, spécialement à l'intérieur d'un module.
3) en cours d'analyse, divers types d'objets sont créés : d'abord des listes de lettres, puis des catégories (ensembles de paires d'attributs-valeurs), puis des "objets sémantiques" (avec la grille de sous-catégorisation instanciée), etc. Chaque expert s'applique

sur un type particulier d'objet ; tant que ce type d'objet n'existe pas, l'expert ne contribue pas au système.

Il ne s'agit pas, dans ces cas, de séquentialité véritable, mais de pertinence d'une information : dès qu'une information devient pertinente pour une règle, et seulement à ce moment-là, celle-ci est appliquée. En d'autres termes, certaines dépendances logiques, intrinsèques à l'analyse linguistique, ont dû être respectées.

Bibliographie

Bouchard, D., L. Bouchard, H. Cedergren, L. Da Sylva, A.-M. Di Sciullo, A. Dugas, L. Emirkanian, B. Klipple, F. Léveillé, H. Perreault, C. Robitaille et J. van Voorst, 1993, *Analyse linguistique parallèle*, Rapport final, Projet ALEX-Linguistique, Centre ATO.CI et département de linguistique, UQAM.

Church, K. et R. Patil, 1982, Coping with Syntactic Ambiguity, or How to Put the Block on the Box on the Table, *American Journal of Computational Linguistics*, Vol. 8.

Desbiens, J., M. Lavoie, S. Pouzyreff, P. Raymond, T. Tamazouzt et M. Toulouse, 1993, CLAP: Un système de programmation objet pour machines parallèles à mémoire distribuée, *ICO Québec*, Automne 1993, pp. 27-37.

Erman, L.D., F. Hayes-Roth, V.R. Lesser et D.R. Reddy, 1980, The HEARSAY-II Speech Understanding System : Integrating Knowledge to Resolve Uncertainty, *ACM Computing Surveys,* 12.

Gazdar, G., E. Klein, G. Pullum et I. Sag, 1985, *Generalized Phrase Structure Grammar*, Oxford, Basil Blackwell.

Miranker D. P., 1987, Treat : A Better Match Algorithm for AI Production Systems, *National Conference on Artificial Intelligence*, AAAI-1987.

Shieber, S., 1986, *An Introduction to Unification-Based Approaches to Grammar*, CSLI Lecture Notes, Chicago University Press.

UN PROTOCOLE POUR LA MISE AU POINT D'ALGORITHMES DE DÉSAMBIGUÏSATION CATÉGORIELLE

François Daoust
Fernande Dupuis

Résumé

Nous présentons dans cet article un aspect du travail réalisé dans le cadre de la commandite de recherche de la compagnie Alex Informatique. Après une présentation générale du projet ALEXATO (traitement parallèle et analyse de texte par ordinateur), nous allons discuter de façon plus élaborée des hypothèses qui sont à la base d'un protocole expérimental pour la désambiguïsation catégorielle. Ce protocole vise à fournir un cadre comparatif pour évaluer la performance d'une variété de modèles de désambiguïsation. Il s'appuie sur la disponibilité de corpus de grande taille et veut faire l'économie d'une désambiguïsation manuelle préalable.

1 Présentation générale du projet ALEXATO

Le projet ALEXATO[1] tire son nom de la contraction de deux particules : «ALEX» désigne évidemment le projet Alex et

[1] L'équipe du projet ALEXATO est composée comme suit:

«ATO» (ATO) fait référence au logiciel SATO[2] et à l'objectif de sa parallélisation.

SATO, le système d'analyse de texte par ordinateur, a été conçu pour fournir un soutien informatique au travail d'analyse de texte. C'est un outil général basé sur une représentation bidimentionnelle du texte. La première dimension rend compte de la dimension lexicale du texte alors que la deuxième dimension rend compte de la dimension séquentielle du texte, c'est-à-dire du déploiement ordonné des items lexicaux. Souvent, on désigne les deux axes de ce plan en termes, respectivement, d'axes paradigmatique et syntagmatique.

Nous souhaitons pouvoir utiliser SATO sur des corpus de grande taille sans augmenter abusivement le temps de calcul. Comme la dimension lexicale d'un corpus tend à plafonner avec l'augmentation de la taille du corpus, la contrainte de complexité, en termes d'effort de calcul, porte davantage sur l'axe syntagmatique dont la croissance est linéaire. Par ailleurs, en raison même de sa dimension séquentielle, l'axe syntagmatique se prête naturellement à une opération de partition en segments ordonnés. Il est donc aussi naturel de concevoir les opérations de calcul sur ces segments de façon parallèle en faisant appel à des processeurs autonomes. Ce traitement parallèle exige cependant une synthèse des résultats qui soit en mesure de restaurer la globalité du corpus. Le premier objectif

---

- Jules Duchastel, responsable du projet, Département de sociologie UQAM;
- François Daoust, chef de projet, Centre ATO-CI, UQAM;
- Josiane Ayoub, Département de philosophie, UQAM;
- Suzanne Bertrand-Gastaldy, EBSI, Université de Montréal;
- Gilles Bourque, Département de sociologie, UQAM;
- Christophe Dansac, stagiaire, Université Pierre Mendes-France, Grenoble;
- Fernande Dupuis, Département de linguistique, UQAM;
- Normand Laurier, analyste de l'informatique, Centre ATO-CI, UQAM;
- Monique Lemieux, Département de linguistique, UQAM;
- Louis-Claude Paquin, Département des communications, UQAM.

[2] Le logiciel SATO est un outil général d'analyse de texte (voir Daoust 1992).

du projet ALEXATO était donc d'explorer la faisabilité de ce modèle de traitement.

Profitant de l'augmentation escomptée de la puissance de calcul associée au traitement parallèle, nous voulions aussi proposer un cadre expérimental pour la résolution d'un problème se situant clairement sur l'axe syntagmatique, à savoir la levée des ambiguïtés catégorielles par l'analyse des contextes immédiats. Plus précisément, la problématique est la suivante. Le savoir lexical est souvent de nature polycatégorielle, c'est-à-dire qu'un lexème peut avoir plusieurs sens ou attributs que l'on peut tenter de décrire par des systèmes de catégories[3]. Souvent, cette polyvalence catégorielle peut être restreinte par des contraintes de cooccurrence. On peut penser aux contraintes syntaxiques de la langue mais aussi aux contraintes sémantiques. Le deuxième objectif du projet ALEXATO était donc de proposer un modèle expérimental, un protocole, pour travailler sur une certaine classe de problèmes d'ambiguïtés. Ce protocole doit nous permettre de mettre en oeuvre et de valider diverses stratégies de désambiguïsation. En particulier, nous visions à explorer diverses méthodes par apprentissage, de type statistique ou connexionniste.

2 Le modèle informatique de traitement parallèle

Dans cette section, nous décrivons nos hypothèses de parallélisation du logiciel SATO. Cette présentation reste générale en évitant d'entrer dans les détails d'implantation informatique, en évitant surtout de nous étendre sur les contraintes particulières liées à l'utilisation des ordinateurs VOLVOX[4].

---

[3] On distingue «système catégoriel» de la notion plus générale d'annotation. Le système catégoriel fait appel à une suite dénombrable et finie de traits, structurés ou pas, et dont l'application sur des données suit une procédure définie et, espère-t-on, reproductible.

[4] Les ordinateurs parallèles VOLVOX (ARCHIPEL SA) utilisés dans le cadre de ce projet sont constitués d'un groupe de 64 «transputers» TS-800 (INMOS) dont la mise en réseau est basée sur une topologie variable. Le système utilise une architecture à mémoire distribuée, la communication entre «transputers» s'opérant au moyen de messages transitant par des ports sériels. Pour une description plus détaillée, voir *Parallel processing made easy!*, Alex informatique Inc., Montréal.

La version actuelle de SATO fonctionne sur des PC d'IBM ou des compatibles avec le système DOS. Le système SATO est composé de deux programmes distincts : SATOGEN, pour la GÉNération du texte sur le plan lexique-occurrences, et SATOINT pour l'INTerrogation du texte, c'est-à-dire le déploiement des stratégies d'analyse.

SATOGEN permet de construire la représentation informatique du corpus avec les propriétés initiales qui lui sont rattachées. On entend par «propriété» dans SATO un système catégoriel, une variable numérique ou un système d'annotation libre qui s'ajoute ou complète le texte intégral. Le programme agit un peu à la manière d'un compilateur en produisant une représentation du texte qui va permettre une exploitation rapide du lexique, de la chaîne des occurrences et des diverses propriétés lexicales ou contextuelles.

SATOINT est la partie interactive du système qui permet l'exploitation des données textuelles dans un contexte interactif tout en donnant accès à un langage de commandes élaboré. L'utilisateur peut à tout moment, en tout ou en partie, reconstituer les textes originaux et effectuer des analyses statistiques sur les mots et les propriétés ajoutées. Il peut ajouter de nouvelles propriétés (systèmes catégoriels, valeurs numériques ou annotations libres) et procéder automatiquement ou manuellement à la catégorisation des mots en contexte ou hors contexte. Il peut aussi produire des dictionnaires à partir de la catégorisation du lexique du texte ou appliquer des dictionnaires existants sur le lexique du texte.

SATO dispose d'un ensemble de fonctions et d'analyseurs dont la combinaison permet à l'analyste du texte de construire des protocoles d'analyse élaborés. Le système fournit un mécanisme de trace qui enregistre dans un journal les opérations réalisées et permet la reprise des opérations pertinentes sous la forme de scénarios de commande.

2.1 Un modèle de traitement coopératif

Pour réaliser une implantation de SATO qui puisse bénéficier des ordinateurs parallèles, nous avons conçu un modèle de

traitement coopératif faisant appel autant aux processeurs parallèles qu'au poste de travail personnel et au serveur. Ce scénario vise à ne transférer sur les ordinateurs parallèles qu'une partie du code, celle qui requiert le plus de ressources de calcul. Ce scénario prudent vise aussi à implanter le modèle en minimisant les efforts de conversion du code source dans le cas où les algorithmes ne sont pas touchés par l'effort de parallélisation.

- Poste de travail : Dans ce scénario de traitement coopératif, le poste de travail (micro-ordinateur) continue à gérer l'interface personne-machine, la lecture des fichiers ASCII, le formatage et l'impression des résultats, le journal et le décodage des commandes. Le poste de travail est muni d'une interface de communication sérielle ou EtherNet.
- Processus central : Le poste de travail est en communication avec un processus central qui agit comme serveur. En plus de ses tâches de serveur, le processus central a pour rôle de répartir sur l'ordinateur parallèle les tâches de calcul les plus lourdes et dont la parallélisation paraît possible et souhaitable.

2.2 La parallélisation du processus central.

La parallélisation du processus central implique trois types de tâches.

La tâche-lexique consiste à gérer le lexique. Cette tâche pourrait être réalisée par le serveur. On pourrait aussi lui dédier un ou plusieurs processeurs.

La tâche-texte porte sur la gestion des textes dans leur représentation SATO (chaîne des occurrences). Cette tâche augmente linéairement avec le nombre de textes dans le corpus. Elle peut être facilement répartie entre plusieurs processeurs qui vont calculer en parallèle sur des segments différents du corpus.

La tâche de contrôle consiste principalement à répartir le travail et à fusionner les résultats (tâche-fusion). Cette tâche implique une liaison étroite entre le serveur et les processeurs parallèles.

Dans la perspective d'une utilisation pleinement intégrée d'un SATO parallèle, on doit envisager la transformation du SATO mono-poste, tel qu'on le connaît actuellement, vers un système multi-requêtes et multi-utilisateurs. Pour employer un vocabulaire courant dans le domaine informatique, on doit donc distinguer un SATO-CLIENT mono-poste et un SATO-SERVEUR qui agit à la manière d'un gérant du corpus doté d'une grande puissance de calcul et apte à satisfaire plusieurs requêtes sur un même corpus centralisé. La parallélisation de SATO implique donc à la fois sa répartition à travers une structure de type client-serveur et la parallélisation du processus central entre un serveur conventionnel de type UNIX, et des processeurs parallèles.

Une première option pour la parallélisation du processus central consisterait à simplement partitionner le corpus en considérant chaque texte de la partition comme un corpus autonome. Conséquemment, on distribue une copie de SATO à chacun des processeurs qui agit alors sur le texte de façon séquentielle. La tâche fusion consiste à distribuer à chacun des SATO une copie de la requête de l'utilisateur et à synthétiser les résultats de manière à donner l'illusion d'un corpus unifié.

Cette première option a l'avantage de conserver intégralement la structure du programme tout en lui superposant une couche supplémentaire. Elle a aussi l'avantage de nous épargner une nouvelle «satogénération» de l'ensemble du corpus à chaque fois que l'on veut lui ajouter un nouveau texte. D'un autre côté, le travail de fusion devient beaucoup plus complexe en raison même de l'éclatement du corpus et de son lexique unique. Le calcul requis pour l'opération de fusion doit être repris intégralement avec chaque requête. Enfin, les opérations sur le lexique doivent être reprises par chaque processeur plutôt que de n'être effectuées qu'une seule fois sur un lexique fusionné. Le dernier problème que pose cette approche tient à la nécessité de convertir la plus grande partie du code sur les processeurs parallèles.

La deuxième option de parallélisation consiste à conserver l'unicité du corpus, en particulier le lexique global. La répartition sur les processeurs parallèles ne touche alors que l'axe syntagmatique pour lequel la charge de calcul est à croissance

linéaire. Si l'on disposait de processeurs parallèles ayant un accès indépendant à une unité de stockage de masse, on pourrait concevoir les processeurs parallèles comme des «disques intelligents».

L'implantation d'un tel modèle pose cependant le problème de la séparation de composantes logicielles qu'il n'était pas pertinent de distinguer dans le cadre du modèle séquentiel. Dans la mesure aussi où plusieurs opérations d'analyse avec SATO impliquent un échange entre les axes lexical et contextuel, une communication rapide entre processeurs demeure une condition pour profiter du parallélisme.

En fait, l'architecture très particulière des ordinateurs VOLVOX, de même que la faiblesse des outils logiciels disponibles ne remplissaient pas les conditions souhaitées. En effet, les ordinateurs dont nous disposions ont une capacité d'entrée-sortie tellement faible que les gains dans la puissance de calcul n'arrivent pas à compenser les lenteurs reliées au flux de données. La modélisation parallèle demeure néanmoins d'une grande pertinence même si les résultats expérimentaux sur ce type de matériel s'avèrent décevants.

3 Un modèle expérimental pour la désambiguïsation catégorielle

Le deuxième objectif du projet ALEXATO était de profiter d'une puissance accrue de traitement sur l'axe syntagmatique pour mettre en oeuvre diverses stratégies d'analyse contextuelle. En particulier, nous voulions expérimenter des modèles par apprentissage qui exigent souvent une puissance de calcul numérique importante. La disponibilité d'un modèle de réseaux de neurones développé par une autre équipe du projet ALEX a inspiré notre démarche (modèle EIDOS, voir Proulx et Bégin 1993)[5].

[5] Pour l'expérimentation des modèles neuronaux et associatifs, nous collaborons avec l'équipe de Robert Proulx, professeur au département de psychologie à l'UQAM. Manon Marcil, professionnelle de recherche de cette équipe a réalisé l'expérimentation statistique exposée dans et article. «Issu des travaux sur le modèle BSB d'Anderson et al. (1977), le modèle EIDOS se distingue toutefois de ce dernier par l'utilisation

Nous pourrions présenter ainsi la classe de problèmes qui nous intéresse. Nous partons de l'hypothèse de la disponibilité d'un savoir lexical sous forme catégorielle. En règle générale, ce savoir est polycatégoriel, c'est-à-dire qu'un lexème peut avoir plusieurs catégories dont la pertinence va varier selon le contexte d'utilisation. Cette disponibilité est très courante en analyse de texte par ordinateur et un logiciel comme SATO est justement conçu pour rendre explicite et facilement manipulable ce savoir lexical.

Quelques exemples suffiront à l'illustrer cette situation. Lorsque l'on veut indexer des segments textuels, on dispose généralement d'un thesaurus de concepts qui renvoie à un certain nombre d'items lexicaux. Cette relation est rarement biunivoque. Un forme lexicale peut renvoyer à plusieurs termes ou à aucun terme selon le contexte d'utilisation du lexème. En analyse de discours, on retrouve une situation analogue. Prenons comme exemple la grille sociologique Bourque-Duchastel-Beauchemin (1994)[6]. Un lexème comme «école» peut renvoyer à plusieurs catégories de la grille : Éducation, Travaux publics (l'école en tant que bâtiment), etc.

Enfin, on peut citer un problème syntaxique très connu, la catégorisation grammaticale : «ferme», par exemple, peut être un nom, un verbe, un adjectif ou un adverbe dépendant de son contexte d'utilisation. Si certains lexèmes sont polycatégoriels, nous parlerons alors d'ambiguïtés, plusieurs lexèmes sont monocatégoriels : par exemple, «fermait» ne peut être qu'un verbe. C'est ce problème de catégorisation grammaticale que nous allons utiliser pour construire notre protocole expérimental.

Dans beaucoup de cas, c'est l'analyse des contextes où sont employés les lexèmes qui permet de lever l'ambiguïté, c'est-à-

---

d'une nouvelle règle d'apprentissage, de même que par l'introduction de deux paramètres généraux, lesquels permettent de régénérer constamment la dynamique du système». (Voir Proulx et Meunier 1994.) On trouvera une présentation du modèle EIDOS dans Bégin et Proulx (1993) et dans Proulx et Bégin (1990, 1993).

6 Voir Bourque, Duchastel et Beauchemin (1994).

dire de sélectionner, parmi les catégories lexicales, celle ou celles qui s'avèrent pertinentes. En d'autres mots, le savoir lexical est généralement accompagné de contraintes d'utilisation. Les contraintes qui nous intéressent ici sont dites «locales», c'est-à-dire qu'elles concernent le contexte formé par les mots qui sont dans l'entourage immédiat. On se limitera aussi aux contraintes qui sont bâties autour de systèmes catégoriels.

3.1 Présentation du protocole

Nous voulons nous servir des contextes faisant appel à des lexèmes monocatégoriels pour tenter de révéler des associations entre la catégorie d'une position cible et les catégories du contexte immédiat. En d'autres termes, nous nous intéressons à cette classe de problèmes d'ambiguïtés pour lesquels l'analyse des contextes immédiats permet une levée, ne serait-ce que partielle, de l'ambiguïté catégorielle du mot en position cible. Plusieurs modèles peuvent être utilisés afin de trouver, dans les catégories portées par les divers lexèmes du contexte, un mécanisme pour confirmer ou éliminer des catégories du lexème à désambiguïser[7].

Le mécanisme le plus courant est celui de la règle, dont les règles de grammaires sont une bonne illustration. Ainsi, par exemple, on dira : si un lexème, qui ne peut être qu'une préposition, est suivi d'un lexème ambigu entre un nom et un verbe, alors ce dernier ne peut pas être un verbe conjugué. Dans la phrase «la femelle construit son nid sous un tas de branches», le mot «branches» ne peut pas être une forme conjuguée du verbe «brancher».

On peut aussi concevoir des dispositifs statistiques. Dans l'exemple précédent, une chaîne de Markov dont les probabilités auraient été estimées par échantillonnage à partir d'un ensemble de contextes non-ambigus aurait aussi permis de conclure que «branches» n'est pas un verbe. De la même façon,

[7] Pour une présentation des modèles linguistiques parmi les plus connus, voir Fujisaki et al. (1989, 1991), Milne (1988) et Smith (1991).

certains types de réseaux de neurones auraient produit le même effet. Les modèles statistiques, markoviens, ou les modèles à base de réseaux de neurones procèdent par apprentissage. Les probabilités, mesures d'association ou poids neuronaux, sont déterminées suite à l'analyse d'un ensemble de contextes désambiguïsants.

Généralement, l'apprentissage est réalisé sur la base de l'analyse d'un corpus échantillon dont on a levé manuellement toutes les ambiguïtés. La construction de tels corpus est une lourde tâche. De plus nous ne croyons pas qu'il soit nécessaire de lever toutes les ambiguïtés pour être en mesure de révéler des associations pertinentes. Par exemple, dans le protocole d'apprentissage utilisé par l'équipe de Robert Proulx avec le modèle EIDOS, la découverte des prototypes est possible même si l'apprentissage a été opéré à partir d'échantillons bruités.

Le protocole expérimental que nous avons élaboré vise donc à nous fournir les éléments de contrôle nécessaires pour valider une variété de modèles en jouant sur divers paramètres du modèle. Le problème test qui a été choisi pour valider le protocole est celui de la catégorisation grammaticale. Il s'agit d'un problème classique pour lequel il existe aussi des solutions «classiques» à base de règles. Nous avons choisi de nous concentrer sur l'ambiguïté verbale. Ce choix tient à des considérations pratiques, à savoir le besoin de compter le nombre de propositions pour évaluer la complexité d'une phrase. Il tient aussi à des considérations théoriques sur l'importance du verbe dans l'analyse de la phrase.

Le problème peut donc être résumé ainsi. Considérant les catégories grammaticales des lexèmes qui précèdent et qui suivent un lexème pouvant être un verbe, quels modèles peut-on construire pour déterminer la catégorie effective de ce lexème ambigu.

Notre protocole expérimental peut être schématisé de la façon suivante.

1 - Constitution d'un corpus témoin.

Nous disposions déjà de corpus validés et représentatifs. Nous avons utilisé des corpus de textes fournis à des élèves de

diverses classes du primaire et du secondaire. Ce corpus, élaboré dans le cadre du projet SATO-CALIBRAGE mené avec le Ministère de l'Éducation du Québec, a l'avantage de nous fournir plusieurs textes dont le niveau est gradué[8]. Il est donc possible de choisir, dans un premier temps à tout le moins, des textes considérés faciles dans lesquels on est susceptible de trouver les structures syntaxiques les plus fréquentes de la langue.

2 - Mise au point d'un dispositif classique à base de règles.

Le savoir linguistique entourant le problème choisi est suffisamment balisé pour qu'il soit possible de construire un système inspiré des grammaires locales de Silberztein (1989)[9]. De plus, nous disposons avec SATO d'un système informatique capable de mettre en oeuvre cette stratégie, de l'appliquer sur notre corpus et d'en valider la performance.

3 - Extraction du corpus témoin de contextes dont les catégories cibles ne sont pas ambiguës.

Autour du projet SATO, nous avons construit une base de données lexicales capable d'effectuer la catégorisation du lexique de notre corpus[10]. De plus, par SATO, il est très facile de repérer les contextes possédant les caractéristiques requises. Les quatre catégories cibles qui ont été retenues pour fins de test sont le verbe conjugué, le nom commun, l'adjectif et l'adverbe.

---

8 Le projet SATO-CALIBRAGE est mené en collaboration avec Léo Laroche et Lise Ouellet du ministère de l'Éducation. Le *Cahier de recherche* no. 3, publié au Centre ATO-CI, décrit ce projet de manière exhaustive.

9 Voir Silberztein (1989).

10 Ce dictionnaire, appelé couramment «la BDL» (base de données lexicales), a été développé au départ par Luc Dupuy dans le cadre du projet SACAO (Système d'analyse de contenu assistée par ordinateur, Programme Actions spontanées, FCAR 1989-91) dirigé par Jules Duchastel alors qu'il était directeur du Centre d'ATO.

4 - Soumission des contextes au modèle associatif (phase d'apprentissage).

Nous nous proposions de tester une variété de modèles statistiques ou neuronaux. En pratique, les contraintes de temps ont fait en sorte qu'un seul modèle a pu être testé, celui de la corrélation simple (Pearson) qui se rapproche d'un réseau de Kohonen[11].

5 - Extraction du corpus témoin des contextes dont les catégories cibles sont ambiguës et pour lesquels nous disposons d'une règle de désambiguïsation.

Cette étape est semblable à l'étape trois. La différence tient aux critères de sélection des contextes qui doivent posséder en position cible des lexèmes possédant plus d'une catégorie grammaticale. Aussi, nous sélectionnons des contextes dont le lexème en position cible est ambigu (polycatégoriel). De plus, nous choisissons, parmi ces contextes, ceux pour lesquels l'ambiguïté peut être levée par une règle. On s'épargne ainsi un lourd travail de catégorisation manuelle et on obtient des contextes pour lesquels on dispose déjà d'un premier dispositif algorithmique éprouvé.

6 - Soumission des contextes au modèle associatif (phase de prédiction).

Lors de cette étape, l'information catégorielle des contextes doit être utilisée pour prédire la meilleure catégorie en position cible. Ainsi, dans notre expérimentation test, on calcule

$$\hat{E} = V \times C$$

où $\hat{E}$ représente le vecteur des catégories cibles estimé par le produit de la matrice de corrélation (V) avec le vecteur représentant le contexte ambigu (C). La valeur la plus forte dans le vecteur $\hat{E}$ sélectionne la variable catégorielle correspondante. Dans cette première expérimentation, nous n'avons pas construit d'intervalles de confiance susceptible de produire une zone d'indécidabilité.

---

[11] Voir Kohonen (1989).

7 - Comparaison de la catégorie prédite par le modèle associatif avec la catégorie prédite par la règle de grammaire.

Dans la mesure où les contextes ambigus sélectionnés ont été choisis parce que l'on disposait déjà d'un dispositif de désambiguïsation, il était possible de comparer l'efficacité du dispositif associatif par rapport à une règle linguistiquement fondée. Nous évitons ainsi de comparer nos résultats avec des décisions humaines pouvant faire appel à des motivations qui dépassent la grille catégorielle fournie au dispositif associatif. On est donc davantage en mesure de faire la distinction entre la performance du système catégoriel et la performance du dispositif associatif.

8 - Reprise de l'expérimentation en faisant varier les paramètres.

Ce protocole expérimental a été conçu pour comparer diverses méthodes et paramètres. Il est donc possible de faire varier les modèles mais aussi les données. Ainsi, on peut tester l'apprentissage en fournissant des jeux de données avec plus ou moins d'ambiguïtés sur les contextes.

3.2 La codification des données

Avant de donner les résultats obtenus lors d'une première expérimentation réalisée dans le cadre du protocole présenté, nous allons décrire le processus de codification qui nous permet de passer du texte intégral aux vecteurs utilisés par les algorithmes numériques.

La catégorisation du lexique du corpus se fait au moyen d'un scénario de commandes SATO (DOGRAMR.CSA). Ce scénario crée une propriété lexicale «gramr» qui définit la grille catégorielle utilisée pour l'étiquetage grammatical. La procédure consulte ensuite un dictionnaire pour inscrire dans la propriété «gramr» les valeurs trouvées dans le dictionnaire pour une entrée lexicale donnée. Finalement, le scénario complète la catégorisation par une analyse morphologique des lexèmes spéciaux tels les nombres. Le tableau qui suit présente un extrait du lexique catégorisé par le scénario DOGRAMR. Dans cet extrait, nous avons sélectionné des lexèmes polycatégoriels.

fréq	gramr	
20	(aux,v_conj)	a
1	(v_conj,ppassé)	admis
2	(v_conj,nomc)	affaire
7	(aux,v_conj)	ai
2	(aux,v_conj,nomc)	aura
2	(aux,v_conj)	avais
1	(aux,v_conj)	avait
1	(v_conj,nomc)	avantages
1	(aux,v_conj)	avons
2	(v_conj,nomc)	change
3	(v_conj,ppassé)	choisis
1	(v_conj,nomc)	coiffe
2	(v_conj,ppassé)	compris
1	(adj,v_conj)	contente
2	(v_conj,nomc,prép)	contre
2	(v_conj,nomc)	contrôle
2	(v_conj,nomc)	costumes
1	(v_conj,nomc)	défends
1	(v_conj,nomc)	demande
1	(v_conj,nomc)	dépenses
2	(v_conj,nomc)	disputes
1	(v_conj,nomc)	doit
1	(adj,v_conj,nomc)	domestiques
1	(v_conj,nomc)	donnes
1	(adj,v_conj)	dure
4	(v_conj,nomc)	élèves
1	(v_conj,nomc)	enquête
1	(v_conj,prép)	entre
1	(aux,v_conj)	es
38	(aux,v_conj)	est
1	(aux,v_conj,nomc)	étais

Ensuite, on définit une deuxième propriété, une propriété contextuelle, que nous nommerons «syntaxe» et qui va hériter des valeurs de la propriété «gramr». En d'autres mots, la propriété «syntaxe» va s'appliquer à chacune des occurrences des lexèmes et va recevoir comme valeur de départ l'ensemble catégoriel s'appliquant au lexème hors contexte.

Finalement, l'application du scénario DESAMBIG va modifier les valeurs de la propriété syntaxe en appliquant des règles de grammaire permettant de lever l'ambiguïté catégorielle[12]. Pour faciliter le travail d'analyse, la procédure va également déposer le numéro de la règle sur le mot désambiguïsé. On retrouve donc, en plus du texte intégral, une propriété grammaticale hors contexte («gramr») et une propriété grammaticale en contexte («syntaxe») qui contient un sous-ensemble des catégories de «gramr». L'extrait de texte qui suit donne un exemple du résultat de l'application des règles sur un paragraphe de texte. Les mots en italique n'ont pu être désambiguïsés alors que les mots en gras, suivis d'un nom de règle, correspondent aux mots pour lesquels la procédure DÉSAMBIG a réussi à lever l'ambiguïté catégorielle. Une règle de type «c» confirme la catégorie verbale alors qu'une règle «d» l'infirme.

L'opinion de KARINE

> Elle est/c2 pour le fait/d1 de laisser les jeunes choisir. La *manière* de s'habiller, c'est/c2 une affaire/d2 de goût personnel. Les jeunes de notre âge savent ce qu'ils aiment mieux_que leurs parents. Il y a/c2 une mode pour les jeunes et une autre pour les adultes. Les jeunes connaissent la mode mieux_que leurs parents. Si certains ont/c2 des goûts excentriques ; c'est/c2 un phénomène passager. Si les jeunes sont/c2 assez autonomes/c2 pour faire des tâches/d2 domestiques/d3, ils le sont/c2 assez pour choisir leur *garde-robe.* Elle se demande/c1 pourquoi les jeunes devraient se priver des vêtements coûteux alors_que certains parents dépensent beaucoup d'argent pour leurs costumes/d1 et leurs habits. Quand les parents choisissent à la place/d1 des jeunes, ceux-ci ne portent pas ce que les parents achètent.

Pour l'application des modèles associatifs, on doit traduire cette information catégorielle en information numérique. Pour ce faire, on utilise la métaphore graphique. Ainsi, on va considérer chaque contexte comme une photo composée d'un ensemble de pixels correspondant à autant de variables binaires. Chaque ligne de l'image correspond à une position dans la

[12] On trouvera une présentation détaillée de la procédure DESAMBIG dans Daoust et Dupuis (1994).

séquence de mots. Dans notre expérimentation, nous avons utilisé des contextes de longueur fixe composés de trois mots avant et de trois mots après la position cible. Nous avons donc des images composées de sept lignes avec, en position centrale, une catégorie cible. Chacune des valeurs possibles (catégories Nom commun, Verbe, etc.) de la propriété «syntaxe» correspond à une variable binaire, c'est-à-dire à un pixel. Donc, un mot qui est ambigu entre un verbe et un nom aura deux pixels en position vrai (+1) et les autres en position fausse (0). Chaque image est donc formée d'un vecteur de 273 variables. C'est SATO qui construit cette représentation binaire de la propriété «syntaxe» en notation octale. Un programme ad hoc construit finalement le vecteur en fonction du format d'entrée du modèle associatif.

Le tableau suivant présente un extrait d'une sortie de SATO dans laquelle on retrouve des contextes avec une représentation binaire (octale) de la propriété «syntaxe». La colonne «#Occ» indique le numéro du mot dans le texte. La colonne «#Lex» indique le numéro du lexème alors que la colonne «syntaxe» identifie les catégories syntaxiques. L'extrait qui suit a servi pour l'apprentissage. Donc, pour chacun des contextes, le mot en position centrale a une catégorie syntaxique non ambiguë.

#Occ	*#Lex*	*syntaxe*	
3	7	#4000000000000	,
4	2916	#10000000020	le
5	22	#200000	10
6	3040	#20000000	mai
7	49	#200000	1991
9	1208	#20004002	chère
10	179	#40000000	Julie
19	3206	#10020000000	moi
20	1183	#100000000000	chez
21	3217	#400000	mon
22	2492	#20000000	grand-père
23	9	#400000000	.
25	2660	#10000000000	il
26	2576	#4000	habite
22	2492	#20000000	grand-père

23	9	#400000000	.
25	2660	#10000000000	il
26	2576	#4000	habite
27	5570	#100000000000	à
28	2876	#10020000020	a
29	2232	#20004002	ferme
29	2232	#20004002	ferme
30	9	#400000000	.
31	3401	#10000000000	nous
32	357	#4000	aiderons
33	3217	#400000	mon
34	2492	#20000000	grand-père
35	5570	#100000000000	à

Dans le cas du modèle de corrélation qui nous a servi de test, nous construisons à partir des vecteurs d'apprentissage une matrice de corrélation de 4 lignes (les 4 catégories de la position cible) par 234 colonnes (6 mots de contexte X 39 catégories possibles) . En rappel, le produit de la matrice avec un vecteur d'entrée nous donne une valeur pour chacune des 4 catégories de la position cible. On retient la catégorie dont la valeur est maximale.

Il s'agit là bien sûr d'un modèle très simple mais qui a suffi pour mettre au point notre chaîne de traitement. Il existe des modèles plus performants. Pour ce modèle simple, on pourrait aussi considérer des mesures d'association plus pertinentes que la corrélation de Pearson qui a le défaut de prendre en compte l'association entre l'absence d'une catégorie dans une position du contexte et l'absence d'une catégorie en position cible. Les valeurs de la matrice de corrélation sont donc très faibles.

Malgré les faiblesses du modèle test, nous avons obtenu des résultats significatifs. Le tableau qui suit présente une synthèse de deux tests effectués avec ce modèle.

SYNTHÈSE DE L'EXPÉRIMENTATION

Nombre de variables catégorielles 39	
Dimension du contexte	*3 mots à gauche* *1 mot cible* *3 mots à droite*

Fichier d'apprentissage *10 235 contextes*

Expérimentation 1

Règle : *Une forme, qui peut être soit un nom soit un verbe conjugué, n'est pas un verbe conjugué si elle est précédée d'une forme qui est strictement une préposition. La préposition peut être suivie facultativement d'un article ou d'un déterminant et d'adjectifs non ambigus.*

Exemple : *La femelle construit habituellement son nid sous un tas de larges branches...*

SATO :
```
Contexte appeler                    **
$*gramr==prép*.                     **
$*gramr=(art$,dét$)*-*.             **
$*syntaxe=v_conj*syntaxe=nomc*syntaxe :-
v_conj*&*.*règle :+d1
```

Rappel : *2 967 contextes*
Erreurs : *0.70%*

Expérimentation 2

Règle : *Une forme, qui peut être un nom ou un verbe, précédée d'une forme qui ne peut être qu'un déterminant démonstratif, n'est pas un verbe conjugué.*

Exemple : *Les béliers ressemblent en plusieurs points à ceux nés sous ce signe astrologique...*

SATO :
```
Contexte appeler                    **
$*gramr==détdém*.                   **
$*syntaxe=v_conj*syntaxe=nomc*syntaxe :-
v_conj*.*règle :+d5
```

Rappel : *190 contextes*

Erreurs : *2.60%*

*Dans les commandes SATO, «$» désigne la troncature à droite sur les caractères du mot ou sur les caractères d'une valeur de propriété. «*gramr», «*syntaxe» et «*règle» introduisent un nom de propriété. «=» teste la présence d'une catégorie. «==» teste la présence d'une catégorie à l'exclusion de toute autre. « :=» affecte une valeur à propriété pour le mot désigné.*

« :-» enlève une valeur à la propriété alors que « :+» l'ajoute. «.» indique que la distance entre deux mots est de 1. Donc, les mots doivent être adjacents. «*-» indique qu'un mot est facultatif. «*&» indique qu'un mot doit être au centre du contexte.*

Certaines des différences entre les catégories prédites par la règle linguistique et le modèle associatif tenaient en fait à des erreurs de catégorisation du lexique. D'autre part, nous avons noté que l'écart entre les deux plus grandes valeurs dans la prédiction du modèle associatif est nettement plus faible dans les cas d'erreurs que pour les bons estimés. Cela laisse entendre qu'il serait possible d'établir une zone d'incertitude qui interdirait au modèle de choisir une catégorie unique lorsque la marge de certitude est trop faible.

Finalement, nous avons pu constater que la grille catégorielle n'était pas assez fine pour plusieurs règles linguistiques. Il faudra donc raffiner la grille pour établir une juste comparaison entre les deux dispositifs. De même, l'ajout de traits de nombre et de genre va permettre de tenir compte des règles d'accord qui peuvent s'avérer très efficaces dans des stratégies de désambiguïsation par les contextes immédiats.

4 Conclusion

Même si l'architecture et l'environnement de développement des ordinateurs parallèles ne nous ont pas permis d'aboutir à un gain d'efficacité dans l'implantation d'un SATO parallèle, le travail de recherche réalisé nous a permis de concevoir une architecture intéressante pour une implantation sur des ordinateurs parallèles qui seraient dotées de capacités satisfaisantes de communication et d'entrée-sortie. Par ailleurs, la stratégie de traitement distribué envisagée dans le cadre de ce projet est aussi réalisable sur des ordinateurs conventionnels, notamment dans le cadre du modèle client-serveur.

Le protocole expérimental réalisé pour l'expérimentation d'algorithmes de désambiguïsation en contexte semble très approprié pour une analyse comparative bien contrôlée. Le protocole a été testé avec un modèle de corrélation simple et la chaîne de traitement s'avère pleinement fonctionnelle.

Les résultats de l'expérimentation sont significatifs. Ils ont permis d'identifier des lacunes dans la grille catégorielle. Dans certains cas, les résultats obtenus ont révélé des erreurs dans la catégorisation du corpus. Plusieurs pistes sont envisagées pour améliorer la performance du modèle utilisé. Aussi, des modèles plus sophistiqués pourraient permettre de confronter ce premier modèle.

Finalement, la stratégie consistant à comparer des modèles associatifs avec des règles linguistiques plus classiques apparaît intéressante. D'abord, elle nous permet de faire l'économie d'une catégorisation manuelle exhaustive. Ensuite, elle permet de contrôler davantage les conditions de l'expérimentation en confrontant les divers modèles à une même grille catégorielle. Elle permet finalement de dépister les faiblesses de la grille catégorielle.

Enfin, au delà du problème type choisi pour cette expérimentation, notre approche devrait pouvoir s'appliquer à d'autres problèmes de désambiguïsation catégorielle. Les modèles construits sont donc susceptibles d'être généralisés et de fournir de nouveaux outils pour l'analyse de texte par ordinateur.

Bibliographie

Anderson , J. A., J. W. Silverstein, S. A. Ritz et R. S. Jones, 1977, Distinctive Features, Categorical Perception and Probability Learning : Some Applications of Neural Model *Psychological Review*, No. 84. pp. 413-451.

Bégin, J. et R. Proulx, 1993, *Categorization in unsupervised neural networks : The EIDOS model*, soumis pour publication.

Bourque, G., G. Duchastel et J. Beauchemin, 1994, *La société libérale duplessiste*, Montréal, Les Presses de l'Université de Montréal.

Daoust, F., 1992, *SATO Manuel de référence*, Centre d'analyse de textes par ordinateur, Université du Québec à Montréal.

Daoust, F. et F. Dupuis, 1994, Le dépistage en contexte des verbes conjugués à l'aide du logiciel SATO, *ICO-QUÉBEC*, vol. 6, no. 1 et 2, pp.106-113.

Daoust, F., L. Laroche et L. Ouellet, à paraître, SATO-CALIBRAGE : un outil d'assistance au choix et à la rédaction de textes, *Revue Québécoise de linguistique*.

Fujisaki, T. F., Jelineq, J. Cocke, E. Black, T. et Nishino, 1989, A Probalilistic Parsing Method for Sentence Disambiguation, présenté au *International Workshop on Parsing Technologies*, CMU, repris in *Current Issues in Parsing Technology*, Masaru Tomita (éd.), Kluwer Academic, 1991, pp. 139-152.

Kohonen, T., 1989, *Self-Organization and Associative Memory*, Berlin : Springer-Verlag.

Milne, R., 1988, Lexical Ambiguity Resolution in a Deterministic Parser in *Lexical Ambiguity Resolution*, Steven L. Small, Garrison W. Cottrel and Michael K. Tanenhaus (éd.), Morgan Kaufman Publishers, pp.45-71.

Proulx, R. et J. Bégin, 1990, A new learning algorithm for the BBS model, *Proceedings of the International Joint Conference on Neural Networks*, in *Neural & Cognitive Sciences Track*. Hillsdate, NJ : Laurence Erlbaum, pp. 704-706.

Proulx, R. et J. Bégin, 1993, *The use of a dual hebbian and anti-hebbian learning rule for categorization in neural networks*, soumis pour publication.

Proulx, R. et J.-G. Meunier, 1994, *Implantation de réseaux neuronaux sur systèmes à base de transputers*, rapport du projet ALEX-UQAM.

Silberztein, M., 1989, *Dictionnaire électronique et reconnaissance lexicale automatique*, Thèse de doctorat en informatique, LADL, Université Paris 7.

Smith, G. W., 1991, *Computers and Human Language*, Oxford University Press.

GÉNÉRATION INTÉGRÉE DE TEXTES ET DE GRAPHIQUES STATISTIQUES

Massimo Fasciano
Guy Lapalme

Résumé

Les graphiques et le texte sont deux médias très différents. Heureusement, lorsque leur intégration est bien effectuée, ils se complètent à merveille : une image permet de montrer alors qu'un texte permet de décrire. Dans le cadre de notre recherche, nous étudions l'interaction entre le texte d'un rapport statistique et ses figures. Nous nous intéressons aussi à l'influence de l'intention du rédacteur (ses buts) sur un rapport. Notre but final est la réalisation d'un système capable de générer automatiquement des rapports statistiques contenant du texte et des graphiques à partir de données brutes.

1 Introduction

1.1 Qu'est-ce qu'un rapport?

Le but du rapport est de produire une synthèse organisée à partir de données ou expériences concrètes. Bien qu'en principe un rapport doive être objectif, cette synthèse est quelquefois biaisée. Le rapport a un but bien précis, il cherche à faire passer le point de vue de la personne ou organisation qui le

rédige. Pour arriver à ce but, il faut souvent "tordre" un peu les faits, c'est-à-dire mettre en évidence ceux qui supportent nos conclusions et en éclipser d'autres.

Le niveau d'organisation d'un rapport est plus élevé que celui d'un texte narratif. En effet, le rapport doit être structuré de façon à ce que son analyse ait plus de poids. Pour un texte narratif, la structure est plus libre et suit l'événement qui est raconté. Les rapports n'offrent toutefois pas une structure aussi stricte que les textes procéduraux : tous les manuels d'instructions sont structurés de façon similaire, de même pour les recettes de cuisine.

On peut classifier les rapports selon le niveau d'organisation de l'information qu'ils présentent. En effet, un rapport peut être construit à partir de données brutes comme des tableaux de nombres ou d'une information beaucoup plus structurée telle un autre texte. Ce dernier type de rapport peut être un résumé d'un texte plus long ou une critique de celui-ci. Ces exemples montrent deux extrêmes, mais il existe d'autres rapports qui partent de données plus ou moins organisées.

Cette classification est très importante car elle indique le degré de liberté dont on dispose pour produire un rapport. Il est clair qu'il y a beaucoup plus de choix en partant de nombres que d'un texte organisé. Un résumé suivra l'ordre et la structure du texte original; on se limite alors au choix de l'information essentielle.

Les rapports qui nous intéressent, les rapports statistiques, ont une particularité intéressante : les données sont nombreuses et peu évocatrices lorsque présentées de façon brute. Sans une analyse statistique préalable pour faire ressortir les points importants ainsi qu'une organisation et présentation très efficaces, on risque de perdre le lecteur.

1.2 Intégration de graphiques au texte

Les graphiques et le texte sont deux médias très différents. Heureusement, lorsque leur intégration est bien effectuée, ils peuvent se compléter à merveille : une image permet de montrer alors qu'un texte permet de décrire. Ainsi, un rapport qui

présente des données peut se servir d'images pour les montrer sous leur forme brute et du texte pour en faire l'analyse.

Évidemment, la frontière montrer/analyser n'est pas absolue et il est tout à fait possible de se servir d'outils graphiques pour présenter les résultats d'une analyse, mais sans un accompagnement textuel, l'analyse risque de ne pas atteindre son plein potentiel.

Inversement, le texte peut aussi servir à "montrer" par une description très détaillée. Cependant, on dit qu'une image vaut mille mots, et dans une telle situation, il est beaucoup plus efficace d'utiliser la puissance visuelle des graphiques plutôt que d'alourdir le rapport avec de longues descriptions. Ainsi, il est préférable de montrer une courbe plutôt que de lister la valeur des points qui la composent. D'un seul coup d'oeil, on voit les tendances importantes sur la courbe.

Dans le cas des rapports statistiques, on dispose d'outils graphiques comme les courbes, les tartes, les barres/colonnes qui sont adaptés à des types particuliers de messages. Par exemple, pour montrer la décomposition d'un ensemble de données, on peut les présenter sous forme de tarte, où l'angle d'un secteur indique la proportion occupée par l'élément associé. C'est efficace car le lecteur décode d'un simple coup d'oeil une information qui ne serait pas aussi évidente si les données étaient sous une forme brute. On peut ensuite associer un paragraphe de texte au graphique pour faire passer notre message de façon plus précise. Par exemple, la phrase "La part de marché de notre compagnie dépasse le total des parts de nos compétiteurs" donnerait au lecteur une interprétation des données qu'il n'aurait peut-être pas eue en ne regardant que le graphique.

1.3 Un exemple

L'information à présenter dans un rapport statistique est souvent donnée sous une forme peu parlante. Par exemple, le tableau 1 contraste l'évolution de 2 compagnies (Xyz inc. et Pqr inc.) entre 1985 et 1994.

	Chiffres d'affaires	
Année	Xyz inc.	Pqr inc.
1985	50	75
1986	60	79
1987	62	76
1988	64	71
1989	69	72
1990	76	70
1991	83	72
1992	89	83
1993	93	98
1994	102	120

Tableau 1 : Évolution des profits de Xyz. et Pqr.

Il est très difficile d'avoir une vue d'ensemble sur les données lorsqu'elles sont présentées sous cette forme. On a beaucoup de difficulté à visualiser l'évolution du chiffre d'affaires de chaque compagnie et il est encore plus laborieux de les comparer. Cependant, on peut lire avec précision les valeurs individuelles.

Pour donner un sens à cette information, on dispose de 2 outils : le texte et les graphiques. La figure 1 montre comment on pourrait l'analyser en utilisant une description textuelle.

> Le chiffre d'affaires de Xyz inc. a augmenté de 52 millions entre 1985 et 1994. Celui de Pqr inc. a augmenté de 45 millions. L'augmentation a été graduelle pour Xyz, alors que celle de Pqr a eu lieu surtout entre 1991 et 1994 (48 millions) après quelques fluctuations les années précédentes. Pqr a commencé et fini en tête mais a été légèrement dépassée entre 1990 et 1992 par son compétiteur.

Figure 1 : Description textuelle pour les données du tableau 1

Cette description rend les données du tableau beaucoup plus utiles mais elle ne résout pas tous les problèmes. En effet, le texte est un peu long, ce qui risque de perdre le lecteur. Celui-ci n'en dégagera donc pas les points essentiels. Il faut essayer de le simplifier sans affaiblir le message. De plus, certaines affirmations du texte sont assez difficiles à vérifier à l'aide d'un simple coup d'oeil au tableau.

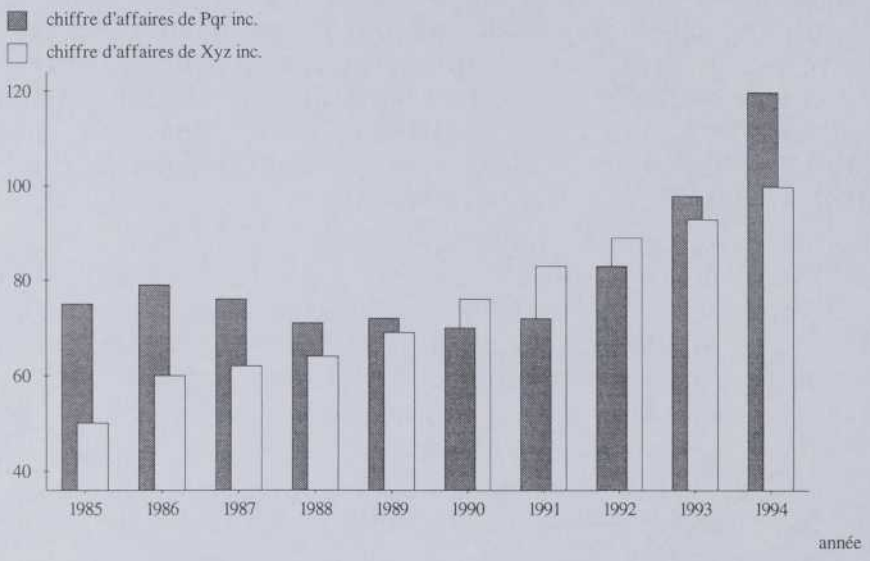
L'ajout d'un graphique au texte permet de le simplifier en transférant certains éléments quantitatifs vers le graphique. Par exemple, il devient inutile[1] de quantifier les augmentations ou d'indiquer si elles ont été graduelles. De plus, tout ce qui est dit dans le texte se vérifie d'un simple coup d'oeil et attire beaucoup plus l'attention du lecteur car l'information n'est pas noyée dans un texte trop long.

La figure 2 présente un cas où le texte montre les points importants de l'analyse, alors que le graphique s'occupe de donner une vue d'ensemble.

Dans d'autres situations, on peut ajouter un texte à un ensemble de graphiques pour établir un lien entre ceux-ci. Cette situation est assez courante lorsqu'on cherche à présenter plusieurs "vues" sur les mêmes données afin de mieux exposer toutes les tendances. À la figure 3, on commence par présenter l'évolution de 2 données pour ensuite montrer qu'elles sont fortement corrélées. Le texte évite donc de "parachuter" une série de graphiques sans description. Il sert à décrire l'utilité de chaque graphique plutôt qu'à compléter ceux-ci par une analyse.

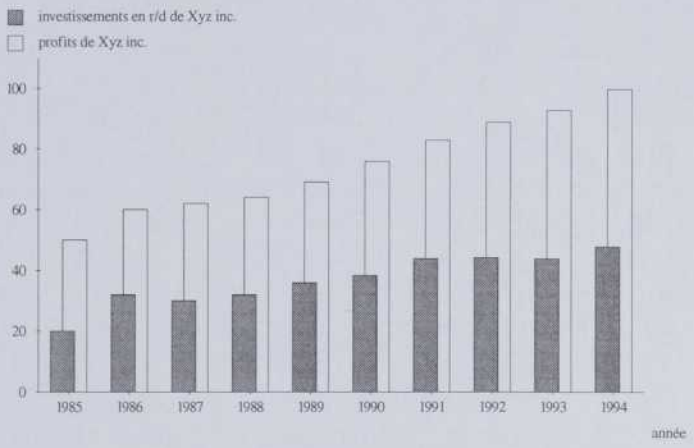
---

[1] à moins de vouloir fortement insister sur ces faits.



Xyz inc. et Pqr inc. ont vu leur chiffre d'affaires augmenter de 1985 à 1994. Pqr est restée en tête sauf entre 1990 et 1992.

Figure 2 : Combinaison texte/graphique pour les données de la figure 1



La figure ci-dessus montre l'évolution des profits de la compagnie Xyz inc. et de ses investissements en recherche et développement entre 1985 et 1994. Tous deux ont augmenté graduellement pendant la période en question (50 et 28 millions respectivement). Il semblerait que l'augmentation des profits soit liée à l'augmentation des investissements en r/d. Cette tendance est confirmée par la figure ci-dessous.

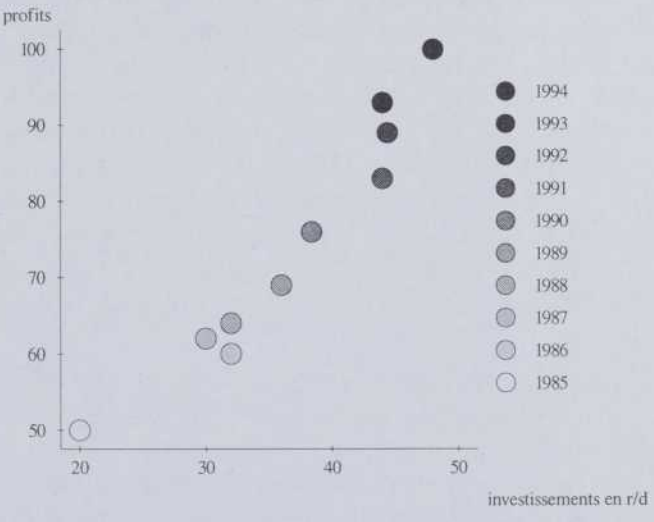


Figure 3 : Combinaison d'un texte et de plusieurs graphiques

2 Caractéristiques des graphiques

Parmi les études sur le choix de graphiques statistiques, nous nous sommes attardés sur celle de MacKinlay (1986) qui a l'avantage de fournir un algorithme basé sur une extension de la classification de Bertin (1983) (variables nominales, ordinales et quantitatives) et qui définit un ensemble de méthodes graphiques pour l'expression de chaque type de variable. Ces méthodes sont ordonnées et l'algorithme fait l'allocation des méthodes aux variables en évitant de générer des solutions impossibles (par exemple, 3 coordonnées spatiales dans un graphique à 2 dimensions).

Nous avons aussi intégré à notre modèle des résultats plus théoriques comme ceux de Tufte (1983, 1990) et de Zelazny (1989).

On retient de ces travaux une classification des propriétés d'un graphique tant au niveau de ses composantes élémentaires (Bertin 1983 et MacKinlay 1986) qu'au niveau global (Zelazny 1989). Les propriétés des composantes élémentaires jouent surtout sur l'efficacité d'un graphique pour exprimer un type donné de variable. Par exemple, les positions spatiales sont plus utiles que les couleurs pour exprimer des variables continues. Les caractéristiques globales d'un graphique sont surtout utiles pour déterminer son rôle dans la transmission du message. Ainsi, des graphiques dont les éléments de base sont similaires peuvent avoir des rôles un peu différents (les barres sont adéquates pour comparer des valeurs, alors qu'on utilise toujours les colonnes ou les courbes pour des données temporelles).

Notre classification des propriétés intègre ces résultats et les applique à un ensemble beaucoup plus général de graphiques. Au niveau des propriétés élémentaires, nous retenons l'efficacité (facilité et précision de lecture) pour chaque variable ainsi que la possibilité de comparer soit les valeurs d'une même variable entre elles soit les valeurs de plusieurs variables.

Au niveau de propriétés globales on retrouve la taille des graphiques utilisés pour présenter les données ainsi que des

facteurs liés au message à présenter. Connaître la taille d'un ensemble de graphiques est très utile pour déterminer s'ils seront utiles dans un contexte donné. Par exemple, pour une courte introduction à un ensemble de données on voudrait éviter de remplir plusieurs pages de figures avant même de commencer le texte.

Pour ce qui est du message[2], il est important de savoir si le graphique transmet bien les informations suivantes:

- évolution des valeurs
- corrélation entre les variables
- répartition des valeurs
- décomposition des valeurs

Parfois, la sélection d'un graphique peut se faire à l'aide d'un critère global. On peut voir à la figure 4 qu'une tarte permet de montrer la décomposition d'un total (ici un pourcentage) beaucoup mieux que des barres. Bien qu'elle soit moins utile pour la sélection, on peut noter aussi une différence au niveau de l'efficacité des 2 graphiques : les barres permettent une lecture et une comparaison des valeurs beaucoup plus précise que la tarte (par exemple, Genco et Xyz inc.).

Dans d'autres situations, on ne peut pas compter sur le rôle du graphique. C'est le cas à la figure 5 où on retrouve des petites différences liées à la lecture et à la comparaison de valeurs mais aucune au niveau du message. Ainsi, on remarque que la figure du haut permet une lecture précise des valeurs à l'aide de son échelle linéaire alors que celle du bas utilise des surfaces, méthode qui ne permet que des lectures relatives. Par contre, celle du bas est plus orthogonale dans ses comparaisons. En effet, il est aussi facile de comparer selon les mois que selon les provinces, alors que la figure du haut encourage plutôt la comparaison entre les mois pour une province donnée (localité des colonnes). Évidemment, la comparaison orthogonale de la figure inférieure n'est possible que lorsque les

[2] Bien que la comparaison soit considérée comme un message dans notre système, elle est utile à un niveau plus élémentaire dans la sélection des graphiques.

valeurs sont très différentes car l'œil est peu sensible aux différences de surface.

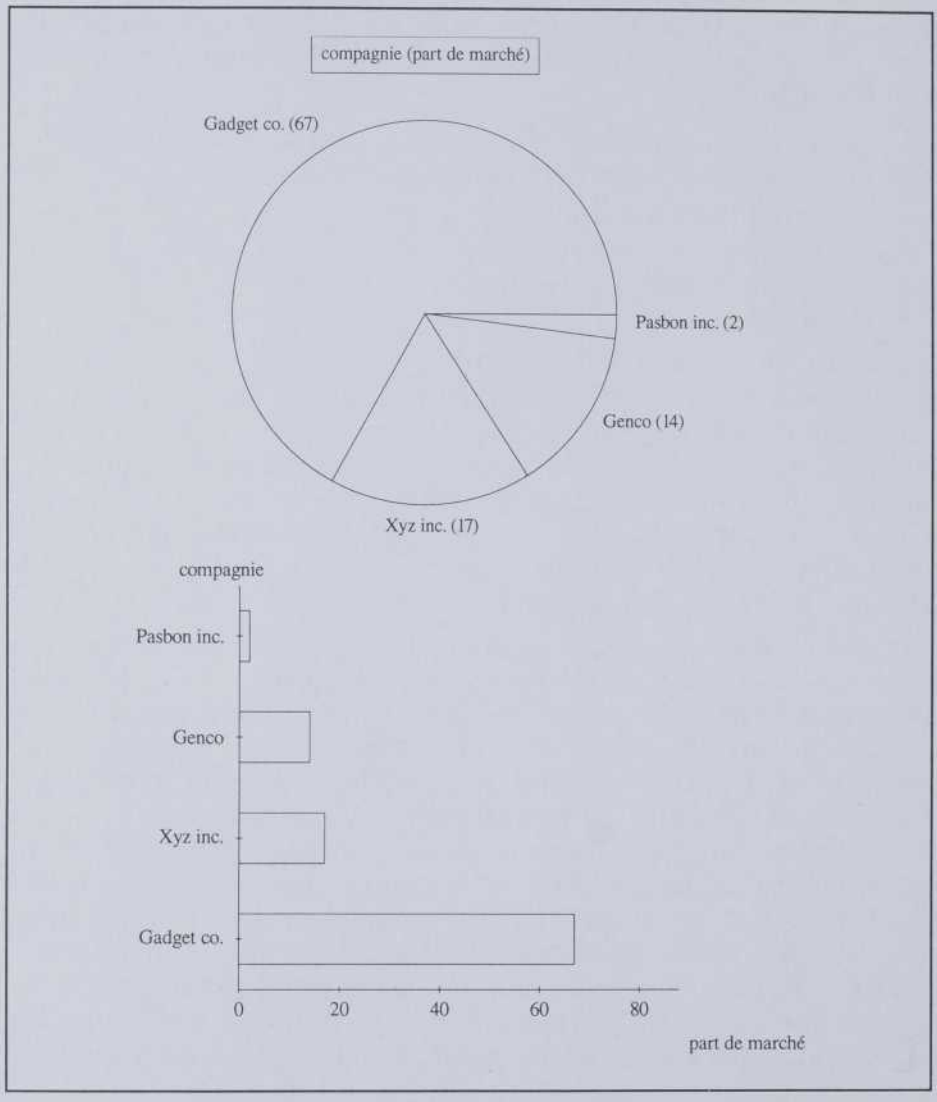


Figure 4 : Comparaison de l'efficacité d'une tarte et d'un graphique en barres

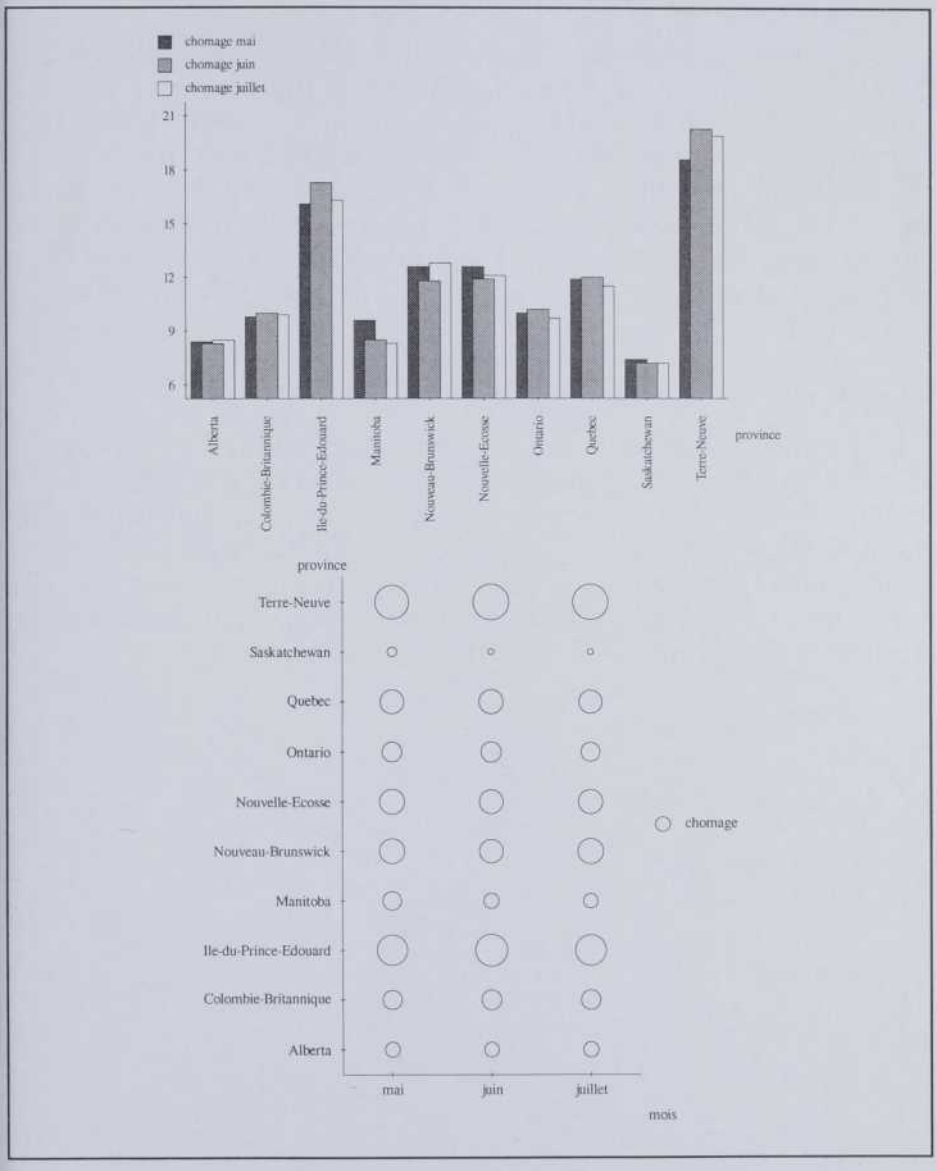


Figure 5 : Comparaison de l'efficacité des colonnes et des surfaces (graphique à 3 variables)

3 Intention du rédacteur

Alors que la recherche porte d'habitude sur le modèle du lecteur (Paris 1991), nous étudions plutôt celui du rédacteur du rapport. Les rapports sont parfois et même souvent biaisés par les intentions et les buts du rédacteur. On doit donc tenir compte des rapports objectifs mais aussi des rapports subjectifs dans lesquels on essaie de faire "mentir" un peu les données. L'intention du rédacteur affecte également la façon dont les graphiques et le texte sont combinés. Cet aspect du problème a été très peu étudié dans le domaine.

3.1 Classification

L'intention du rédacteur affecte directement le message à présenter dans un rapport. Une très bonne étude sur la correspondance entre le message et le graphique a été faite par Zelazny (1989). Dans cet ouvrage, il identifie 5 messages[3] de base (la décomposition, la position, l'évolution, la répartition, la corrélation) et 5 graphiques (voir figure 6) très utilisés dans les rapports et il explique le lien entre ces deux ensembles.

[3] Zelazny les appelle "types de comparaison", mais il affirme qu'ils correspondent directement aux messages que le rédacteur peut transmettre.

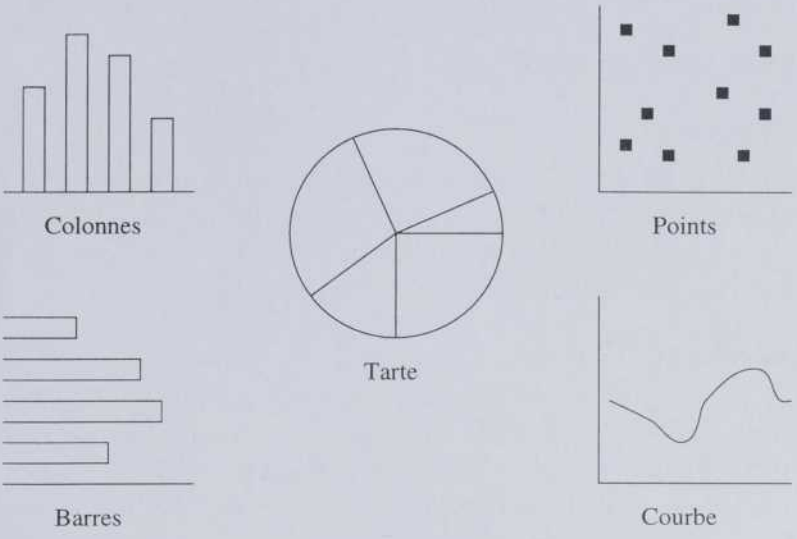


Figure 6 : Les 5 graphiques de base selon Zelazny

3.1.1 La décomposition

Dans une décomposition, on cherche à montrer des valeurs en tant que fraction d'un tout. La tarte, de par sa nature, est presque toujours le meilleur graphique pour montrer une décomposition.

3.1.2 La position

Ce type de message s'applique au classement d'éléments entre eux. Par exemple, comparer les profits de plusieurs entreprises. Le graphique à barres est le plus approprié pour ce type de message. Il rend la comparaison des valeurs facile (surtout si elles sont triées). Le graphique en colonnes est à éviter car il donne une impression d'évolution (on perçoit mieux le temps sur un axe horizontal).

3.1.3 L'évolution

Ce type de message vise à montrer des changements dans le temps, comme par exemple, la variation des profits d'une entreprise au fil des années. Le graphique en colonnes (données

non-continues) et la courbe (données continues) sont appropriés pour ce type de message.

3.1.4 La répartition

C'est une répartition de fréquence dans laquelle on cherche à montrer combien d'échantillons se trouvent dans chacun des intervalles considérés. Les graphiques qui correspondent à ce type de message sont les colonnes et la courbe.

3.1.5 La corrélation

Une comparaison de corrélation montre la relation entre 2 variables. Elle permet de tester si le comportement d'une variable est lié à celui d'une autre. On utilise souvent le graphique en points (avec une courbe de régression) pour montrer une corrélation.

Cette étude est un très bon point de départ mais elle laisse beaucoup à désirer à deux points de vue : tout d'abord les 5 graphiques traités par l'auteur ne sont pas suffisants pour de vrais rapports, et ensuite les messages utilisés, bien que très utiles semblent manquer d'organisation (décomposition et position ont des points communs qui ne sont pas traités, évolution est trop générale, etc.). Nous avons donc étendu son étude à des graphiques plus complexes tout en organisant ses messages en une hiérarchie qui en illustre mieux les points communs.

La figure 7 présente cette hiérarchie et la figure 8 montre la différence entre 4 messages spécifiques de la branche "évolution" : augmentation, diminution, stagnation et récapitulation. Les données décrivent la variation des profits (en millions de $) d'une compagnie entre 1971 et 1978.

- comparaison
  - décomposition
  - position
- évolution
  - variation
    - * augmentation
    - * diminution
  - stagnation
  - récapitulation
- corrélation
  - forte
  - faible
- répartition
  - intervalles fixes (classes d'âge)
  - répartition désirée (uniforme, cloche, ...)

Figure 7 : Hiérarchie des messages

Année	Profits
1971	10
1972	11
1973	11
1974	18
1975	16
1976	16
1977	15
1978	19

AUGMENTATION: Les profits de la compagnie ont augmenté de 8 millions entre 1971 et 1974, puis de 3 millions entre 1975 et 1978 après une chute de 2 millions.

DIMINUTION : Les profits de la compagnie ont diminué de 3 millions entre 1974 et 1977.

STAGNATION : Les profits de la compagnie ont stagné entre 1971 et 1973, puis entre 1975 et 1977 après une grosse augmentation en 1974.

RECAPITULATION : Les profits de la compagnie ont subi 3 ans d'augmentation, 2 ans de stagnation et 2 ans de diminution entre 1971 et 1978.

Figure 8 : Adaptation du texte à l'intention du rédacteur

3.2 Effet

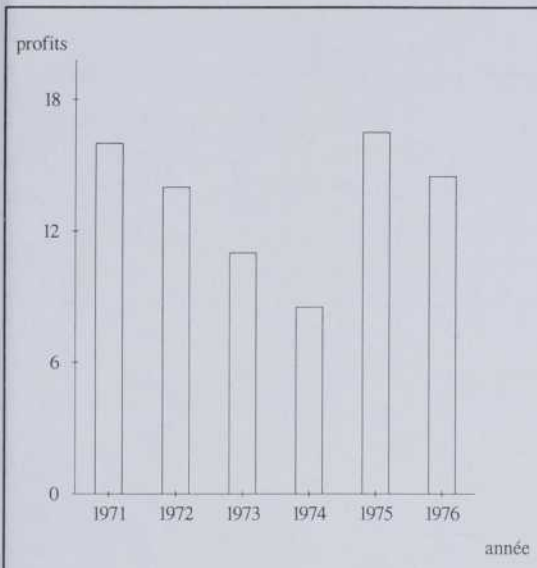
L'intention du rédacteur a un effet très important sur l'utilisation du texte et des graphiques. Cette influence n'est pas tout à fait la même dans un rapport subjectif que dans un rapport objectif.

Au niveau des rapports objectifs, l'effet de l'intention du rédacteur est assez claire. Certains graphiques sont mieux adaptés à certains messages (voir Zelazny) :

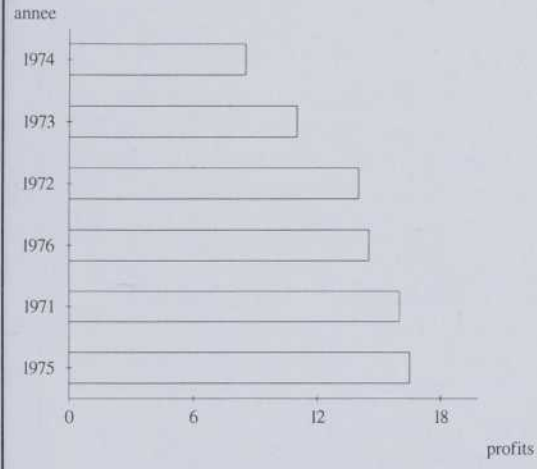
- décomposition : tarte
- évolution : colonnes ou courbe
- corrélation : points

Au niveau du texte, le contenu change complètement, comme on peut le voir à la figure 9. Cet exemple présente les mêmes données (profits d'une compagnie entre 1971 et 1976) à travers deux messages différents. La partie supérieure montre les données à travers un message d'évolution alors que celle du bas insiste plutôt sur la comparaison des années.

Pour ce qui est des rapports plus subjectifs, les effets sont beaucoup plus subtils et dépendent aussi du lecteur. Ce qui est certain, c'est qu'un graphique subjectif est plus difficile à réaliser qu'un texte subjectif car toute omission dans un graphique est suspecte. Dans un texte, l'omission est pratique courante (pour le rendre plus compact). Toutefois, comme on peut le voir à la figure 10, 2 graphiques similaires peuvent tout de même transmettre des messages très différents.



ÉVOLUTION : Globalement, les profits ont diminué malgré une forte reprise de 1974 à 1975.



POSITION : Les profits ont été à leur plus fort en 1975 et en 1971. Ils ont été à leur plus bas en 1974, avec environ la moitié de leur valeur de 1975.

Figure 9 : Effet des messages évolution et position

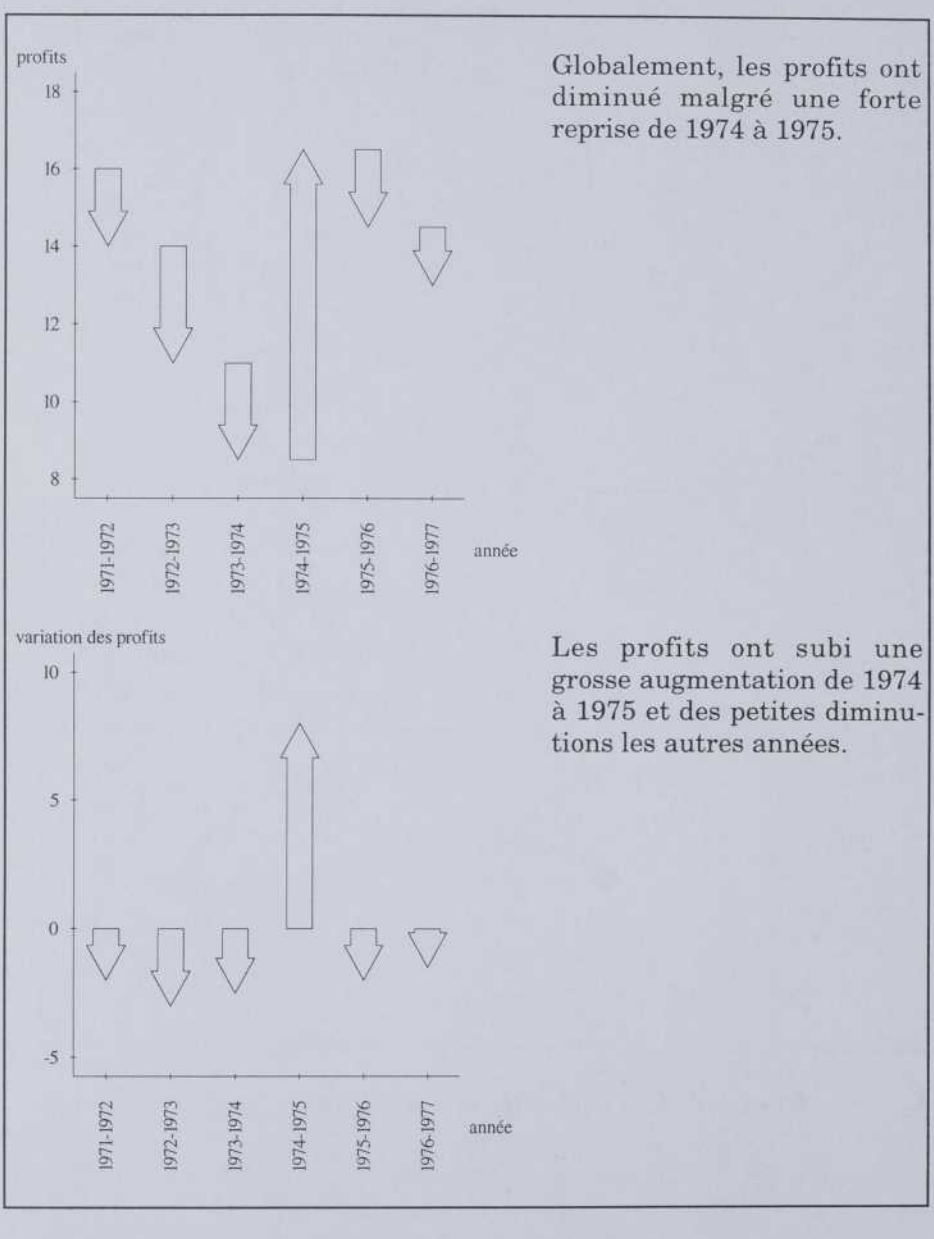


Figure 10 : Intégration texte/graphique pour un message subjectif

La partie du haut insiste beaucoup sur les fluctuations et combine texte et graphique de façon conventionnelle. Le message qu'on y présente est honnête tant au niveau du texte que du graphique. Par contre, le mariage du texte et du graphique dans la partie du bas est assez spécial. Le message présenté dans le texte est extrêmement biaisé, à un tel point qu'il ne correspond plus beaucoup aux données. En effet, il insiste sur la grosse augmentation et parle d'une série de petites diminutions, oubliant cependant de mentionner que la somme de ces diminutions dépasse l'augmentation. Le choix d'un graphique pour accompagner ce texte est assez délicat : si on omet le graphique, le lecteur se doute qu'on lui cache des choses et si on en montre trop dans le graphique, le lecteur voit bien que notre texte n'est pas honnête. On a choisi de reprendre le graphique du haut en alignant les variations à 0, ce qui rend beaucoup plus difficile la sommation de toutes les petites diminutions, tout en présentant assez d'information pour ne pas soulever de doutes.

La figure 10 montre donc une intégration texte/graphique hors de l'ordinaire : utiliser le graphique pour rassurer un peu le lecteur tout en essayant de le déjouer en présentant les données de façon à appuyer le mieux possible le texte. On voit donc que les techniques d'intégration conventionnelles ne s'appliquent pas toujours aux cas subjectifs.

4 L'implantation d'un prototype

Nous travaillons sur l'implantation d'un prototype intégrant les résultats théoriques présentés dans cet article. Ce système devra donc être en mesure de produire des graphiques et du texte pertinents et de les combiner.

Le processus de génération d'un rapport texte/graphiques fait intervenir un ensemble de modules différents, chacun capable de fournir son expertise aux autres. Selon nous, il est important d'isoler ces modules et de les rendre disponibles directement à l'utilisateur de notre système, plutôt que de lui fournir une "boîte noire" qui produit un rapport complet. La raison est simple : l'utilisateur peut avoir besoin de l'expertise

de notre système pour faire un ou des choix bien particuliers sans nécessairement vouloir un texte complet.

Un premier module s'occupe de choisir les relations intéressantes à exprimer. Il s'agit de la première phase de la planification, le choix du contenu (ex : effet de la hausse des prix sur le taux de chômage).

Ensuite, un module détermine les schémas d'expression les plus appropriés pour une relation. Parmi les schémas d'expression les plus utiles on retrouve : un texte de présentation, un graphique accompagné d'un texte d'analyse, un texte qui relie plusieurs graphiques. C'est la seconde phase de planification, le choix de la structure.

Une fois le schéma identifié, un autre module s'occupe de préciser ses composantes. C'est à ce niveau-ci qu'on détermine quel graphique utiliser (tarte, courbe, etc.) et quelle information inclure dans le texte.

La phase suivante organise l'information à l'intérieur des composantes des schémas. Par exemple, c'est ici que l'on décide si les données seront triées sur les échelles des graphiques. C'est aussi à cette étape que l'ordre des idées du texte est déterminé.

Finalement, le module de réalisation (génération de surface) se charge de dessiner les graphiques (interne à notre système) et de produire le texte (utilisation projetée d'un générateur externe).

Le langage d'implantation que nous avons choisi est Prolog à cause de ses qualités au niveau de la gestion logique de bases de règles et de données relationnelles.

Le prototype travaille avec des données relationnelles en entrée et utilise des annotations pour exprimer les unités utilisées (prix en $, poids en kg) ainsi que l'intention du rédacteur (s'il en a une) pour chacune des relations.

Le traitement des unités nécessite une certaine connaissance du monde. Celle-ci est encodée dans un graphe d'héritage multiple. À l'aide de ce graphe, dont on peut voir un extrait à la figure 11, nous pouvons organiser les propriétés des unités dans un ensemble de catégories générales (contraintes, type, organisation, etc.). Avec une telle organisation, il devient

beaucoup plus simple d'exprimer des conditions dans le système. Par exemple, un choix d'échelle peut dépendre de la propriété "ordonnée" qui englobe par héritage les variables ordinales et quantitatives.

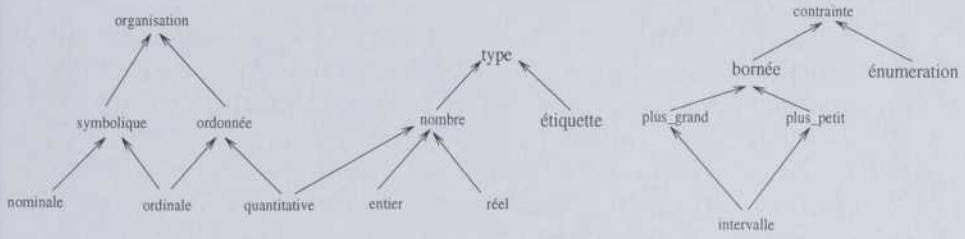


Figure 11 : Extrait du graphe d'héritage des unités

De plus, pour pouvoir exprimer les unités de façon plus naturelle, on doit traiter leur décomposition. Cela peut être aussi simple que de dire "3 millions de $" au lieu de "3 000 000 $" ou un peu plus subtil, comme par exemple lorsqu'on décide "d'approximer" 129 jours par 4 mois et 1 semaine.

Présentement, le prototype réalise une partie des modules "experts" décrits auparavant. Les couches de haut niveau traitant le problème de la planification sont encore à l'état embryonnaire, mais les modules inférieurs, en particulier le choix et l'organisation interne des graphiques sont opérationnels et le module de réalisation a produit tous les exemples graphiques de cet article. Le texte n'est pas encore traité; pour cette étape, nous comptons interfacer notre système avec un générateur de texte déjà existant (Gagnon 1993).

5 Conclusion

Les rapports statistiques sont une application intéressante de la génération intégrée de textes et de graphiques. En effet, ils cherchent à présenter une information très dense tout en s'assurant que le lecteur en retient les points essentiels. L'union des deux médias facilite énormément la tâche car les graphiques nous donnent une vue d'ensemble alors que le texte

permet de cibler les détails intéressants. Un autre aspect intéressant de ce domaine d'application est le degré de liberté à notre disposition. Puisqu'on part de séries de nombres et que l'on veut produire un rapport cohérent, on a beaucoup de choix à faire. Ceci est autant un avantage (liberté) qu'un problème (complexité).

Les rapports sont très souvent biaisés par le point de vue de la personne qui les rédige. Il faut tenir compte de cet aspect du problème autant à un niveau théorique, pour reproduire le style des rapports qu'on imite, qu'au niveau pratique, pour produire un outil qui sera utilisé par les rédacteurs. L'étude de l'effet de l'intention du rédacteur sur le format et le contenu du rapport est donc fondamentale.

Le système que nous proposons agira comme un expert capable de conseiller l'usager dans la rédaction d'un rapport. Sa flexibilité lui permettra de produire un rapport complet sans aucune intervention ou de demander au rédacteur de faire des choix subjectifs aux moments opportuns.

Bibliographie

Bertin J., 1983, *Semiology of Graphics*, The University of Wisconsin Press, traduit par William J. Berg.

Gagnon M., 1993, *Expression de la localisation temporelle dans un générateur de texte*, Thèse de doctorat, Département d'informatique et de recherche opérationnelle, Université de Montréal, Publication 888.

Mackinlay J.D., 1986, *Automatic Design of Graphical Presentations*, Thèse de doctorat, Computer Science Department, Stanford University.

Paris C. L., 1991, The role of the user's domain knowledge in generation, *Computational Intelligence*, 7, pp. 71-93.

Tufte E. R., 1983, *The Visual Display of Quantitative Information*, Graphics Press.

Tufte E. R.,1990, *Envisioning Information*, Graphics Press.

Zelazny G., 1989, *Dites-le avec des graphiques*, InterÉditions.

LES APPLICATIONS EN TERMINOTIQUE À LA DIRECTION DE LA TERMINOLOGIE ET DE LA DOCUMENTATION

Christine Leonhardt

Résumé[1]

Le développement et la mise au point d'outils destinés aux spécialistes de la langue sont au nombre des activités habituelles de la Direction de la terminologie et de la documentation (DTD) du Bureau de la traduction à Travaux publics et Services gouvernementaux Canada. Dans certains cas, c'est le changement technologique qui est à l'origine du travail de mise au point. Dans d'autres cas, ce travail répond aux exigences croissantes des clients et des gestionnaires. Cette communication présente un survol des principales activités de la DTD relatives à TERMIUM®, banque de données linguistiques du Gouvernement du Canada, et à LATTER, poste de travail du terminologue, et met en lumière les défis rencontrés lors du développement de ces outils[2].

---

1 Mise à jour et traduction de la communication suivante :
Malcolm Williams, "Terminology and LATTER: An Update", *TKE'93, Terminology and Knowledge Engineering. Proceedings. Third International Congress on Terminology and Knowledge Engineering*, Frankfurt, 1993, pp. 412 – 419.

2 Les membres suivants du personnel de la DTD ont travaillé aux projets présentés dans la communication :
John Carey – TERMIUM® sur CD–ROM, Roger Racine – TERMIUM® en direct, Gilbert Dupuis – LATTER et Yvan Cloutier – LATTER.

1 TERMIUM®

1.1 TERMIUM sur CD-ROM

Depuis 1987, le même logiciel a servi à la production de la version de TERMIUM® sur disque optique. Au moment où le premier projet à ce sujet a vu le jour, il existait peu de produits sur disque optique.

Les dernières années ont été le théâtre de changements technologiques importants et d'une stabilisation industrielle dans le domaine de l'édition sur disque optique. De façon à s'assurer que TERMIUM® bénéficie de la meilleure technologie possible et que la DTD reçoive le meilleur service possible, le Gouvernement du Canada a lancé un appel d'offres pour la production de *TERMIUM® sur CD-ROM*, et un contrat a été accordé en mai 1993.

Le serveur retenu a créé l'application en adaptant aux besoins de TERMIUM® le logiciel dont il est le propriétaire exclusif. L'édition de mai 1994 de TERMIUM® est la première édition produite par le nouveau serveur. Même si l'interface utilisateur fait l'objet de certaines modifications quant à son apparence, les fonctionnalités et les techniques d'interrogation qui ont fait le succès du produit jusqu'à maintenant sont demeurées intactes. De plus amples renseignements sur cette édition revue de *TERMIUM® sur CD-ROM* seront disponibles dans les prochains mois.

1.2 TERMIUM® en direct

C'est également le changement technologique qui est principalement à l'origine du nouveau développement de la version en direct de TERMIUM®, projet qui débuta en 1990. Quoique l'accès à la banque centrale durant la période de développement soit réservé au personnel du Bureau de la traduction, la version en direct de TERMIUM® constitue toujours un moyen essentiel de gérer et de diffuser des renseignements terminologiques qui soient à jour.

Quand le concepteur du logiciel sur lequel se fonde TERMIUM® en produisit une version très différente, il devint évident qu'il fallait redessiner en conséquence toutes les composantes de la banque. La programmation de la quatrième génération de TERMIUM® débuta donc en mars 1992. La première phase du projet de la mise à niveau de l'application est maintenant terminée.

1.2.1 Cadre du projet de développement

En raison de circonstances particulières, le projet de mise à niveau de l'application a été limité à la reproduction des caractéristiques et des fonctionnalités de TERMIUM® III à l'aide du nouveau logiciel BASISplus. La Direction décida toutefois d'apporter certaines améliorations au produit de façon à répondre aux besoins exprimés par les utilisateurs de TERMIUM® III. Parmi les objectifs essentiels poursuivis lors de la mise à niveau figurent une excellente interactivité (temps de réponse), une grande facilité de maintenance et la création d'une interface utilisateur semblable à celui de TERMIUM® III.

Le travail de mise à niveau de l'application a été effectué à l'interne en vue d'un meilleur contrôle des coûts et de l'obtention de résultats concrets aussi rapidement que possible. La DTD a mis sur pied une équipe composée de membres de son personnel et d'un ingénieur principal des systèmes embauché à contrat. La rédaction de spécifications détaillées s'est avérée inutile, l'équipe DTD connaissant bien TERMIUM®. Étant donné que la nouvelle édition du logiciel était très peu connue, le respect de telles spécifications aurait pu empêcher une utilisation pleine et adéquate des fonctionnalités du système et retarder de plusieurs mois le début du travail de programmation. Aussi l'équipe a–t–elle choisi une conception par prototypage de façon que l'utilisation du logiciel permette de dégager la conception et la programmation idéales. Ce sont les spécifications fonctionnelles préparées par le concepteur du logiciel BASISplus en consultation avec l'équipe de mise à niveau de TERMIUM® qui ont permis d'orienter la phase exploratoire.

L'équipe de mise à niveau a fait face à de nombreux défis, tant au plan administratif que technique, dont certains sont dus à la relative nouveauté du logiciel. Par exemple, il a été impossible de trouver un consultant expérimenté dans la conception d'applications à l'aide de ce produit. Le logiciel possédait certains défauts dont plusieurs, heureusement, ont pu être corrigés grâce à une nouvelle version reçue pendant les travaux de développement. On a aussi découvert que le module BASISplus préconisé pour la programmation de l'interface et de la composante permettant la saisie était peu documenté et faisait, de fait, l'objet d'un développement continu. Un accès direct au personnel R et D des concepteurs du logiciel a permis de franchir ces obstacles.

1.2.2 Conception

Le plus grand défi, cependant, a consisté à choisir des options qui permettaient d'obtenir un équilibre entre une utilisation maximale des caractéristiques du logiciel et le maintien d'une excellente performance, l'un des aspects importants de TERMIUM®. BASISplus est un système relationnel de gestion de l'information textuelle et les bases de données relationnelles sont consommatrices de ressources. Cependant, comme le logiciel ne requiert pas une conception entièrement relationnelle, il fut possible de tirer parti de l'architecture seulement quand nécessaire.

Dans un système relationnel, les éléments de données sont disséminés dans une série de tableaux (ou, selon la terminologie de BASISplus, de « types de fiches »), en fonction de leurs relations avec les autres éléments de données. Les relations permettent de tirer des renseignements figurant sur des types de fiches sélectionnées en vue de leur rassemblement sous forme variée. Bien que cette approche permette l'élimination des données redondantes, la réduction des exigences en espace d'emmagasinement et la création de diverses vues des fiches pour faire face aux besoins d'utilisateurs variés, elle peut ralentir considérablement une application.

La base de données TERMIUM® contient 46 types de fiches. Dans certains cas, il a été très facile d'organiser les données en tableaux individuels sans craindre qu'il y ait des répercussions négatives sur la performance du système. Par exemple, il était clair qu'il fallait prévoir un tableau pour les codes de projet, un autre pour l'information relative aux sources bibliographiques et un autre pour les profils d'utilisateurs. Dans d'autres cas, cependant, la recherche d'une configuration qui réponde également aux exigences relatives à la performance, à la recherche et à l'emmagasinement, s'est avérée plus difficile.

L'organisation des données terminologiques en est un exemple. L'une des exigences de TERMIUM® IV consiste en la possibilité d'avoir accès à l'information en plusieurs langues sur la même fiche. (En TERMIUM® III, toutes les fiches contiennent au maximum deux langues, dont l'une doit être soit l'anglais soit le français.) La nouvelle approche a pour but de faciliter les échanges avec d'autres banques de données et de garantir que le principe de l'uninotionnalité prévalant lors de la création d'une fiche est respecté. (En vertu de ce principe, la base de données contient une fiche par notion et chaque fiche ne représente qu'une seule notion.) Plusieurs clients de TERMIUM® préfèrent toutefois avoir accès aux renseignements terminologiques seulement en anglais et en français. La solution originale envisageait de répondre à ces besoins divergents en créant deux bases de données, l'une contenant des données en anglais et en français et l'autre contenant des renseignements dans toutes les langues qui pouvaient être emmagasinées dans TERMIUM®, y compris l'anglais et le français.

Cette approche semblait acceptable au moment où elle fut proposée étant donné que la DTD emmagasinait des données multilingues provenant d'autres organismes, sans vraiment gérer l'information recueillie. Quand il s'avéra que la DTD commencerait à gérer activement le contenu multilingue de la banque, une fois TERMIUM® IV en place, la possibilité d'emmagasiner des fiches dans deux bases de données a été éliminée.

Des facteurs additionnels comme une limite de la longueur des fiches, des exigences d'indexation particulières et une facilité de maintenance ont été pris en compte dans la conception de la fiche multilingue définitive. Comme conséquence, une fiche multilingue est, de fait, une fiche « logique », qui peut être générée de façon dynamique au moment de l'affichage. Physiquement, les données terminologiques dans chaque langue sont emmagasinées dans un type de fiches réservé à cette langue. Par exemple, il existe un type de fiches pour les termes anglais et leurs contextes d'utilisation, un autre pour le français, un autre pour l'allemand, etc. Il existe également un type de fiches contenant des champs réservés aux domaines, aux sources, au fichier, au fonds, aux codes de projet et aux autres éléments de nature administrative. Ce type de fiche sert de pivot à la fiche « logique » et reflète en quelque sorte l'existence d'une notion donnée dans la base de données.

Cette fiche pivot ainsi que chacune des fiches traitant d'une même notion dans chaque langue reçoivent le même numéro matricule. À l'aide de ce numéro, les fiches physiques dans toutes les langues sont mises ensemble pour l'affichage de la fiche de terminologie complète. Il est également possible de n'afficher que les fiches existant dans les combinaisons de langues intéressant l'utilisateur. On peut remplir n'importe quelle combinaison de fiches linguistiques et on peut sélectionner n'importe quelle combinaison de langues lors de l'affichage. La fiche pivot est toujours présente.

On pourrait penser que la génération de la fiche logique lors de l'affichage nécessiterait un temps considérable. De fait, l'extraction de fiches à l'aide d'une clé d'accès unique est une tâche que le logiciel accomplit de façon remarquablement rapide et il n'existe aucun délai à l'affichage des fiches dû à la nécessité d'établir la fiche logique.

Bien que cette approche permette techniquement de répondre aux besoins des clients, les défis méthodologiques attendent toujours les terminologues de la DTD, étant donné que la plupart d'entre eux travaillent uniquement en anglais et en français et qu'on ne peut s'attendre à ce qu'ils puissent réellement travailler dans toutes les langues que TERMIUM® peut

accueillir. (Le choix des langues traitées est déterminé par le jeu de caractères disponibles dans TERMIUM® et qui correspond à peu près au jeu de caractères Latin – 1. À l'aide de ce jeu de caractères, on peut emmagasiner des données dans les langues de l'Europe du Nord et de l'Ouest.) Les terminologues devront alors consulter des experts dans d'autres langues pour s'assurer que la même notion est traitée correctement dans toutes les langues présentant des liens entre elles. De plus, lorsque la DTD recevra des fiches multilingues provenant d'autres organismes, il faudra comparer celles–ci avec des fiches déjà emmagasinées dans TERMIUM® avant de les ajouter à la base de données. Un groupe de travail de la DTD est chargé d'étudier les besoins méthodologiques et d'autres aspects du travail terminologique dans un contexte multilingue.

1.2.3 Tiroirs

Lors de la conception du logiciel, un autre défi intéressant a consisté en la création de subdivisions de la banque de données dans lesquels des organismes autres que la DTD puissent emmagasiner et gérer des données terminologiques multilingues. De plus, ces subdivisions, appelées « tiroirs », fournissent un moyen aux organismes et aux particuliers de partager leurs renseignements terminologiques à l'échelle mondiale en ayant accès au réseau existant de diffusion de TERMIUM®.

Les tiroirs doivent être dotés de diverses cotes de sécurité selon les exigences des organismes qui les exploitent. Dans certains cas, l'organisme peut disposer de données qu'aucun autre client de TERMIUM® ne devrait avoir la permission de lire ou de modifier. Il était aussi nécessaire de s'assurer que chaque tiroir puisse être aisément reconnu et traité comme une entité distincte, surtout qu'il existe une possibilité élevée qu'il y ait emmagasinement de fiches en double ou de fiches présentant des contradictions entre elles.

L'équipe de mise à niveau a étudié diverses options pour la mise en place des tiroirs et en a rejeté plusieurs, généralement pour des raisons de performance ou de complexité. La solution retenue consiste en la création de bases de données séparées

pour contenir les tiroirs. La structure des données dans chaque tiroir est presque identique à celle de la Base de données linguistiques, soit la partie de TERMIUM® que gère la DTD. De fait, un tiroir est une série de types de fiches, y compris la fiche pivot et les fiches propres à chaque langue. Les noms donnés aux types de fiches réfèrent au tiroir dont elles font partie. Dans le cas des types de fiches linguistiques, seules les fiches pertinentes au travail du propriétaire de tiroirs doivent être définies.

La validation dans les tiroirs est moins rigoureuse que celle de la Base de données linguistiques. Elle peut aussi être adaptée aux besoins des propriétaires de tiroirs. Par exemple, si le propriétaire d'un tiroir possède son propre système de classement de domaines, la validation peut être conçue de telle façon que seuls les codes des domaines permis puissent être emmagasinés. On peut également effectuer certains changements de la configuration, comme la définition de champs additionnels.

La saisie des données dans les tiroirs se fait généralement en lots par l'emmagasinement d'un fichier séquentiel adéquatement formaté ou encore de façon interactive ou par l'intermédiaire de la DTD. Grâce à l'existence des menus et du profil de l'usager, on peut contrôler l'accès aux tiroirs à des fins d'interrogation.

On n'a pas encore pu se pencher sur les aspects administratifs reliés à l'offre de tiroirs. Les enjeux méthodologiques doivent également faire l'objet d'un examen approfondi. Un projet pilote sur l'utilisation des tiroirs débutera bientôt.

Lors de la première phase du nouveau développement de TERMIUM®, l'équipe de mise à niveau a relevé les défis décrits ici ainsi que bien d'autres. Durant la seconde phase, elle mettra au point et complétera les caractéristiques et les fonctionnalités requises pour TERMIUM®. De plus, l'équipe procédera à l'analyse des nouvelles exigences des utilisateurs et à la mise en place de nouvelles fonctionnalités au besoin. Ainsi, elle se penchera sur les problèmes propres à l'emmagasinement, à la gestion et à l'extraction de données phraséologiques et linguistiques ; à l'emmagasinement et à l'exportation des images ;

ainsi qu'à l'emmagasinement, à la gestion et à l'extraction des données plein texte.

2 LATTER, poste de travail du terminologue

La DTD a créé LATTER en vue de répondre aux besoins de rationalisation des ressources et de dégraissage du travail nécessaire à la création de produits terminologiques. LATTER fournit aux terminologues le moyen de rassembler, de stocker, de partager, d'analyser et de synthétiser les renseignements terminologiques en vue de faciliter et d'accélérer l'entrée des fiches dans TERMIUM® ainsi que la production de lexiques et de vocabulaires. En définitive, le poste de travail est un ensemble intégré d'outils nécessaires tant à la recherche terminologique qu'à la gestion des données en vue de répondre aux exigences fonctionnelles de terminologues professionnels.

2.1 Structure et fonctionnalités

À ce stade, le poste de travail du terminologue consiste en de nombreuses applications tournant indépendamment les unes des autres, rassemblées sur un micro–ordinateur (386 ou une version plus récente). L'installation de Microsoft Windows[3] a aidé à l'intégration de ces applications, qui comprennent WordPerfect[3] (logiciel de traitement de textes), un logiciel de communication appelé PROCOMM PLUS[3], FASTBACK PLUS[3] (utilisé pour la sauvegarde), et le logiciel LATTER lui–même.

L'application LATTER comprend un système de base de données terminologiques locale, doté d'une gestion utile et de capacités d'échanges, et dont l'interface est conçue pour un environnement à fenêtres. La structure des données du poste de travail correspond de près à celle de TERMIUM®, en partie pour tenir compte des méthodes habituelles de travail et en partie pour s'assurer que les échanges de données avec la banque centrale en sont facilités. Un certain nombre d'éléments de données propres à LATTER sont aussi définis. Ils sont requis

---

[3] Windows, WordPerfect, PROCOMM PLUS and FASTBACK PLUS sont les marques de commerce de leurs propriétaires respectifs.

pour une gestion efficace des fiches LATTER pendant que les données sont recueillies, analysées et emmagasinées dans le poste de travail.

La conception de la fiche LATTER est dynamique et très flexible. On peut créer des fiches unilingues, bilingues et multilingues. La fiche d'entrée s'ajuste au fur et à mesure que les données sont saisies. On peut ainsi gérer des champs de longueur indéfinie et répéter autant de champs que nécessaire. Il est possible d'effectuer certaines opérations, comme l'impression et la suppression de données, tant sur des fiches individuelles que sur des ensembles de fiches. La création et le maniement d'ensembles de fiches sont des caractéristiques particulièrement intéressantes de LATTER.

2.2 Échanges de données

L'emmagasinement de fiches peut se faire de façon interactive ou en lots. Jusqu'à maintenant, trois formats d'importation de données ont été mis au point pour l'emmagasinement en lots. L'un est destiné à l'importation de fiches extraites de TERMIUM®, tandis qu'un autre est utilisé pour l'importation des données mises au point dans d'autres bases de données LATTER (situées soit sur le même poste de travail ou sur le poste de travail d'un autre terminologue). Le troisième format sert à l'importation de données provenant d'une variété de sources. Des caractères délimiteurs marquent le début d'une nouvelle fiche et le début d'un nouveau champ. Des codes mnémoniques permettent de repérer les champs où des données ont été emmagasinées. Ce format est habituellement utilisé pour l'importation de fiches créées à l'aide de WordPerfect. On peut réviser les fiches, les traiter et en terminer la rédaction. Il est aussi possible de convertir des données provenant d'autres systèmes de gestion terminologique sur micro–ordinateur en un format d'importation dans LATTER (et finalement dans TERMIUM®), dans la mesure où les données sont disponibles en fichiers ASCII séquentiels dans lesquels les éléments de données ont été clairement identifiés.

On peut également copier des fiches LATTER dans des fichiers en vue de leur exportation dans une autre base de

données LATTER ou dans TERMIUM®. Le format de la fiche ASCII créée pour l'envoi des fiches à TERMIUM® peut servir au transfert des données à d'autres systèmes de gestion des données terminologiques, même s'il faut effectuer un peu d'éditique à cette fin.

L'interface utilisateur LATTER, à menus, présente un judicieux mélange de fenêtres qui s'ouvrent au moment voulu, de boîtes de dialogue et d'opérations fonctionnelles. Les données relatives à l'affichage, aux défauts et aux préférences d'interrogation et d'importation/exportation sont emmagasinées sur chaque copie de l'application et sont modifiables par l'utilisateur.

2.3 Réactions des utilisateurs

Durant un an, dix terminologues ont utilisé la Version 1 de LATTER dans un contexte de production pour divers types de travaux. L'utilisation du poste de travail sur une base régulière a mis en lumière certaines instabilités du logiciel ainsi que les forces de LATTER et sa pertinence pour certains types de travaux. En vue de tenir compte des diverses expériences d'utilisation de LATTER et pour donner suite aux commentaires et aux suggestions de ce groupe d'utilisateurs, on a priorisé les améliorations à effectuer pour la Version 2. Des versions à jour du logiciel utilisé pour la création de LATTER furent incorporées dans l'application, ce qui a contribué à résoudre la plupart des instabilités notées. On a également effectué des changements à la structure des données et aux fonctionnalités d'exportation des données afin qu'il y ait compatibilité entre LATTER et TERMIUM® IV. Le même groupe d'utilisateurs vient de recevoir la dernière version de LATTER. Un plus grand nombre de terminologues commenceront à utiliser peu à peu le poste de travail.

Lors de la première diffusion de LATTER, les terminologues utilisaient le poste de travail principalement pour emmagasiner des fiches prêtes à être saisies dans TERMIUM®. Au fur et à mesure qu'ils se sont familiarisés avec les diverses fonctionnalités du logiciel, ils ont trouvé des façons d'en faire usage à d'autres fins dans leur travail. Par exemple, des ensembles de

fiches TERMIUM®, dont les terminologues ont la responsabilité, ont été importés dans les bases de données LATTER pour révision. Selon le résultat de l'analyse effectuée, les fiches individuelles ont été modifiées et retournées à TERMIUM® ou annulées dans TERMIUM®. Certains terminologues ont commencé à effectuer une mise à jour systématique de TERMIUM® à l'aide de LATTER en extrayant de la banque centrale les résultats de l'interrogation de TERMIUM® par liste de termes qu'ils ont ensuite importés dans LATTER. Ils ont alors procédé à la révision, à la comparaison des fiches et, si nécessaire, à leur modification avant de les réexporter dans TERMIUM®.

L'exemple ci–après démontre que LATTER est un véritable outil de gestion de la terminologie. L'un des terminologues de la DTD a mis au point un outil résultant de la combinaison de la technique du balayage et de la reconnaissance optique de caractères, de l'utilisation des macro–instructions WordPerfect et de LATTER en vue du dépouillement terminologique et, en définitive, de la création de fiches destinées à TERMIUM®. Des textes présentant un intérêt certain ont été ainsi dépouillés et les fichiers ASCII qui en ont résulté ont été importés dans WordPerfect. Le terminologue a marqué à la main les termes pour lesquels des fiches LATTER devaient être créées, soit manuellement avant le balayage soit à l'aide de WordPerfect après le balayage. Il a ensuite utilisé les macros pour extraire chaque terme, son contexte d'utilisation et la source et les présenter sous le format d'importation de LATTER. Les fiches ont alors été entrées dans la base de données LATTER du terminologue où elles ont fait l'objet de recherches complémentaires en vue de leur éventuelle exportation dans TERMIUM®. Cette approche est maintenant utilisée par d'autres terminologues.

Les terminologues ont eu des réactions diverses devant LATTER. Au début, les problèmes que présentaient le matériel et le logiciel engendraient un certain découragement. Une fois ceux–ci plus ou moins résolus, la rétroaction fut meilleure, passant des plaintes au sujet de la perte de données à des requêtes en vue d'améliorations concrètes à l'interface utilisateur et aux fonctionnalités de l'application. On a perçu ces requêtes comme étant positives puisqu'elles démontraient

l'intérêt des terminologues à poursuivre l'utilisation du poste de travail. Un commentaire qui revenait fréquemment avait trait à la nécessité d'accélérer certaines opérations (par exemple, la procédure de sauvegarde d'une fiche ou le passage d'une fiche à l'autre à l'intérieur d'un ensemble de fiches).

De façon générale, la réaction des terminologues a varié selon la nature du travail effectué. Dans un cas, une terminologue a utilisé LATTER uniquement pour finaliser des fiches destinées à TERMIUM® et a trouvé que, comparé à l'utilisation d'un format d'entrée spécialement conçu à cette fin dans WordPerfect, LATTER était plus lent, plus complexe à utiliser et présentait des déficiences au chapitre de l'édition. Les forces de LATTER n'ont pu être mises en lumière étant donné que cette terminologue n'avait pas besoin de gérer les données. Aussi a–t–elle décidé de ne pas utiliser l'application. Par contre, les terminologues qui utilisent LATTER pour la collecte, l'organisation et l'analyse des résultats de leurs recherches en vue de la création d'une fiche destinée à TERMIUM®, sont enchantés des possibilités qu'offre LATTER. Leur degré de satisfaction face à l'outil a augmenté au fur et à mesure qu'ils ont mieux compris les fonctionnalités de LATTER et ils ont modifié leurs méthodes de travail en conséquence.

De fait, le logiciel LATTER représente davantage qu'une base de données terminologiques. La DTD prévoit intégrer d'autres outils dans l'application de telle façon que le poste de travail puisse être utilisé à des stades antérieurs du processus de recherche terminologique. L'automatisation de l'extraction de termes lors du dépouillement et de l'organisation notionnelle, par exemple, constitueront des améliorations appréciables. La capacité de recueillir et de gérer d'autres types de données, d'ordre linguistique et phraséologique, revêt une importance accrue. On effectuera bientôt une analyse des besoins et des possibilités de solution dans ce domaine.

Dans l'intervalle, l'intégration du poste de travail dans la chaîne de travail des terminologues de la DTD se poursuit. De plus, étant donné que LATTER a été conçu comme outil de gestion des données terminologiques, les collaborateurs de TERMIUM® en perçoivent l'utilisation comme outil de maintien

de leur propre base de données terminologiques, laquelle peut être emmagasinée dans un tiroir de TERMIUM®, ce qui permet l'échange de l'information.

3 Conclusion

En vue de suivre les innovations technologiques qui se produisent rapidement et de répondre aux exigences et aux attentes des clients, la DTD continue à développer et à raffiner des applications destinées aux spécialistes de la langue. En améliorant les outils utilisés par les terminologues et en encourageant la participation des organismes partenaires, la DTD s'assure que le contenu de ses produits terminologiques est à jour et présente autant d'intérêt que les outils automatisés utilisés dans leur production.

Remerciements

L'auteure désire remercier les personnes qui ont fait des commentaires sur ce texte, principalement Helen Hutcheson et Diane Michaud. Elle remercie également Michèle Valiquette pour la traduction vers le français et Silvia Pavel pour avoir accepté de présenter sa communication au 62e Congrès de l'ACFAS.

Bibliographie

Leonhardt, Christine, 1992, LATTER, The Terminologist's Workstation, In *Proceedings from the International Symposium on Terminology and Documentation in Specialized Communication*, Ottawa, Supply and Services, pp. 257–275.

Leonhardt, Christine, 1994, TERMIUM® and LATTER : An Update, In *Terminology Update*, Vol. 27,1, Ottawa, Supply and Services, pp. 15–18.

L'ANALYSE DE TRADUCTION ET L'AUTOMATISATION DE LA TRADUCTION

Pierre Isabelle, Marc Dymetman, George Foster
Jean-Marc Jutras, Elliott Macklovitch, François Perreault
Xiabo Ren, Michel Simard

Résumé

Nous avançons que le concept d'*analyse de traductions* peut servir de point de départ à une nouvelle génération d'aides à la traduction. Nous montrons que les traductions peuvent être analysées et versées dans une *mémoire traductionnelle structurée* et décrivons le concordancier bilingue TransSearch que nous avons mis au point pour permettre aux traducteurs d'exploiter cette mémoire traductionnelle. Nous affirmons que les analyseurs de traductions peuvent contribuer à la détection des *erreurs de traduction* dans les premiers jets et nous présentons les résultats d'une expérience portant sur la détection des faux amis, réalisée dans le cadre du projet TransCheck. Nous soutenons enfin que l'analyse de traductions peut faciliter la transcription directe de traductions dictées et présentons le nouveau projet TransTalk.

1 Introduction

En 1951, Y. Bar-Hillel, qui fut le premier chercheur à se consacrer à temps plein au domaine de la traduction automatique (TA), écrivait :

> *«Dans le cas des domaines cibles où la précision absolue est une condition sine qua non, il faut renoncer à la TA pure en faveur de la TA mixte, c'est-à-dire un processus traductionnel faisant intervenir l'intelligence humaine. Ce qui soulève la question suivante : Quelles étapes du processus devrait-on confier à un partenaire humain?»(Bar-Hillel (1951), p. 230)*

Quarante-deux ans et trois «générations» de systèmes plus tard, la TA pure n'est pas plus généralement applicable qu'elle ne l'était à cette époque[1]. Plus décourageant encore, la TA mixte ne l'est guère plus. Il n'existe pas de données précises aisément accessibles à ce sujet, mais on peut dire à coup sûr que la part actuelle de la TA, pure ou mixte, se situe bien en deça de 1 p. cent du marché global de la traduction. Force nous est donc d'en conclure que les chercheurs en TA n'ont pas encore réussi à proposer de réponses réalistes et pratiques à la question posée par Bar-Hillel au sujet de la division du travail entre l'homme et la machine.

Bar-Hillel lui-même proposa l'idée d'un tandem homme-machine dans lequel le partenaire humain interviendrait soit avant, soit après le processus mécanique, «mais de préférence pas quelque part au milieu». C'est donc dire que la machine se chargerait de l'essentiel du processus traductionnel. Depuis, la «TA assistée par l'homme» est restée le paradigme prédominant dans le milieu de la TA, où l'on a continué à demander aux machines d'accomplir une tâche qu'elles n'arrivent pas à bien faire, c'est-à-dire traduire. Et où l'on a continué à demander aux traducteurs d'exécuter des tâches auxquelles ils préféreraient se soustraire, comme insérer des codes bizarres dans des textes sources, répondre à des questions inattendues

---

[1] Même s'il a été démontré que la TA peut être remarquablement performante dans le cas plutôt marginal de certains sous-langages extrêmement restreints, comme celui des bulletins météorologiques (Isabelle 1987).

sur le parenthésage des syntagmes ou réorganiser d'étranges fouillis de mots en langue cible. Résultat : le marché n'a jamais manifesté beaucoup d'enthousiasme pour ce genre de *modus vivendi* homme-machine.

Il est devenu évident qu'en général, les **machines ne réussissent toujours pas à maîtriser la partie essentielle du processus de traduction**. Déjà, en 1980, Martin Kay (1980) plaidait en faveur d'un renversement des rôles qui aurait pour effet de remettre la machine à «sa place», c'est-à-dire celle d'assistant du traducteur humain :

> *«Je veux préconiser une approche au problème selon laquelle on permettrait à la machine de prendre en charge graduellement, presque imperceptiblement, certaines fonctions du processus général de traduction. La machine assumerait d'abord des tâches non essentiellement reliées à la traduction. Puis, peu à peu, elle s'attaquerait à la traduction comme telle. Tout serait une question de modestie. À chaque étape, on ne confierait à la machine que ce qu'elle sait bien faire. Et ainsi, petit à petit, l'oiseau ferait son nid!»(p. 11)*

C'est précisément pour cette approche réaliste et modeste que le Centre canadien de recherche sur l'informatisation du travail (qui porte maintenant le nom de Centre d'innovation en technologies de l'information - CITI) a opté, en 1987, lorsqu'il a lancé le projet de poste de travail de traducteur (Macklovitch 1993). Dans sa plus récente incarnation, le poste de travail du CITI met à la disposition du traducteur un environnement multi-fenêtres qui lui donne simultanément accès à plusieurs outils, comme le traitement de texte en écran partagé, la vérification orthographique, la consultation terminologique et lexicographique, la comparaison de fichiers, le compte de mots, l'extraction de textes intégraux, etc. (Macklovitch 1993). Il faut reconnaître que ces fonctions ont plus à voir avec la bureautique qu'avec l'automatisation de la traduction elle-même. Mais suivant le scénario proposé par Kay, nous pouvons maintenant tirer profit de cette base informatique en l'enrichissant progressivement d'outils orientés vers la traduction. Dans cette optique, la question centrale peut être formulée comme suit : **au-delà de la bureautique intégrée, mais en deça de la**

traduction automatique, que peut-on faire de plus pour faciliter la tâche du traducteur?

Nous avançons, dans la suite de cet article, que le concept d'*analyse de traductions* peut servir de fondement à l'élaboration d'une nouvelle génération d'outils informatisés à l'intention du traducteur. La deuxième section de cet article est consacrée à une description générale de la notion d'analyse de traductions. Les sections 3, 4 et 5 décrivent les travaux réalisés au CITI sur trois applications : une mémoire traductionnelle, un vérificateur de traductions et un système de dictée pour traducteurs.

2 L'analyse de traductions

Dans la documentation récente (Dymetman et Macklovitch 1988, par exemple), la traduction est souvent conceptualisée comme une relation $tr_{L1L2}(s, t)$, dont l'extension est un ensemble de paires $<S, T>$, où S est un texte en langue L_1 et T un texte en langue L_2. Comme il existe, dans chacune de ces langues, un nombre infini de textes, la relation tr_{L1L2} doit être définie de façon récursive, ce qui aura pour conséquence que cette relation aura un caractère compositionnel: jusqu'au niveau d'un certain ensemble fini d'éléments élémentaires, S et T seront décomposés respectivement en des ensembles d'éléments $\{s_1 , s_2 , ... , s_n\}$ et $\{t_1 , t_2 , ..., t_n\}$, de telle façon que pour tout i, la relation $tr_{L1L2} (S_i , T_i)$ sera également satisfaite.

Un système de TA ordinaire incorpore une spécification quelconque (éventuellement partielle) de la relation traductionnelle tr_{L1L2}, de même qu'un procédure qui produira, pour n'importe quelle valeur de S, une ou plusieurs valeurs T pour lesquelles $<S, T>$ appartiendra à tr_{L1L2}. Un système de TA réversible (voir, par exemple, Dymetman 1992, Van Noord 1993) peut en outre calculer, pour n'importe quelle valeur de T, les valeurs S pour lesquelles $<S, T>$ appartiendra à tr_{L1L2}.

Les systèmes de TA tentent de résoudre le problème de la production de traductions, mais, comme l'a fait remarquer Debili (1992), nous pouvons aussi envisager les traductions du point de vue de la reconnaissance. Nous appellerons donc

accepteur de traductions toute procédure qui, à partir d'une paire particulière de textes <*S, T*>, peut déterminer si la relation $tr_{L1L2}(S, T)$ est toujours vérifiée. Et nous appellerons *analyseur de traductions* toute procédure récursive *at*(<*S, T*>, *AAT*) qui attribue, aux paires <*S, T*> qui satisfont la relation $tr_{L1L2}(S, T)$, un arbre d'analyse traductionnelle *AAT*. Un AAT rend explicite la structure compositionnelle de la relation traductionnelle. Par exemple, moyennant une définition appropriée de la relation traductionnelle anglais-français, un analyseur de traductions pourrait produire un AAT comme celui qui est illustré à la Figure 1.

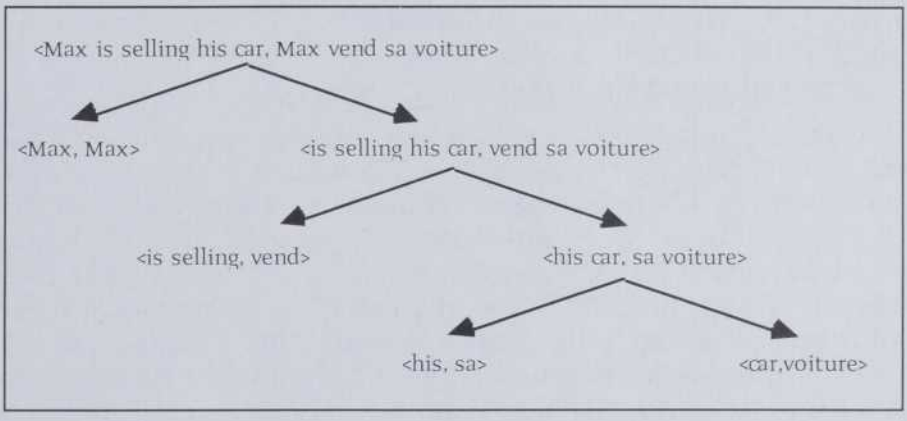


Figure 1 : Un arbre d'analyse traductionnelle (AAT)

Isabelle (1992) utilise le terme *bi-texte* pour désigner une structure qui, comme l'AAT, sert à décomposer les traductions en leurs correspondances constitutives. Les AAT sont des descripteurs structuraux des analyses de traductions au même titre que les arbres d'analyse ou de parsage sont des descripteurs structuraux des analyses grammaticales.

L'analyse de traductions et la TA posent, en principe, des problèmes très semblables : le calcul est basé sur la même relation abstraite tr_{L1L2}. La différence réside uniquement dans les modes de calcul. Cela signifie-t-il que les analyseurs de traductions et les systèmes de TA sont, en pratique, assujettis

exactement aux mêmes contraintes? Plus spécifiquement, cela signifie-t-il qu'il n'est possible de réaliser des analyseurs de traductions efficaces qu'à la seule condition qu'il soit également possible de réaliser des systèmes de TA efficaces?

Il est évident que non. Dans les rares cas où la TA de haute qualité est possible, il devrait évidemment être possible de construire un analyseur de traductions pour les sorties du système de TA. Dans les cas où la TA n'est pas possible, nous soutenons, et c'est ce qui importe, qu'il est malgré tout possible d'élaborer des dispositifs capables d'analyser les traductions réalisées par des humains et que ces analyseurs auront de nombreuses utilités. Cette différence découle des exigences pratiques que des tâches différentes (la TA par opposition à l'analyse de traductions) imposent au niveau de précision de la caractérisation formelle de la relation $tr_{L1L2}(S, T)$.

Considérons, par exemple, le modèle qui sous-tend la méthode d'alignement de phrases proposée par Brown et al. (1991). Sur le plan conceptuel, ce modèle génère des séquences de paires de textes $<S, T>$ qui présentent les caractéristiques suivantes : a) S est une séquence $<s_1 , s_2 , ..., s_n>$ dans laquelle chaque s_i est lui-même une séquence de zéro, une ou deux «phrases» et T est une séquence semblable $<t_1 , t_2 , ..., t_n>$; b) une «phrase» est une chaîne d'unités lexicales terminée par une unité de ponctuation ; c) une unité lexicale est une chaîne de caractères encadrée par des caractères délimitateurs ; d) la longueur $l(s_i)$ de chaque s_i (en termes du nombre d'unités contenues) présente une corrélation avec la longueur $l(t_i)$ de la «phrase» correspondante t_i, selon une distribution de probabilités $pr(l(s) \mid l(t))$ et e) cette distribution de probabilités peut être estimée à partir des fréquences observées dans des corpus de traductions, comme le corpus bilingue du Journal des débats de la Chambre des communes.

Ce modèle saisit bien l'un des aspects spécifiques de la relation traductionnelle établie entre deux langues, à savoir la corrélation de longueur qui existe entre les phrases qui sont des traductions réciproques. En ce sens, il constitue un modèle de traduction, aussi faible soit-il.

Si l'on appliquait un modèle de ce genre à la traduction anglais-français, une phrase anglaise *e* se traduirait plus ou moins par une séquence aléatoire de caractères *f*, dont la seule propriété notable serait d'avoir une longueur *l(f)* qui est typique de la traduction française d'un phrase anglaise de longueur *l(e)*. Dans la pratique, un tel «système de TA» semblerait parfaitement inutile.

Si, d'autre part, à l'instar de Brown et al., on applique ce modèle à l'analyse de traductions, on obtient un système capable de décomposer des traductions existantes en des représentations qui rendent leur structure compositionnelle explicite jusqu'au niveau de la phrase. Le résultat sera un arbre d'analyse ayant la forme illustrée à la Figure 2, où les textes S et T sont décomposés en n paires successives de blocs de phrases s_i et t_i.

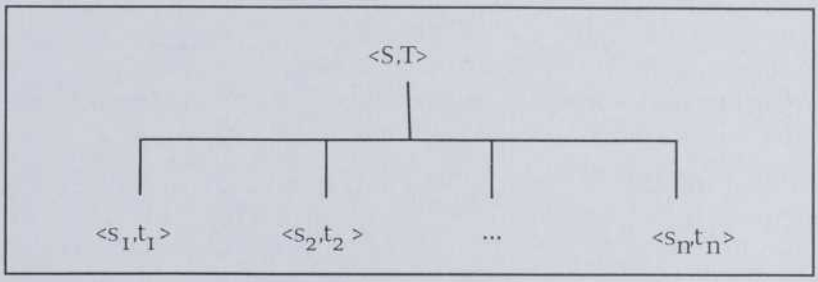


Figure 2 : L'alignement de phrases en tant qu'arbre d'analyse simple

L'analyse est, à l'évidence, très grossière : aucune correspondance n'est établie au-delà du niveau de la phrase. Pourtant, comme nous le verrons bientôt, ces bi-textes à «faible résolution» peuvent servir de base à certains outils informatisés très pratiques.

Évidemment, des analyses plus raffinées offriraient encore plus de possibilités à cet égard. En fait, il n'est pas trop difficile d'imaginer des familles de modèles de traduction un peu plus puissants qui, tout en demeurant insuffisants pour les fins de la TA, pourraient néanmoins être utilisés pour expliciter

davantage de structures dans des traductions existantes (comme les correspondances entre syntagmes ou mots).

Pour ce qui est de leur architecture générale, les modèles utilisés pour l'analyse de traductions peuvent être très proches de ceux utilisés pour la TA. La possibilité la plus évidente est peut-être le modèle tripartite illustré à la Figure 3.

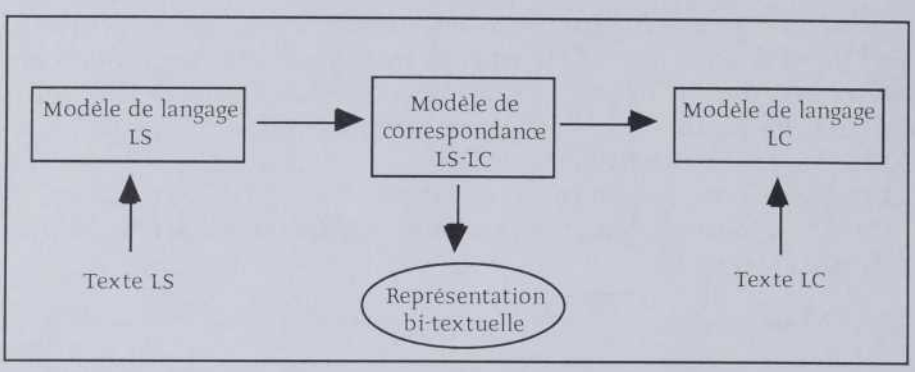


Figure 3 : Un modèle tripartite pour l'analyse de traductions

Comme le modèle de transfert bien connu de la TA, ce modèle comprend deux composantes propres à chacune des langues (les modèles de langage) et une composante «contrastive», propre aux paires (le modèle de correspondance). Les deux composantes unilingues fonctionnent en mode analytique et les représentations linguistiques qu'elles produisent sont traitées par le modèle de correspondance, qui les relie en une représentation bi-textuelle unique où les correspondances traductionnelles sont rendues explicites. Que ses composantes soient mises en oeuvre au moyen de techniques à base de règles ou à base de corpus, ce modèle demeure un modèle naturel. En fait, la meilleure façon de conceptualiser la méthode d'alignement simple basée sur la longueur des phrases, que nous avons déjà mentionnée, est de l'envisager comme une instance de ce modèle.

Certaines indications suggèrent que les modèles probabilistes se révéleront extrêmement utiles pour l'élaboration d'analyseurs de traductions d'utilité générale. Tandis que les

méthodes à base de règles conviennent bien à l'élaboration de modèles «profonds» dans des domaines restreints, les méthodes probabilistes semblent particulièrement bien adaptées au développement de modèles superficiels potentiellement capables de produire des analyses partielles relativement précises de traductions générales.

Quoi qu'il en soit, l'essentiel de notre propos est que l'analyse de traductions, même basée sur des modèles de traduction faibles, constitue un point de départ approprié pour le développement d'une nouvelle génération d'aides à la traduction. Nous examinerons maintenant certains de ces outils.

3 La mémoire traductionnelle

3.1 Les traductions existantes en tant que ressource

La tendance aux approches à base de corpus en TA découle en partie de la constatation que le fonds de traductions existantes constitue une ressource d'une très grande richesse dont le potentiel n'a pas encore été pleinement exploité. En fait, il est évident que les **traductions existantes renferment infiniment plus de solutions à plus de problèmes de traduction que tout autre outil de référence.**

Mais les traducteurs ne pourront exploiter les richesses enfouies dans leurs traductions antérieures que lorsqu'ils disposeront des outils leur permettant de les gérer comme des données de traduction plutôt que comme des données de traitement de texte. Et c'est précisément à cela que sert un analyseur de traductions : transformer des données de traitement de texte en des structures bi-textuelles qui rendent explicites les correspondances traductionnelles.

Une fois les traductions existantes structurées en bi-textes, les segments correspondants en langue source et en langue cible sont systématiquement interreliés. Plus particulièrement, tout segment qui renferme une occurrence d'un problème de traduction est relié à un segment qui renferme une solution toute faite à ce problème. S'ils disposent des moyens nécessaires

pour créer, stocker et interroger de telles structures bi-textuelles, les traducteurs pourront transformer leur production antérieure en une mémoire traductionnelle exploitable et extrêmement efficace.

3.2 TransBase

Pour rendre accessibles les résultats des analyses traductionnelles de grandes quantités de texte, nous avons conçu un modèle simple de mémoire traductionnelle structurée, que nous avons appelé TransBase. Ce modèle possède les mêmes caractéristiques de base que les systèmes d'extraction de textes intégraux : il peut gérer des quantités arbitraires de texte, il peut être augmenté de façon incrémentielle et il assure un accès rapide au contenu textuel de la base de données. Ce qui le distingue essentiellement de ces systèmes, c'est sa capacité à stocker des représentations bi-textuelles.

Une base de données TransBase se construit à l'aide d'un analyseur de traductions semblable à celui qui est illustré à la Figure 2. Chacun des textes d'une paire de traductions réciproques fait l'objet d'une analyse linguistique qui le décompose en ses éléments structuraux (paragraphes, phrases, etc.) et détermine son contenu lexical. Cette information est stockée dans deux composantes distinctes de la base de données, propres à chacune des langues en présence, et indexée de façon à permettre l'accès rapide à n'importe quelle partie du texte. Un «analyseur de correspondances», basé sur les techniques décrites dans Simard, Foster et Isabelle (1992), utilise ensuite ces analyses linguistiques pour construire une «carte traductionnelle» au niveau de la phrase, qui est également stockée dans la base de données. Cette structure et le mode de construction de la base de données sont illustrés à la Figure 4.

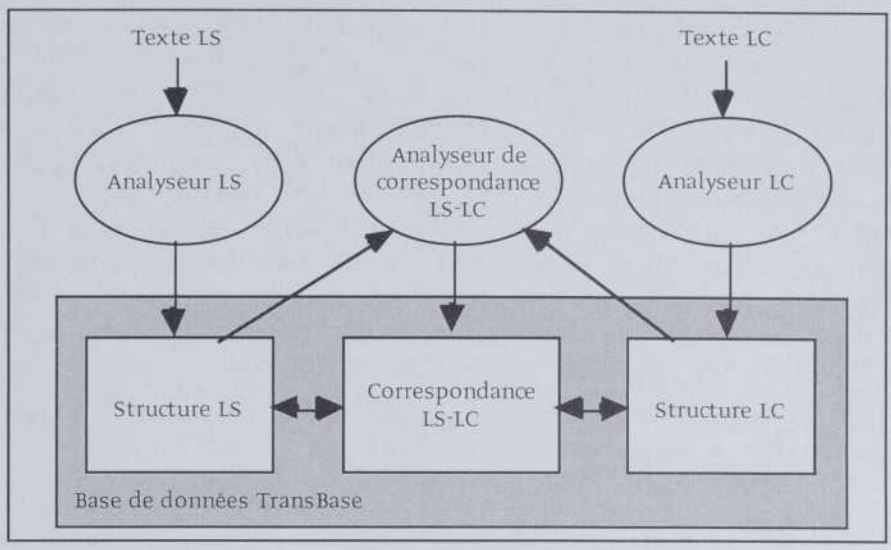


Figure 4 : Structure générale de la base de données TransBase

Dans la base de données, les textes source et cible sont traités de façon symétrique. Cependant, comme la directionalité de la traduction peut être importante pour l'utilisateur, TransBase peut enregistrer laquelle des deux langues est la langue source.

3.3 TransSearch

Il existe de nombreuses façons d'exploiter cette mémoire traductionnelle. La première qui vient à l'esprit, et celle qui est probablement la plus universellement utile, consiste à fournir aux traducteurs des outils qui leur permettront d'interroger le contenu textuel de la base de données TransBase. Certains auteurs (voir, par exemple, Church et Gale 1991) ont déjà suggéré qu'un outil capable de produire des *concordances bilingues* serait utile aux lexicographes bilingues. Il est plutôt évident qu'un concordancier bilingue serait également utile aux traducteurs. Il se peut, par exemple, qu'en rencontrant, dans un texte de départ anglais, un idiotisme tel que *to be out to lunch* ou *to add insult to injury*, le traducteur ne soit pas

certain de l'équivalent français approprié. Il se peut aussi qu'il ne trouve pas de réponse satisfaisante dans les dictionnaires bilingues traditionnels. S'il disposait d'un concordancier bilingue, il pourrait alors interroger une base de données bitextuelles du genre de TransBase et en extraire des exemples de ces expressions, accompagnées de leurs traductions. Cela serait pratique non seulement pour les idiotismes, mais aussi pour la terminologie spécialisée ou pour les tournures et formules propres à certains domaines (*To whom it may concern..., Attendu que...*). On trouvera, dans Macklovitch (1992), un exposé plus détaillé de cette question.

Le logiciel TransSearch est justement cet outil : il permet d'extraire de la base de données des occurrences d'«expressions» précises et de les afficher à l'intérieur de leur contexte bilingue. Étant destiné principalement aux traducteurs, qui sont susceptibles de l'utiliser simplement comme source de référence supplémentaire, ce logiciel est conçu pour être exploité en mode interactif et pour donner des réponses en temps réel, ce qui est assuré par l'inclusion d'index de mots dans le modèle TransBase.

Comme la plupart des traducteurs ne sont pas des experts en informatique, nous avons accordé beaucoup d'attention à la convivialité de l'interface de TransSearch. En utilisant un langage d'interrogation intuitif et à orientation graphique, il est facile pour l'utilisateur d'effectuer des recherches complexes dans la base de données. Chacune de ces interrogations définit une expression logique portant sur des séquences de mots : lorsque l'interrogation est lancée, le système extrait de la composante alignement de la base de données tous les couples qui correspondent à cette expression. L'inclusion de dictionnaires et de descriptions morphologiques du français et de l'anglais permet en outre à TransSearch de repérer automatiquement les formes fléchies des éléments recherchés.

Les résultats d'une interrogation sont normalement présentés en deux colonnes, les traductions réciproques étant affichées côte à côte. L'utilisateur peut alors examiner chacune des solutions proposées à l'intérieur du document dont elle a été extraite, ou rassembler toutes les solutions repérées accompa-

gnées d'une petite portion de leur contexte immédiat, selon le mode de présentation habituel des concordances.

La Figure 5 présente un exemple type des résultats obtenus avec TransSearch. Dans cet exemple, l'utilisateur a interrogé le système pour trouver des occurrences de l'expression anglaise *to take X to court* qui ne sont pas traduites en français par *poursuivre X* ou *intenter un (ou des) procès à X* et la base de données interrogée était constituée des traductions du Journal des débats de la Chambre des communes de 1986. Tous les traducteurs à qui nous avons montré le fonctionnement de ce système en ont conclu qu'un concordancier bilingue leur serait très utile.

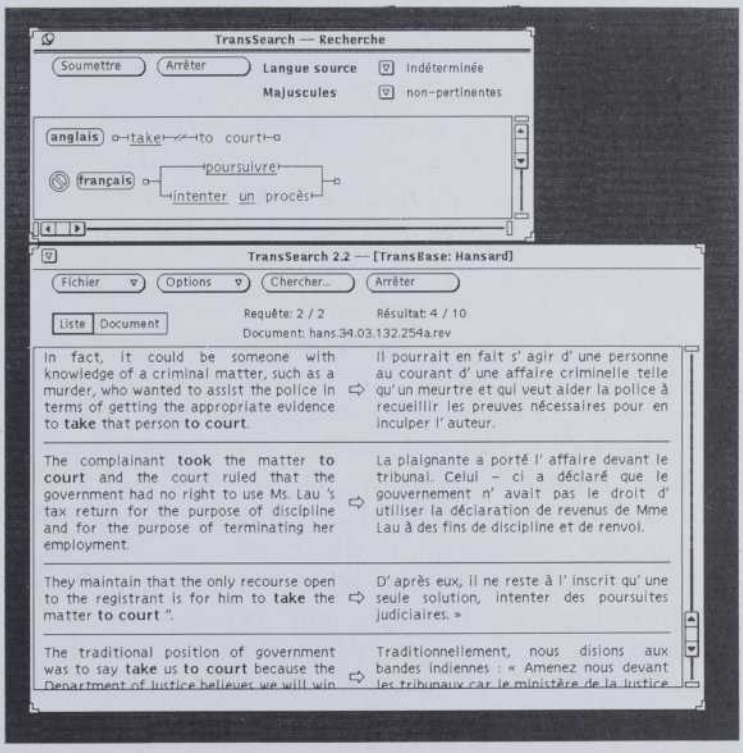


Figure 5 : Exemple des résultats obtenus avec TransSearch

4 La vérification de traductions

4.1 L'analyse de traductions et la détection d'erreurs

Depuis quelques années, on voit apparaître sur le marché du logiciel des outils critiques conçus pour aider les rédacteurs à améliorer leurs textes, grâce à la détection des problèmes potentiels d'orthographe, de grammaire et même de style. Si ces outils peuvent, en principe, aider les traducteurs à corriger les erreurs de rédaction contenues dans leurs traductions, ils ne peuvent aucunement les aider à corriger les *erreurs de traduction* au sens strict du terme, c'est-à-dire les correspondances erronées entre le texte source et le texte cible. Par exemple, ils ne peuvent pas les aider à repérer les *contresens* ou les *faux sens*, c'est-à-dire les cas où les deux textes sont individuellement corrects et signifiants, mais ne sont pas, en l'occurrence, des traductions réciproques. De telles erreurs ne peuvent être détectées que par un dispositif qui examine simultanément le texte source et le texte cible : en d'autres termes, un dispositif qui incorpore un analyseur de traductions.

Moyennant un analyseur de traductions capable de reconstruire un sous-ensemble quelconque C_{set} des correspondances observables dans les résultats d'une opération de traduction, et moyennant un ensemble quelconque de contraintes C auxquelles doivent satisfaire les correspondances admissibles, un vérificateur de traductions est un dispositif qui aide le traducteur à s'assurer que le sous-ensemble C_{set} respecte effectivement les contraintes C. Cette capacité exige la présence d'un analyseur de traductions basé sur un modèle de traduction «robuste», un modèle qui est capable de détecter les correspondances réelles susceptibles de dévier de la norme définie par C.

Le problème général de la qualité des traductions est une question manifestement complexe et frustrante. Nous n'avons certainement pas l'intention de proposer une mesure ou une méthode globale pour évaluer les traductions. Notre but est plus modeste : nous désirons seulement cerner certaines

caractéristiques particulièrement simples qui sont recherchées par la plupart des traducteurs et mettre au point quelques outils qui les aideront à vérifier si leurs traductions possèdent ces caractéristiques.

Une première caractéristique souhaitable et plutôt évidente est l'*exhaustivité*. En effet, toutes les parties d'un texte source devraient normalement avoir un équivalent dans le texte cible. Mais il arrive parfois aux traducteurs de faire des *erreurs d'omission*, en oubliant par exemple de traduire une phrase, un paragraphe, voire un page entière. Dans de tels cas, un analyseur de traductions efficace devrait être en mesure d'établir qu'un segment du texte source est mis en correspondance avec un segment vide dans le texte cible. Le dispositif de vérification pourrait alors en informer le traducteur, en lui signalant l'existence d'un problème éventuel dans son premier jet.

Une deuxième caractéristique candidate est la *cohérence ou l'uniformité terminologique*. En traduction technique, il est de rigueur d'utiliser systématiquement le même terme pour traduire toutes les occurrences d'un terme particulier du texte source. Un processus d'analyse permettant de faire ressortir toutes les correspondances terminologiques entre une traduction et sa source devrait vraisemblablement aider les traducteurs à respecter le principe de l'uniformité terminologique.

Autre caractéristique : les traductions sont censées être exemptes *d'interférences linguistiques* provenant de la langue source. Dans certains cas, ces interférences mènent à des constructions boiteuses en langue cible, que l'on peut souvent détecter sans même se référer au texte source. Par exemple, si le mot anglais *address* est traduit en français par *addresse* (avec deux d), un vérificateur orthographique ordinaire devrait être mesure de détecter cette erreur. Mais il y a aussi des cas où l'interférence mène non pas à des mots mal orthographiés, mais bien à des erreurs de traduction. Les *faux amis*, par exemple, ont tendance à provoquer ce genre d'interférence.

Le mot m_e de la langue L_e et le mot m_f de la langue L_f sont des *mots apparentés* lorsque leur forme est semblable en raison d'une étymologie commune. C'est le cas, par exemple, du mot anglais *government* et du mot français *gouvernement*. Le plus

souvent, ces mots sont non seulement des homonymes translinguistiques, mais aussi des synonymes. Dans certains cas, cependant, il n'y a pas de synonymie. Par exemple, les mots apparentés anglais/français suivants ont des sens complètement différents : *<actual / actuel>*, *<library / librairie>*, *<physician / physicien>*. Ces mots apparentés sont dits des «faux amis» parce que leur ressemblance morphologique crée une attente sémantique qui peut induire en erreur. La phrase *Max se rendit à la librairie* est parfaitement correcte en français, mais comme traduction de *Max went to the library*, elle constitue un cas flagrant d'erreur de traduction. Si un analyseur de traductions était capable de reconnaître, dans une traduction, une correspondance établie entre des mots apparentés reconnus comme des faux amis, il pourrait marquer cette correspondance comme une erreur possible que le traducteur pourrait ensuite vérifier.

Il existe probablement plusieurs autres types d'erreurs de traduction qu'un analyseur de traductions pourrait aider à déceler. La recherche dans ce domaine ne fait que commencer. Afin d'avoir un meilleur aperçu du potentiel pratique de cette approche, nous avons réalisé une expérience portant sur la détection des faux amis dans des traductions réelles.

4.2 Une expérience portant sur la détection des faux amis

Les faux amis (FA) peuvent être subdivisés en faux amis *absolus* ou *partiels*. Les FA absolus, comme ceux qui sont cités dans les exemples précédents, se caractérisent par le fait que leurs significations sont complètement disjointes ; ils ne peuvent donc jamais être utilisés comme des traductions réciproques. Les FA partiels, par contre, ont des sens qui se recoupent partiellement et ils peuvent être des équivalents traductionnels dans un sous-ensemble particulier de leurs emplois possibles. Par exemple, le verbe français *examiner* est parfois l'équivalent (≡) et parfois le non-équivalent (≠) du verbe anglais *to examine* :

The doctor examined his patient ≡ *Le médecin examina son patient*

The professor examined his students ≠ Le professeur examina ses étudiants

En nous concentrant dans un premier temps sur le problème plus facile des FA absolus, nous avons réalisé une expérience visant à évaluer 1) l'ampleur du problème dans des traductions réelles et 2) l'efficacité de certaines méthodes de détection simples.

Nous avons élaboré un analyseur de traductions AT1 qui instancie le modèle de la Figure 1 de la façon suivante : les modèles de langage pour le français et l'anglais sont réduits à des processus de segmentation en mots (*tokenisation*) et d'analyse morphologique (basés sur un dictionnaire et sur un ensemble de règles de flexion). Ces modèles de langage produisent une représentation morphologique simple du texte d'entrée : chaque unité lexicale est représentée comme l'ensemble des formes des entrées lexicales dont elle est une instance possible. Le modèle de correspondance utilisé dans AT1 est simplement le programme d'alignement de phrases mis au point par Simard, Foster et Isabelle (1992). La représentation qu'il produit est une séquence $<<e_1 , f_1>, <e_2 , f_2>, ...<e_n , f_n>>$, dans laquelle chaque e_i est une séquence de zéro, une ou deux phrases du texte anglais représentées morphologiquement, chaque f_i est une séquence de zéro, une ou deux phrases du texte français représentées morphologiquement, et chaque $<e_i , f_i>$ est une correspondance traductionnelle.

Nous avons extrait de van Roey et al. (1988) une liste de 145 paires de mots classifiés comme des faux amis absolus : *<accommodate / accommoder>*, *<actually, actuellement>*, etc.[2]. Nous avons ensuite réalisé un vérificateur simple qui a pour fonction d'examiner les résultats produits par AT1 et d'identifier, pour chaque paire de mots $<M_e , M_f>$, l'ensemble de paires de phrases $<e_i , f_i>$ satisfaisant la condition $M_e \in e_i$ (c'est-à-dire que e_i contient le mot m_e) et $M_f \in e_f$. Cette condition peut

[2] Nous ne savons pas encore quelle proportion du problème réel est prise en compte par ces 145 paires, mais nous avons de fortes de raisons de soupçonner qu'il s'agit seulement de la pointe de l'iceberg.

évidemment être satisfaite par des paires de phrases où m_e et m_f sont présents sans toutefois être utilisés comme des traductions réciproques.

Nous avons ensuite testé ce dispositif relativement simpliste sur un corpus composé des traductions du Journal des débats de la Chambre des communes couvrant une période d'un an. Après vérification manuelle des résultats, nous avons constaté qu'un grand nombre d'erreurs de traduction réelles avaient été détectées, comme dans l'exemple suivant :

> The peace movement in Canada is composed of **physicians**, members of the church, [...]
>
> _ Le mouvement canadien pour la paix compte dans ses rangs des **physiciens**, des ecclésiastiques, [...]
>
> (Journal des débats, 1987/09/29)

> There are parts of this bill which concern **librarians** and the artistic community.
>
> _ Quelque part dans ce projet de loi, il est question des **libraires** et des artistes.
>
> (Journal des débats, 1987/11/30)

Mais, comme on peut le voir dans le Tableau 1, les résultats présentaient aussi un niveau de «bruit» très élevé.

	Nombre de cas	Pourcentage
Erreurs réelles	57	7,4
Bruit	718	92,6
Total	775	100

Tableau 1 : Résultats de l'extraction de FA dans les sorties de AT1

Ce bruit provenait de trois sources différentes. Premièrement, il y avait des cas où la «fausseté» (par allusion à faux amis) de $<M_e , M_f>$ était attribuable à la catégorie grammaticale (POS) des deux mots en présence. Par exemple, le nom français *local* et le nom anglais *local* sont des faux amis absolus, mais leurs homographes adjectivaux n'en sont pas. Le vérificateur ne

tenant pas compte de l'information grammaticale, il a relevé des cas non pertinents. Deuxièmement, une certaine proportion de bruit était générée par des citations non traduites. Par exemple, le mot anglais *agenda* et le mot français *agenda* sont des FA absolus. Comme ils sont parfaitement identiques, notre vérificateur a été incapable de les distinguer et a donc extrait des cas où le mot *agenda* apparaissait des deux côtés, pour la simple raison que l'un des textes le renfermait sous la forme d'une citation non traduite de l'autre langue. Troisièmement, il y avait des cas où M_e et M_f apparaissaient effectivement dans des phrases qui étaient des traductions réciproques, mais où ces mots n'étaient pas eux-mêmes utilisés comme des équivalents traductionnels. Notre modèle de correspondance (c'est-à-dire d'alignement de phrases) était simplement trop grossier pour éliminer ces cas[3]. Ces sources de bruit sont ventilées dans le Tableau 2.

Catégorie de bruit	Nombre de cas	Pourcentage
Mauvaise catégorie grammaticale	703	97,9
Citation non traduite	6	0,8
FA non utilisés comme des équivalents traductionnels	9	1,3
Total	718	100

Tableau 2 : Catégories de bruit relevées dans les FA extraits des sorties de AT1

Ces résultats indiquaient clairement qu'il fallait incorporer la catégorisation grammaticale. Nous avons donc remplacé l'analyseur de traductions AT1 par un nouvel analyseur, AT2, qui se distinguait du premier en ce que ses deux modèles de langage incorporaient le programme de catégorisation grammaticale de Foster (1991). On a ensuite modifié le processus de

[3] À noter, cependant, qu'aucun élément de «bruit» n'était attribuable à des alignements erronés, notre algorithme assurant un taux de réussite d'environ 98 p. cent.

recherche pour tenir compte de l'information grammaticale associée à nos 145 paires de FA. Cette méthode a donné de bien meilleurs résultats, comme on peut le constater dans les Tableaux 3 et 4.

	Nombre de cas	Pourcentage
Erreurs réelles	56	76,7
Bruit	17	23,3
Total	73	100

Tableau 3 : Résultats de l'extraction de FA dans les sorties de AT2

Catégorie de bruits	Nombre de cas	Pourcentage
Mauvaise catégorie grammaticale	6	35,3
Citation non traduite	2	11,8
FA non utilisés comme des équivalents traductionnels	9	52,9
Total	17	100

Tableau 4 : Catégories de bruit relevées dans les FA extraits des sorties de AT2

La catégorisation grammaticale a réduit considérablement le niveau de bruit, tout en n'exerçant qu'un effet marginal sur le nombre de cas rappelés (perte d'un cas). Cette amélioration spectaculaire est en grande partie attribuable à la résolution des problèmes liés à un petit nombre de mots utilisés fréquemment (comme le cas du mot *local*, mentionné précédemment). Une partie du bruit restant est due à des erreurs de catégorisation grammaticale, mais la majeure partie découle

maintenant de la grossièreté de notre modèle de correspondance.

Il ne fait aucun doute que de meilleurs modèles permettraient d'améliorer la détection des faux amis. Cependant, la performance de la méthode de faible coût algorithmique que nous avons testée ici pourrait fort bien s'avérer suffisante pour les applications réelles.

5 La dictée de traductions : TransTalk

Nous sommes revenus quelques fois, dans cet article, sur le fait que des modèles de traduction faibles, s'ils sont employés de façon réaliste, peuvent offrir au traducteur des outils efficaces qui ne lui imposent pas de contraintes artificielles. Considérant que de nombreux traducteurs préfèrent dicter leurs textes plutôt que de les dactylographier eux-mêmes, un module de dictée automatique constituerait une addition fort utile au poste de travail de traducteur (Gurstein et Monette 1988).

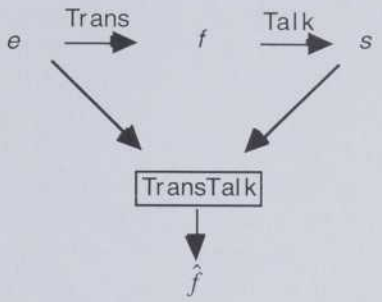
La technologie de la reconnaissance automatique de la parole est actuellement limitée aux vocabulaires restreints et elle est, de ce fait, inapplicable à la tâche de la plupart des traducteurs. Une possibilité intéressante, cependant, serait le *couplage* d'un module de reconnaissance de la parole et d'un modèle de traduction (faible). Le modèle de TA servirait alors à faire des prédictions probabilistes quant aux énonciations susceptibles d'être produites librement par le traducteur, afin de réduire dynamiquement le «vocabulaire réel probable» envisagé par le module de reconnaissance de la parole pour chaque unité de dictée (phrase ou paragraphe) et ce, jusqu'au point où la reconnaissance complète de ces unités pourrait être tentée.

Il est évident, par exemple, que la composition probabiliste du vocabulaire considéré par un module de reconnaissance de la parole qui tente de décoder la phrase *Ces impôts cachés doivent être acquittés par les pauvres aussi bien que par les riches* sera très différente selon que sa source anglaise *The poor as well as the rich have to pay these hidden taxes* est ou non connue du module. Il est beaucoup plus probable, par exemple, que la traduction française de cette phrase anglaise

renferme le mot *impôts* qu'une phrase française choisie au hasard. Il semble donc raisonnable d'espérer qu'un modèle de traduction faible puisse rendre cette composition accessible au module de reconnaissance de la parole.

Cette idée a été avancée par Dymetman, Foster et Isabelle (1992) ainsi que par Brown et al. (1992). Nous avons entrepris un projet de collaboration avec le groupe de reconnaissance de la parole du Centre de recherches informatiques de Montréal (CRIM), le projet *TransTalk*, qui vise à démontrer la faisabilité de cette approche, en utilisant l'anglais comme langue source et le français comme langue de dictée. Nous avons l'intention de nous limiter, au début, à la dictée de mots isolés, puis de passer progressivement à la parole continue. Les projets TransSearch et TransCheck décrits précédemment comportaient l'élaboration d'analyseurs de traductions incorporant des modèles de langage pour le français et l'anglais et un modèle de correspondance français-anglais (alignement de phrases) qui ont été entraînés sur le corpus du Journal des débats de la Chambre des communes. Ce corpus, ou domaine, est donc un choix tout naturel pour le projet TransTalk, puisque les modules existants constitueront alors des ressources fondamentales pour TransTalk. On peut en fait envisager TransTalk comme un système incorporant un analyseur de traductions fort semblable à ceux que nous avons déjà décrits, sauf qu'il a la capacité de traiter une langue cible parlée, plutôt qu'écrite.

TransTalk est basé sur un modèle probabiliste p de dictée de traduction qui met en relation une unité textuelle anglaise écrite e, sa traduction française écrite f (pour simplifier, nous supposerons que ces unités textuelles sont des phrases) et s, la réalisation acoustique de f. Les unités e et s sont toutes deux connues du système, qui a pour tâche de produire une estimation $\hat{f}$ de l'unité f effectivement formulée par le traducteur.



L'on est donc amené à définir $\hat{f}$ comme :

$$\hat{f} = \text{argmax}_f\, p(f \mid e, s)$$

c'est-à-dire que $\hat{f}$ est la phrase française la plus probable selon le modèle p, étant donné la phrase source anglaise et la réalisation acoustique de la phrase française.

En utilisant la formule de Bayes, on peut réécrire cette équation comme suit :

$$\begin{aligned}\hat{f} &= \text{argmax}_f p(s \mid e, f)\ p(f \mid e) \\ &= \text{argmax}_f p(s \mid f)\ p(f \mid e)\end{aligned}$$

la dernière égalité étant une conséquence de l'assomption modérée suivante : une fois que f est connue, les connaissances supplémentaires sur e n'ajoutent rien à la détermination de s.

Cette équation rappelle fortement la «formule fondamentale» de la reconnaissance statistique de la parole (Bahl et al. 1983) :

$$\hat{f} = \text{argmax}_f p(s \mid f)\ p(f)$$

où les distributions $p(s \mid f)$ et $p(f)$ sont appelées respectivement le modèle acoustique et le modèle de langage. Dans la situation envisagée ici, le modèle de langage pur $p(f)$ a été remplacé par un «modèle de langage conditionnel» $p(f \mid e)$, dans lequel la connaissance de e «affine» la structure statistique du modèle de langage, en le forçant en particulier à «concentrer» son attention sur un sous-ensemble lexical restreint de la langue. On peut mesurer quantitativement cet «affinement» au moyen de la *perplexité*, une quantité relative à la théorie de l'information qui mesure l'incertitude moyenne qu'un modèle de langage

manifeste à l'égard du prochain mot devant apparaître dans un texte naturel, après avoir vu les mots précédents : moins un modèle est perplexe, plus il est prédictif (Jelinek 1990). Brown et al. (1992) décrivent les résultats d'une expérience qu'ils ont réalisée sur le Journal des débats en utilisant un de leurs modèles de traduction les plus simples (du français à l'anglais, dans leur cas). Ces résultats révèlent que la perplexité par mot de leur modèle de langage pur (anglais) s'établit en moyenne à 63,3, tandis que la perplexité de leur modèle de langage conditionnel chute à une moyenne de 17,2. Ces résultats sont très encourageants pour la dictée, car ils signifient que le module acoustique devrait pouvoir faire un choix, étant donné un mot anglais parlé, parmi en moyenne 17,2 candidats équiprobables proposés par le modèle de langage conditionnel, par opposition à 63,3 candidats équiprobables proposés par le modèle de langage pur.

Plusieurs approches sont possibles pour la modélisation de $p(f \mid e)$. *Une première approche*, proposée par l'équipe IBM, consiste à utiliser la formule de Bayes pour écrire, par analogie à la formulation standard du problème de reconnaissance de la parole :

$$p(f \mid e) \propto p(e \mid f)\, p(f)$$

où (selon leur terminologie) $p(e \mid f)$ correspond au «modèle de traduction», qui joue un rôle semblable à celui du modèle acoustique dans la reconnaissance de la parole. Cela nous amène par conséquent à une formule symétrique pour l'ensemble du modèle de dictée de traduction, dans laquelle $p(f)$ est le modèle de langage, $p(s \mid f)$ est le modèle acoustique et $p(e \mid f)$ est le modèle de traduction. Cette méthode présente deux avantages : (1) elle repose sur un seul modèle de langage pour le français et (2) les travaux réalisés chez IBM sur la TA statistique semblent indiquer que des approximations même grossières de $p(e \mid f)$, *lorsqu'elles sont couplées à un bon modèle de langage pour le français*, donnent des approximations acceptables pour le modèle de langage conditionnel $p(f \mid e)$. C'est comme s'il y avait une «division du travail» entre $p(f)$, qui est responsable de la structure correcte des phrases françaises, et $p(e \mid f)$, qui est chargé d'apparier les phrases anglaises et

françaises (d'où le terme quelque peu trompeur de «modèle de traduction»), sans trop tenir compte de la structure interne du français ou de celle de l'anglais[4] (voir Dymetman et al. 1992, pour plus de détails). Cette méthode présente toutefois une lacune importante au niveau du traitement : elle exige la réalisation d'une recherche étendue parmi les phrases f pour maximiser $p(e \mid f)\, p(f)$ (sans compter le facteur $p(s \mid f)$, ce qui ne fait qu'aggraver les choses). On sait que cela pose de sérieuses difficultés pratiques en termes de résultats de recherche non optimaux ainsi qu'en termes de temps de traitement, ce dernier facteur étant évidemment de toute première importance pour une application de dictée.

Une deuxième approche à la modélisation de $p(f \mid e)$ consiste à considérer *a priori* une certaine famille paramétrisée de modèles de langage $p\lambda(f)$ pour le français, à décrire une mise en correspondance $e \rightarrow \lambda(e)$, puis à définir le modèle de langage conditionnel au moyen de :

$$p(f \mid e) = p_{\lambda(e)}(f)$$

Bien qu'elle présente l'inconvénient d'utiliser plus qu'un modèle de langage de référence pour le français, cette approche peut être mise en œuvre efficacement si la famille $p\lambda(f)$ est bien choisie. Nous examinons actuellement la possibilité d'adapter un modèle de langage proposé dans Foster 1991. Ce modèle est une sorte de modèle markovien caché «tri-POS», qui dépend de deux familles de paramètres. La première famille $a_{i,\, j,\, k}$, donne la probabilité de générer un mot de catégorie grammaticale POS_k, les mots de catégories grammaticales POS_i et POS_j ayant déjà été générés. La deuxième famille, $b_{i,\, m}$, donne la probabilité qu'une catégorie grammaticale donnée POS_i soit associée au mot m. Cela signifie, sur le plan conceptuel du

[4] En fait, il est facile de voir que, pour les fins de la traduction de l'anglais au français, il est équivalent d'utiliser $p(e \mid f)$ ou $p(e \mid f) \,/\, p(e)$ comme «modèle de traduction». Cette dernière quantité est l'exponentielle de l'information mutuelle entre e et f, une quantité symétrique dans e et f, qui ne possède aucune mémoire de la structure statistique interne de e ou de f, mais seulement de leur relation statistique.

moins, que le modèle génère d'abord des chaînes de catégories grammaticales, en utilisant la fenêtre du contexte des deux catégories grammaticales déjà générées, puis «décore» chaque catégorie grammaticale avec un mot donné, dépendant uniquement de cette catégorie grammaticale. Les paramètres $a_{i,\ j,\ k}$ représentent une approximation de la structure «grammaticale» du français, tandis que les paramètres $b_{i,\ m}$ représentent une approximation de sa structure «lexicale».

Nous nous proposons de faire l'essai d'un schème où ces paramètres varieront dynamiquement selon la phrase source observée e. Une possibilité intéressante serait de maintenir les «paramètres grammaticaux» à leurs valeurs globales fixes en langue française (sans tenir compte de l'influence de la composition grammaticale de la phrase anglaise sur sa traduction), tout en modifiant les paramètres «lexicaux» selon la composition lexicale de la phrase anglaise. La première famille de paramètres peut être estimée de façon fiable sur un corpus français suffisamment étendu, tandis que la deuxième famille de paramètres, qui dépend de e, peut être estimée si l'on fait certaines hypothèses simplificatrices s'apparentant au modèle de traduction 1 de Brown et al. (1993). Essentiellement, chaque $b_{i,mf}\ (e)$ est considéré être la moyenne des contributions $p\ m_f \mid m_e\ , POS_i)$ faites par chaque mot anglais m_e de e à la probabilité de réaliser la catégorie grammaticale POS_i dans le mot français m_f. Pour estimer les paramètres $p\ \ m_f \mid m_e\ , POS_i)$, il faut partir d'un corpus d'apprentissage pré-aligné composé de bi-textes anglais-français (voir section 3). Il est alors possible de faire des estimations initiales pour les paramètres $p\ (m_f \mid m_e\ , POS_i)$, puis d'utiliser des techniques de réestimation standard (voir Brown et al. 1993) sur ce corpus d'apprentissage pour maximiser l'efficacité prédictive de ces paramètres, tout en maintenant les paramètres grammaticaux à des valeurs fixes.

Le principal avantage de cette approche est que, pour chaque phrase source e, le modèle de langage conditionnel se réduit, en fait, à un simple modèle markovien caché $p_{\lambda(e)}(f)$; le problème de la dictée de traduction adopte alors une forme familière en reconnaissance de la parole, soit :

$$\hat{f} = \text{argmax}_f p_\lambda(e)(f)\ p\ (s \mid f)$$

pour laquelle il existe de puissantes techniques de recherche (Bahl et al. 1983).

6 Conclusion

Une nouvelle génération d'aides à la traduction se profile déjà à l'horizon. Grâce au développement des techniques d'analyse de traductions, les postes de travail de traducteurs pourront bientôt mettre à la disposition de leurs utilisateurs des outils qui dépasseront les simples fonctions de bureautique. Les traducteurs seront bientôt en mesure de tirer profit du vaste potentiel inexploité que recèle leur production antérieure. Ils disposeront bientôt d'outils de vérification qui les aideront à détecter les erreurs de traduction présentes dans leurs premiers jets. Et il y a de bonnes chances que la transcription automatique de la parole se concrétise en traduction bien avant qu'elle ne devienne une réalité pour les applications unilingues.

Nous ne serions pas surpris de voir cette liste d'applications basées sur le concept de l'analyse de traductions s'allonger rapidement. Nous ne souhaitons que du bien à la TA classique, mais nous croyons que c'est dans le domaine des aides à la traduction que se produiront les véritables progrès, et ce, pendant de nombreuses années encore!

Bibliographie

Bar-Hillel, Y., 1951, The State of Machine Translation in 1951, in *American Documentation*, vol. 2, pp. 229-237.

Bahl, L., Jelinek, F. et R. Mercer, 1983, A maximum likelihood approach to continuous speech recognition, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-5(2), pp. 179-191.

Brown, P., Lai, J. et R. Mercer, 1991, Aligning sentences in parallel corpora, *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, Berkeley (Californie).

Brown, P., Chen, S., Della Pietra, S., Della Pietra, V., Kehler, S. et R. Mercer, 1992, *Automatic speech recognition in machine aided translation*.

Brown, P., Della Pietra, S., Della Pietra, V., et R. Mercer, 1993, The Mathematics of Machine Translation: Parameter Estimation, *Computational Linguistics*, vol. 19, pp. 263-311.

Church, K. et W. Gale, 1991, Concordances for Parallel Texts, in *Proceedings of the 7th Annual Conference of the UW Centre for NOED and Text Research*, Oxford.

Debili, F. et E. Sammouda, 1992, Appariement des phrases de textes bilingues français-anglais et français-arabes, in *Proceedings of COLING-92*, Nantes.

Dymetman, M., 1992, *Transformations de grammaires logiques et réversibilité en traduction automatique*, thèse d'État, Université de Grenoble 1, France.

Dymetman, M., Foster, G. et P. Isabelle, 1992, *Towards an Automatic Dictation System for Translators (TransTalk)*, rapport technique, CITI (CITI), Laval (Québec), Canada.

Foster, G., 1991, *Statistical Lexical Disambiguation*, mémoire de maîtrise, McGill University, School of Computer Science.

Gurstein, M. et M. Monette, 1988, *Functional Specifications for a Translator's Workstation*, rapport technique 12SD.36902-5-0003, Socioscope Inc., Ottawa, Canada. Rapport présenté au Centre canadien de recherche sur l'informatisation du travail (Centre d'innovation en technologies de l'information).

Isabelle, P., 1987, Machine Translation at the TAUM Group, in Margaret King (éd.), *Machine Translation Today: The State of the Art*, Edinburgh University Press.

Isabelle, P., 1992, Bi-Textual Aids for Translators, in *Proceedings of the Eighth Annual Conference of the UW Centre for the New OED and Text Research*, University of Waterloo, Waterloo, Canada.

Isabelle, P., Dymetman, M. et E. Macklovitch, 1988, CRITTER: a Translation System for Agricultural Market Reports, in *Proceedings of COLING-88*, Budapest.

Jelinek, F., 1990, Self-Organized Modeling for Speech Recognition, in Alex Waibel et Kai-Fu Lee, éd., *Readings in Speech Recognition*, pp. 450-506, Morgan Kaufman, San Mateo, Californie.

Macklovitch, E., 1992, Corpus-Based Tools for Translators, in *Proceedings of the 33rd Annual Conference of the American Translators Association*, San Diego.

Macklovitch, E., 1993, *A Third Version of the CWARC's Workstation for Translators*, rapport technique, CWARC (CITI), Laval (Québec), Canada.

Kay, M., 1980, *The Proper Place of Men and Machines in Translation*, CSL-80-11, Xerox PARC.

Sato, S. et M. Nagao, 1990, Toward Memory-Based Translation, in *Proceedings of COLING-90*, pp. 247-252.

Simard, M., Foster, G. et P. Isabelle, 1992, Using Cognates to Align Sentences in Parallel Corpora, in *Proceedings of the 4th International Conference on Theoretical and Methodological Issues in Machine Translation*, Montréal.

Van Noord, G., 1993, *Reversibility in Natural Language Processing*, CIP-Gegevens Koninklijke Bibliotheek, La Haye.

Van Roey, J., Granger, S. et H. Swallow, 1988, *Dictionnaire des faux-amis français-anglais*, Paris, Duculot.

LE TRAITEMENT DES TEXTES PRIMAIRES ET SECONDAIRES POUR LA CONCEPTION ET LE FONCTIONNEMENT D'UN PROTOTYPE DE SYSTÈME EXPERT D'AIDE À L'ANALYSE DES JUGEMENTS

Suzanne Bertrand-Gastaldy, Louis-Claude Paquin,
Gracia Pagola, François Daoust

Résumé

Afin d'assister les conseillers juridiques de la Société Québécoise d'Information Juridique (SOQUIJ), un prototype de système expert pour l'aide à la sélection, à la classification, à la lecture et à l'indexation des jugements a été implanté sur ACTE (Atelier Cognitif et TExtuel) développé au Centre de recherche en information et cognition ATO.CI. À partir d'un corpus d'apprentissage de textes déjà analysés et grâce à des traitements statistico-linguistiques sur SATO (Système d'Analyse de Textes par Ordinateur) et SPSS, on a confronté les données issues de l'analyse humaine à celles des textes intégraux; des tendances et des anomalies ont pu être décelées qui ont servi à questionner les outils et les pratiques ainsi qu'à réorienter ou à corroborer l'enquête cognitive des savoir-faire. Une fois identifiés les types d'unités linguistiques et leurs propriétés généralement retenus par les spécialistes pour chacune des opérations d'analyse, on a mis au point une chaîne de traitements qui, pour chacun des modules du système expert et à partir du plus grand nombre possible de sources de connaissances, dépiste et décrit les indices pertinents, puis les transforme en faits. Ceux-ci s'avérant distincts les uns des autres, un diagnostic peut être porté, à chaque

étape, selon un principe de convergence. Toutefois, quelques difficultés concernant le cumul des coefficients de certitude et l'intégration de statistiques nécessitent des études plus poussées.

Introduction

Au Québec, la cueillette, la sélection, le traitement et la diffusion de la jurisprudence sont sous la responsabilité principale d'un organisme parapublic : SOQUIJ. La loi constituant la Société québécoise d'information juridique, entrée en vigueur le 1er avril 1976, lui confie le mandat de "promouvoir la recherche, le traitement et le développement de l'information juridique en vue d'en améliorer la qualité et l'accessibilité au profit de la collectivité." À ce titre, SOQUIJ est le serveur des bases de données du ministère de la Justice ainsi que le producteur-serveur de plusieurs autres bases de données, dont celles qui concernent la jurisprudence.

Or, la saisie électronique des jugements à la source mise en place progressivement par le ministère de la Justice du Québec multipliera presque par cinq le nombre de jugements acheminés à SOQUIJ : la quantité annuelle prévue est de 50 000. Pour maintenir son niveau de service sans accroître indûment son personnel, cet organisme a envisagé de recourir à des méthodes automatiques pour assister certaines des opérations intellectuelles d'analyse effectuées par des conseillers juridiques et a confié à notre équipe de recherche le mandat de concevoir un prototype de système expert. Celui-ci a été réalisé sur le logiciel ACTE (Atelier Cognitif et TExtuel) développé au Centre ATO.CI.

Après avoir expliqué en quoi consistent les différentes fonctions d'analyse qu'il nous a fallu modéliser, nous évoquerons les éléments théoriques sur lesquels nous avons appuyé notre démarche et nous montrerons comment celle-ci combine des approches complémentaires (statistiques, linguistiques et cognitives); nous l'illustrerons ensuite par quelques exemples d'indices extraits pour chaque type d'analyse. Puis nous mentionnerons les principaux enrichissements du thésaurus nécessités par les nouvelles fonctions qu'il est appelé à remplir dans un système automatique. Finalement, nous exposerons

l'implantation des stratégies d'analyse dans un programme de chaîne de traitement et l'intégration d'informations de sources et de valeurs différentes.

1 La modélisation des tâches d'analyse

1.1 Les fonctions d'analyse à assister

Après entente avec les représentants de SOQUIJ, nous avons convenu de concevoir un système qui assisterait les tâches suivantes : 1) élimination à la source de certains jugements (étape de la sélection); 2) détermination du (ou des) domaine(s) du droit et, le cas échéant, du sous-domaine (selon un plan de classification préétabli) auquel chaque décision retenue appartient (étape du tri et de la classification); 3) prise de connaissance du contenu des textes en vue de la rédaction d'un résumé informatif; 4) sélection de termes d'indexation à partir du résumé rédigé par les conseillers juridiques.[1]

1.2 Théories sous-jacentes

La méthodologie que nous avons élaborée est sous-tendue par au moins trois orientations théoriques : le texte comme objet sémiotique, les analyses documentaires comme des applications particulières d'un processus de lecture, et finalement l'intertextualité.

1.2.1 Le texte comme objet sémiotique

Le texte est envisagé comme un objet sémiotique complexe dans lequel un lecteur humain ou informatique sélectionne, à des niveaux multiples, des indices pertinents en fonction de ses objectifs d'analyse (Meunier 1992, Meunier et al. 1994). Les chaînes de caractères ou «mots» sont autant de porteurs de traits signifiants. Les processus cognitifs d'interprétation

---

[1] Le mandat qui nous avait été confié incluait plusieurs contraintes dont celle de respecter à la fois les outils documentaires actuels (thésaurus et plan de classification) et les «habitudes» des experts - habitudes cependant très faiblement documentées.

humaine étant fonction de nombreux éléments dont la plupart ne sont guère formalisables (systèmes de croyance, intentions, connaissances du contexte textuel et extra-textuel, etc.), un système entièrement automatique est impossible à envisager : seul un mécanisme d'aide à l'interprétation est réalisable qui permet d'identifier et de manipuler certains des indices pertinents décelables par divers analyseurs.

1.2.2 Les analyses documentaires comme des applications particulières d'un processus de lecture

Les opérations d'analyses effectuées dans un service documentaire sont envisagées comme des lectures particulières dirigées par des tâches spécifiques à accomplir : attribution d'une rubrique de classification, assignation de mots-clés, condensation du texte, entre autres. Ces lectures mettent en jeu diverses opérations cognitives de sélection, rejet, généralisation (Van Dijk 1977), stratégies de confirmation et contrôle, etc. (David 1990) portant sur des indices ou configurations d'indices dont la pertinence varie en fonction de chaque type de lecture. À chaque tâche d'analyse correspond donc un parcours particulier du texte. La décision d'inclure un document dans une base de données - ou de le rejeter - n'exige pas la prise en compte du même nombre ni des mêmes types d'indices que l'opération d'indexation. La rédaction d'un résumé requiert une prise de connaissance plus approfondie du contenu textuel que l'attribution d'une rubrique de classification, mais exige un examen moins attentif cependant que la comparaison des thèses défendues par un texte avec celles d'un autre texte.

1.2.3 L'intertextualité

De par la nature même de leur condition de production, les textes secondaires sont en relation d'intertextualité avec les textes primaires dont ils sont issus (Beacco et Darot 1984) ainsi qu'avec les outils documentaires - thésaurus et plan de classification - servant à effectuer l'analyse (Begthol 1986). Comme nous l'avons exposé dans Bertrand-Gastaldy (1993), la comparaison des propriétés des éléments présents dans les textes de départ et retenus dans les différents textes d'arrivée (rubriques

de classification, termes d'indexation, résumés) avec celles des éléments qui ont été éliminés permet de découvrir des tendances et des anomalies qui servent à orienter ou approfondir l'enquête cognitive auprès des experts.

1.3 Notre approche

Nous n'avons donc pas cherché à mettre au point un outil d'analyse qui serait performant en dehors de tout contexte (par exemple un analyseur morphologique, un extracteur de lexies complexes), mais bien au contraire de comprendre en quoi le contexte de la tâche faisait varier les objets textuels et les propriétés des objets susceptibles de retenir l'attention des experts. Notre approche a dès lors consisté d'une part à modéliser les stratégies cognitives mises en oeuvre par les experts du domaine lors des différentes lectures effectuées en fonction des produits attendus (liste des documents à éliminer, tri et classification, résumé, indexation), d'autre part à faire évoluer les outils documentaires pour les rendre aptes à répondre à l'utilisation automatique que nous voulions en faire.

1.3.1 Les sources de données

Pour mener à bien notre travail, nous disposions de deux types de sources. Nous avions accès à une demi-douzaine de conseillers juridiques avec lesquels nous avons tenu plusieurs sessions de travail afin d'arriver à connaître les critères explicites ou implicites auxquels ils recourent pour prendre leurs décisions aux différentes étapes de leur analyse. D'autre part, les données de nature linguistique se trouvaient déjà presque toutes sur support informatique. Il s'agit des produits issus des différentes opérations d'analyse : textes intégraux rejetés ou retenus, notices bibliographiques accompagnées des résumés, index, ainsi que des outils utilisés pour l'analyse : plan de classification et thésaurus.

1.3.2 Les traitements sur les données linguistiques

Les caractéristiques attribuées aux données, en contexte ou hors contexte, ont consisté en l'ajout d'informations de nature diverse décrivant le statut sémiotique des constituants du texte

et enrichissant les chaînes de caractères immédiatement accessibles à l'ordinateur. Ces caractéristiques proviennent de connaissances générales de la langue (type de langue, nature grammaticale des lexèmes), de connaissances générales sur la structure des textes (phrases, paragraphes), d'informations de nature éditique (conventions typographiques - capitales, caractères gras ou italiques - dans les enregistrements), de connaissances spécifiques au domaine (vocabulaire de spécialité, structure des jugements et de leurs résumés, mention de loi, de jurisprudence et de doctrine), de connaissances "documentaires" (champs d'une notice, appartenance ou non des lexèmes aux langages documentaires), de propriétés statistiques (fréquence absolue ou relative, indice de répartition, valeur discriminante, chi^2, etc.). Ces informations ont été obtenues par des algorithmes développés avec le logiciel SATO (Système d'Analyse et Textes par Ordinateur) et ont fait l'objet d'un marquage approprié (propriété et valeur de propriétés dans SATO). On pourra consulter une publication du Centre ATO.CI pour plus de détails (Bertrand-Gastaldy et al. 1993).

Nous présentons ci-dessous un extrait de texte dans lequel apparaissent diverses propriétés et leurs valeurs :

- les caractères typographiques ***typo**, avec les valeurs *italique* et *nil.*
- les subdivisions ***par**, (avec les valeurs *manchette, litige, contexte, décision*), l'appartenance aux outils documentaires d'où sont tirés les mots-clés (***term**) avec les valeurs *Ta* pour descripteurs du thésaurus acceptés, *Tr* pour descripteurs rejetés du thésaurus, *Tl* pour termes libres du domaine tels qu'identifiés par les experts, *Clas* pour rubrique de classification (ces valeurs peuvent être spécifiées par les premières lettres du sous-domaine du droit auquel appartiennent les termes : par exemple, *TlAss* pour terme libre caractérisant le domaine Assurances);
- la numérotation des phrases ***phr** et leur ordre ***ord** (*pr* pour première, *deux* pour deuxième, *ad* pour avant-dernière, *de* pour dernière);
- la position (***marque**) des mots dans la macrostructure (manchette, litige, contexte, décision) : un terme portant la

valeur *mancondéc* se trouve donc à la fois dans la manchette, dans le contexte et dans la décision.

On remarque également, dans le texte qui suit, certains résultats du prétraitement automatique ou semi-automatique : le doublement des traits d'union séparant deux éléments grammaticaux différents (soit--elle), le doublement des points d'abréviation et l'ajout d'une barre oblique devant les majuscules de noms propres (\C..\C..). La détection de termes complexes dans les outils documentaires ou la liste de termes libres du domaine préparés en cours d'expérimentation a résulté en la substitution d'un trait d'union au caractère blanc figurant entre les composants des termes. Tous les ajouts sont inscrits en caractères gras dans notre exemple :

***par=ident*typo=nil**<ND>91-3 ***par=prov**<HD>COUR_D_'APPEL

par=manchette** ASSURANCEterm=(TaAss,ClasAss) *marque=mandéc** -- assurance_de_responsabilité***term=ClasAss *marque=man** -- recours***term= Clas *marque=mancondéc** contre le tiers responsable -- option***term=Ta *marque=mancondéc** -- article 2603 C.C. -- interdiction_de_cumul***temr-T1** -- amendement***term=(Ta,Clas) *marque=mancondéc**.

par=litige *phr=1 *ord=(ad,pr)** Appelterm=(Ta,Clas) *marque=li** d'un jugement***term=Ta *marque=li** de la \Cour supérieure ayant accueilli une requête_en_irrecevabilité. ***term=T1 *phr=2 *ord=(de,deux)** Rejeté, avec dissidence.

par=contexte *phr=1 *ord=pr** Le 18 février 1988, l'appelante a intenté une actionterm=Tr *marque=condéc** contre la mise_en_cause ***term=Tr *marque=con** \Fontaine, lui réclamant 23 688$ à titre de dommages***term=Class *marque=con** à la suite d'un incendie ***term=TaAss *marque=condéc** provoqué par sa négligence***term-Tr *marque=con phr=2 *ord=deux**. Quelques mois plus tard, l'appelante a fait signifier une déclaration***term=Ta *marque=con** amendée qui ajoutait la compagnie_d'_assurances***term=T1Ass** intimée à titre de défenderesse et qui concluait à la condamnation conjointe et solidaire des codéfenderesses.
phr=3 *or=au** L'intimée a alors présenté une requête_en_irrecevabilitéterm=T1** fondée sur le fait que l'appelante n'avait aucun recours***term=Clas *marque=mancondéc** contre elle puisque, en poursuivant \Fontaine, elle avait exercé l'option***term=Ta *marque=mancondéc** prévue à l'article 2603 \C..\C.. . ***phr=4 *ord=au** La requête_en_irrecevabilité***term=T1** a été accueillie malgré la demande verbale d'amendement*term=(Ta,Clas) *marque=mancondéc présentée par l'appelante visant à modifier la désignation des parties et à ne maintenir que l'intimée à titre de défenderesse,

reléguant \Fontaine au rang de mise_en_cause*term=Tr*marque=con. [...]

*par=décision *typo=italique *phr=1*ord=pr \Mme la juge*term=Ta *marque=condéc \Tourigny et \M.. le juge*Term=Ta *marque=condéc \Proulx : *typo=nil Les dispositions du *typo=italique Code de procédure_civile*term=(Ta,Clas) *marque=déc *typo=nil relatives à l'amendement doivent recevoir une interprétation aussi large que possible. *phr=2 *ord=deux Cependant, une interprétation, aussi large soit--elle, ne peut écarter une disposition de droit substantif incluse dans le *typo=italique \Code civil. *typo=nil *phr=3 *ord=au Le législateur a voulu que, en intentant un recours*term=Clas *marque=mancondéc, la partie demanderesse fasse un choix, ainsi que l'a confirmé \M.. le juge*term=Ta *marque=condéc \Mayrand dans l'arrêt *typo=italique \L'\Union québécoise, mutuelle [...]

En se positionnant sur un mot, on peut visionner, à l'aide d'une commande du logiciel SATO, toutes les valeurs de propriétés (ou traits) qui lui ont été attribuées, y compris celles qui résultent de calculs statistiques effectués par le logiciel :

assurance_de_responsabilité			**mise_en_cause**		
*alphabet	=	fr	*alphabet	=	fr
*fréqtot	=	3	*fréqtot	=	2
*longueur	=	27	*longueur	=	13
*term	=	ClasAss	*term	=	Tr
*marque	=	man	*marque	=	con
*par	=	manchette	*par	=	contexte
			*phr	=	(1,4)
			*ord	=	(pr,au)
*poids	=	27	*poids	=	26
*discri	=	92	*discri	=	133
*chi^2	=	2958	*chi^2	=	4890
*gramr	=	tcomposé	*gramr	=	tcomposé

Une fois caractérisées, les données ont été filtrées en fonction des différents indices et soumises à une analyse de discrimination sur SPSS qui a fait ressortir les meilleurs prédicteurs pour expliquer les résultats des diverses opérations d'analyse.

1.4 L'enquête cognitive

Les résultats des analyses ont ensuite été confrontés aux données recueillies dans une première phase de l'enquête cognitive, puis soumis aux experts du domaine qui avaient pour tâche de confirmer les tendances observées et d'expliquer les anomalies, et surtout de décider d'une éventuelle réingénierie des processus ainsi que d'une éventuelle modification des outils documentaires.

L'enquête cognitive (comportant entrevues, observation et recueil de commentaires sur les résultats de nos traitements) a donc permis à la fois de compléter les analyses de données exposées précédemment et de les orienter. Nous cherchions les techniques et les stratégies employées pour parcourir un texte, les différentes parties du texte examinées pour prendre une décision de sélection, de tri-classification, de résumé et d'indexation, les connaissances utilisées (importance de tel ou tel tribunal, poids à accorder à la nature des parties en cause, valeur discriminante de telle ou telle mention de loi, de tel ou tel lexème, marqueurs du raisonnement du juge; références au contenu de la base de données, aux besoins des utilisateurs, à l'actualité, etc.), les catégorisations effectuées, les inférences faites pour passer des expressions en langue naturelle à leurs équivalents dans le thésaurus. Nous avons donc procédé selon une boucle : textes --> conseillers juridiques --> textes.

1.5 Complémentarité des approches

La complémentarité des approches utilisées pour la modélisation, recommandée à plusieurs reprises (Chaumier et Dejean 1992, Doszkocs 1986, Blosseville et al. 1992, Meunier et al. 1987), permet de tenir compte de la multiplicité des connaissances mises en oeuvre pour l'analyse du matériau textuel orientée vers des fins documentaires. Elle constitue, nous semble-t-il, un heureux compromis qui tient compte de caractéristiques exigeant parfois des solutions contradictoires, dans l'état de développement actuel des technologies : matériau textuel très complexe à analyser, mais nécessitant néanmoins des approches de nature linguistique et cognitive, volume important des données prohibant des analyses très fines et

pouvant bénéficier des effets de nombre, savoir-faire de plusieurs experts à expliciter, selon des méthodes appropriées à leur mode d'inscription, de façon à respecter la culture de l'organisation.

1.6 Exemples d'indices pertinents retenus pour modéliser les différentes opérations d'analyse

Nous donnons ci-dessous quelques exemples d'indices utilisés par les conseillers juridiques pour chacune des opérations d'analyse qu'il nous a fallu modéliser et nous fournissons des indications sur la place qui leur a été réservée dans le prototype.

1.6.1 La sélection

L'annexe 2 au Règlement sur la cueillette et la sélection des décisions judiciaires (*Loi sur la Société québécoise d'information juridique* (L.R.Q., chap. S-20, art. 21) indique qu'une décision peut être sélectionnée si elle contient un des éléments suivants : 1) un point de droit nouveau; 2) une orientation jurisprudentielle nouvelle; 3) des faits inusités; 4) une information documentaire substantielle; 5) une problématique sociale particulière.

Notons tout de suite que tous les jugements de la Cour suprême sont gardés ainsi que tous les jugements de la Cour d'appel à moins que ces derniers ne soient pas motivés. Pour les autres cas, les conseillers juridiques nous ont fourni, pour chacun des critères mentionnés dans le règlement, au moins un exemple d'indice textuel, mais, à part le nombre de citations aux lois et à la jurisprudence, ce sont des indices qui se détectent difficilement par une analyse automatique. Nous avons été à même de constater que l'étape de la sélection repose sur des opérations cognitives complexes mettant en jeu de nombreuses connaissances spécialisées.

La détection de la plupart des indices pertinents nécessite une compréhension du sens des phrases ou de plus larges portions du texte (par exemple, lorsque le juge exprime son désaccord - critère n°2) et des connaissances sur le monde, en

particulier sur l'actualité (en Responsabilité civile, il faut détecter le fait inusité - critère n°3 - comme un traitement médical nouveau ou la chute d'une personne aveugle sur un trottoir). Ou bien il faut identifier, à l'intérieur des textes, certaines catégories d'information et apprécier leur importance relative, par exemple la nouveauté du jugement par rapport à ceux qui ont été publiés dans des numéros antérieurs, etc. C'est pourquoi la prise de décision restera toujours la prérogative des conseillers juridiques.

En plus de consulter les experts, nous avons procédé par apprentissage sur corpus. Après examen d'un certain nombre de jugements rejetés (disponibles sur support papier seulement) et d'une série de jugements retenus, nous sommes arrivés à la conclusion que quelques critères formels simples permettent néanmoins de déclarer candidats au rejet un certain nombre de textes : 1) les jugements sont courts; 2) les jugements sont de type formulaire; 3) ils proviennent de la Cour des petites créances; 4) ils entérinent une convention. Une liste de types de requêtes ne faisant généralement pas l'objet de sélection a été constituée, mais n'est pas encore validée définitivement. Le prototype inclut seulement le premier et le dernier critères et nous envisageons d'y rajouter celui qui s'appuie sur la structure physique des jugements. La tâche sera d'autant plus facile que les textes seront saisis selon la norme SGML (Standard General Markup Language), ce que malheureusement n'a pas prévu pour le moment le ministère de la Justice.

1.6.2 Le tri-classification

Un document *Savoir-faire des conseillers juridiques pour le tri* a été constitué à partir des entrevues effectuées auprès des conseillers juridiques. Il explicite les critères de tri utilisés pour chacune des 57 grandes classes du plan de classification.
Il s'avère que l'appartenance d'un jugement à un domaine du droit peut être décelée, dans plusieurs cas (par exemple : DROIT PÉNAL, FAMILLE, TRAVAIL), d'après quatre types de renseignements contenus dans la première page : le tribunal, le nom des parties ou la procédure entreprise, le numéro de greffe, l'intitulé du jugement le cas échéant.

- Ainsi, un jugement provenant de la Chambre d'expropriation de la Cour du Québec traitera assurément du domaine de l'expropriation. Par ailleurs, un jugement dont l'une des parties est un syndicat pourrait vraisemblablement aborder le droit du travail. Enfin, un jugement qui mentionne qu'il s'agit d'une requête en irrecevabilité à l'encontre d'une action en dommages-intérêts pourrait être classé en procédure civile.
- Dans certains cas, le numéro de greffe permet de classer immédiatement le jugement sous la bonne rubrique : par exemple, lorsque le chiffre qui suit le premier tiret est 11 (500-11-222222), il s'agit de FAILLITE, 41 pointe vers FAMILLE-PROTECTION DE LA JEUNESSE, 12 ou 04 vers FAMILLE, 43 vers FAMILLE-ADOPTION.

Mais comme il existe des chevauchements entre plusieurs rubriques de classification (par exemple, le DROIT CIVIL recoupe OBLIGATIONS, VENTE, CONTRATS, entre autres) et comme plusieurs rubriques (quatre au maximum) peuvent être attribuées à un même jugement en vertu des politiques implicites de classification, il est parfois nécessaire de consulter le texte du jugement, pour prendre connaissance soit des lois ou articles du code civil cités, soit du vocabulaire employé par le juge (on retrouve dans ce vocabulaire beaucoup des termes répertoriés dans le thésaurus ou le plan de classification). Pour le domaine ASSURANCE, par exemple, sur la première page, le tribunal qui rend la décision n'est pas un bon indice. Si le nom d'une des parties désigne une compagnie d'assurances, il est possible mais pas certain qu'il faille classer le jugement dans ASSURANCE; une compagnie d'assurances qui a indemnisé son assuré peut, en effet, poursuivre la personne qui lui a causé des dommages et il faudrait alors classer le jugement dans RESPONSABILITE. Le fait que, dans le texte du jugement, les articles 2468 à 2676 du *Code civil* ou bien la *Loi sur les assurances* soient cités, vient renforcer le second indice. Si, de surcroît, le jugement comporte les termes comme : «assurance-automobile, assurance collective, assurance de choses» ou ses spécifiques : «assurance-incendie, assurance-vol, assurances de personnes» ou à nouveau les spécifiques de ce dernier terme :

«assurance-vie, assurance-invalidité, assurance-accident», ou encore «assurance (de) responsabilité, assurance maritime», alors on peut prendre la décision de le classer dans ASSURANCE avec une quasi-certitude de ne pas se tromper.

Notre enquête cognitive a révélé l'utilité d'autres types de combinaison d'indices comme la présence d'un terme associée à sa position (par exemple, «requête en liquidation d'une compagnie» qui, se trouvant dans les premières pages du jugement, entraîne la décision de le classer dans COMPAGNIES, de même que «demande en divorce» qui permet de le classer dans FAMILLE) ou la co-présence et la proximité de deux termes surtout dans le cas où l'un des termes est vague et peut pointer vers plusieurs domaines du droit («délégation» près du terme «obligation» est un bon indice pour le classement sous la rubrique OBLIGATIONS, de même que «divorce» et «pension alimentaire» ou «prestation compensatoire» pour FAMILLE). Mais ceci n'a pas été implanté dans le prototype actuel.

Enfin, notons que certains indices permettent de classifier immédiatement le jugement dans une sous-rubrique sans attendre une analyse plus approfondie de la part du conseiller juridique responsable du domaine : ainsi, le nom du tribunal «Cour du Québec - Chambre de la jeunesse» et la mention de la «Loi sur la protection de la jeunesse» pointent sans ambiguïté vers la sous-rubrique PROTECTION DE LA JEUNESSE dans la rubrique FAMILLE.

On constate que l'analyseur textuel doit repérer plusieurs indices différents. Il faut pour cela que ces éléments fassent l'objet d'une fouille appropriée, ce qui est réalisé grâce au système de marquage de propriétés dans SATO et pourrait l'être au préalable avec SGML, dans certains cas comme les citations de lois et de jurisprudence, par exemple.

Nous avons ajouté à cette approche linguistico-cognitive, une approche purement statistique qui a consisté en une analyse discriminante (effectuée avec SPSS) des mots par rapport à

un corpus d'apprentissage : l'algorithme utilisé est capable de produire un indice de confiance dans le résultat obtenu. [2]

1.6.3 La prise de connaissance du contenu du jugement en vue de la rédaction du résumé

En observant les conseillers juridiques en train de parcourir et d'annoter les textes de jugements et en recueillant les commentaires qu'ils ont bien voulu faire pendant ou après l'exécution de leur tâche, nous avons pu brosser un portrait de la façon dont ils prennent connaissance du contenu. Nous avons constaté que certains éléments utiles pour le tri-classification peuvent ensuite être réutilisés avec d'autres indices pour la rédaction du résumé et l'indexation. Nous avons ensuite pu établir une liste des éléments textuels importants pour chacun des experts selon les domaines de droit dans lesquels il œuvre.

Chaque spécialiste possède un schéma de la structure d'exposition des jugements dans tel ou tel domaine et recherche les énoncés-clés dans les parties réputées les contenir : questions de droit au début du jugement, motifs d'accusation («motifs suivants», «chefs d'accusation») et peine dans les premières lignes, énoncés des faits («les faits se résument comme suit») au début du jugement, moyens de procédure.

Les unités lexicales, particulièrement celles qui figurent dans le thésaurus et, le cas échéant, dans les listes de termes supplémentaires élaborées par quelques conseillers ainsi que les expressions pouvant indiquer qu'il y a discussion, lien de causalité, interprétation, etc. semblent constituer de bons

---

[2] L'analyse discriminante établit le «portrait-robot» de chaque classe à partir du profil statistique de chacun des individus de la classe. Le meilleur résultat que nous ayons obtenu dans nos expérimentations avec un grand nombre de variables sélectionnées en fonction de leur chi^2 et de certaines autres caractéristiques comme la langue a donné un taux de réussite de 92% en apprentissage et de 68% pour le groupe-test. Nous prévoyons que le dépistage des multitermes et des lois citées améliorera encore la précision des résultats. Nous envisageons aussi de procéder à la sélection de termes identifiés par les experts pour chaque domaine, car l'expérience de classement à l'aide de SATO a montré la pertinence de cette approche.

déclencheurs dans certains domaines, mais aussi la citation d'un article de loi, la mention du *Code civil* ou du *Code de procédure civile*, de la *Charte québécoise des droits et libertés*, etc.

Certaines divergences de lecture tiennent tout simplement au style cognitif des conseillers juridiques, mais peuvent en même temps être déterminées par la plus ou moins grande complexité du domaine ou les possibles recoupements entre domaines dans lesquels les jugements peuvent être classés : là où une personne lit intégralement le texte pour en prendre connaissance, plusieurs autres se contentent d'une lecture rapide favorisant qui le début et la fin du texte, qui le début et la fin de chaque paragraphe.

Dans le prototype de système expert, nous avons, pour le moment, retenu trois profils de lecture qui conviennent à tous : 1) les termes appartenant aux outils documentaires (thésaurus et plan de classification); 2) les intervenants (Juge, cour, parties, etc.); 3) les sources du droit (lois, articles, jurisprudence, etc.). Des couleurs différentes les mettent en relief et l'on peut les visualiser, au choix, dans le texte intégral, dans leur contexte immédiat, dans les phrases dans lesquelles ils sont insérés ou encore dans les paragraphes.

Pour une étape ultérieure de notre recherche, nous envisageons, en outre, une aide à la lecture personnalisée en fonction du domaine de droit dans lequel le jugement aura été préalablement classé, cette aide consistant tout simplement à mettre en relief par des couleurs les indicateurs particuliers à ce domaine. Par exemple, en RESPONSABILITÉ, le système surlignerait : «liens de causalité», «faute», «dommage exemplaire», «Charte des droits et libertés», etc.

Finalement des études exploratoires nous ont montré qu'il serait possible de mettre en lumière de façon différenciée, les parties du jugement qui traitent du litige, du contexte et de la décision, d'après des constantes de vocabulaire observées dans les résumés où ces trois parties sont très nettement distinguées (occurrences de lexèmes très différents, temps des verbes, etc.).

1.6.4 L'indexation

La tâche d'indexation, plus complexe que les tâches précédentes, n'a pas été aussi bien explicitée par les experts et nous avons dû nous appuyer sur la littérature - très peu diserte cependant - pour formuler des hypothèses en vue des traitements. En effet, l'étude cognitive des opérations d'analyse documentaire ne bénéficie pas d'une longue tradition en sciences de l'information (Bertrand 1993, Bertrand-Gastaldy et al. 1994, David 1990, Endres-Niggemeyer 1990, Farrow 1991).

Mais il est clairement apparu que l'assignation des termes à insérer dans la manchette puis dans l'index est effectuée d'après le résumé. D'ailleurs, pour expliquer ses choix, une personne nous a précisé : «En lisant le résumé, il y a des mots qui clignotent. Question d'expérience, de flair.» Le dispositif auquel nous recourons pour mettre les termes importants en valeur permet justement de faire clignoter les termes marqués.

Le type de termes retenus, leur localisation dans le résumé, leur forme, leur ordre d'inscription semblent répondre à de très nombreuses règles mises au point par chacun au fil de l'expérience, selon les domaines. Si le système expert doit reproduire ces règles, la tâche va être longue et surtout va nécessiter de entrevues supplémentaires : chaque cas est un cas particulier ou presque. Par contre, c'est à ce prix que le système pourra faciliter la cohérence, - du moins la cohérence intra-indexeur -, alléger le fardeau des conseillers juridiques et les libérer pour les prises de décision les plus délicates, notamment pour les cas-frontières.

En attendant de pouvoir approfondir cette enquête cognitive, nous nous sommes livrés à une étude comparée des propriétés des termes présents ou pas dans les résumés et retenus ou pas dans les manchettes. Toutes nos études ont pris appui sur les phénomènes d'intertextualité entre les résumés, les manchettes et les outils documentaires. Nous avons, entre autres, examiné l'importance de critères comme la position des termes dans la macro et la meso-structure des résumés, leur fréquence, leur valeur discriminante. Pour concevoir une aide à l'indexation directement à partir des textes intégraux,

notamment pour ceux qui ne feront pas l'objet de résumé (c'est une perspective envisagée par SOQUIJ à plus ou moins long terme), il faudrait inclure ceux-ci dans l'étude des phénomènes d'intertextualité.

L'enquête cognitive a révélé, en outre, que, dans plusieurs domaines, l'indexation obéit à une sorte de grille implicite : le premier descripteur est chargé d'apporter tel type d'information, le second tel type de précision, etc. Par exemple, en droit pénal, on respecte l'ordre suivant : la rubrique, la sous-rubrique, le type d'infraction commise, les principes de droit étudiés dans la décision, les mentions sur l'appelant, le contexte de l'infraction, la peine imposée, alors qu'en procédure civile, on retient successivement : l'identification de la procédure, le moyen de procédure, le type de défense.

Sachant que, pour les experts de SOQUIJ, l'ordre d'inscription des termes a une signification, il nous sera possible de mettre au point, dans une phase ultérieure, des traitements plus complexes permettant de comparer, dans chaque domaine du droit, les listes de termes assignés en première, deuxième, troisième positions, etc. pour faire surgir des grilles utilisées de façon peut-être inconsciente. Pour le moment, le prototype de système expert d'aide à l'indexation ne fait que surligner de façon différenciée les différents mots-clés potentiels et produire une liste de ces mots-clés triés selon le domaine de classification et classés par ordre de fréquence décroissante.

Avant d'implanter toutes les fonctionnalités envisagées (prise en compte de la valeur discriminante, de la position dans la macro-structure et la micro-structure), il faut que les experts prennent plusieurs décisions sur leurs politiques d'indexation en fonction de nos observations et recommandations et se prononcent également sur les modifications des outils documentaires. Nous pensons qu'en les confrontant aux résultats produits par un système expert encore rudimentaire, nous les aiderons à expliciter davantage les choix qu'ils feront à partir des suggestions de la machine.

1.7 Les propositions d'enrichissement des outils documentaires

Tout au cours de notre projet, nous avons été amenés à étudier l'utilisation des outils documentaires et à introduire des modifications qui facilitaient le travail de marquage automatique, modifications qui pourraient même être utiles dans le contexte d'une analyse humaine.

1.7.1 Les modifications à apporter au plan de classification

D'après le taux d'utilisation des différentes rubriques et sous-rubriques, le plan de classification nous a semblé répondre au volume et au rythme de publication des analyses de jugements dans *Jurisprudence Express*. Nous avons simplement recommandé d'examiner la possibilité de subdiviser deux classes fortement représentées et de recourir davantage aux subdivisions pour le repérage des notices dans une base de données automatisée, de façon à permettre une sélection relativement fine aux utilisateurs ayant un domaine particulier en tête. Le besoin de sélectivité n'est pas le même dans les publications imprimées, surtout celles qui paraissent à un rythme hebdomadaire comme *Jurisprudence Express*.

1.7.2 Les études effectuées sur l'utilisation du thésaurus

Le marquage des termes nécessaire à plusieurs de nos traitements, de même que le désir de mieux évaluer dans quelle mesure le thésaurus répondait aux besoins d'indexation tels qu'implicitement fixés par les conseillers juridiques, nous ont conduits à effectuer une série d'études complémentaires. Nous avons, par exemple, recherché les variantes morphologiques et les variantes syntaxiques, la présence de descripteurs et non-descripteurs à l'intérieur des mots-clés libres dans les manchettes (qui constituent environ 60% des mots-clés), étudié les structures les plus fréquentes pour la formation de ces mots-clés libres, fait la liste des descripteurs jamais employés ou jamais employés seuls, etc. Nous avons aussi tenu compte des cooccurrences des différents termes (descripteurs et non-descripteurs, mots clés libres) entre eux et avec les rubriques de classification. En outre, en calculant la force d'association des

termes avec les rubriques (selon une méthode qui tient compte de la fréquence), nous sommes désormais en mesure d'amorcer une structuration du vocabulaire par domaine de droit et donc de concevoir une réconciliation de deux outils documentaires (qui se recoupent et se contredisent parfois).

Bref, la richesse des analyses effectuées permet d'offrir une multitude de points de vue sur l'utilisation effective (plutôt que souhaitée lors de la conception) de ces outils, au fil des ans, par plusieurs personnes. Nous avons donc soumis à SOQUIJ non seulement un portrait des outils et des pratiques actuelles, mais des suggestions très détaillées pour l'enrichissement et la modification de ces outils et de ces pratiques.

Les résultats de nos différentes études nous ont amenés à conclure que le thésaurus devait être enrichi; d'abord pour contrôler une indexation qui s'avère, dans les faits, plus spécifique que ce que permet l'outil actuel, ensuite parce que le système expert doit pouvoir repérer toutes les formes possibles d'un descripteur dans les résumés et éventuellement, dans les textes intégraux pour les ramener aux formes souhaitées pour l'indexation, enfin parce que, une fois les descripteurs organisés selon une hiérarchie stricte, il devient possible d'opter pour différents niveaux de généricité selon les produits documentaires (en fonction notamment de leur périodicité, de leur support et de leur couverture du domaine).

2 La réalisation du prototype de système expert

La section précédente exposait la modélisation (méthode et résultats) effectuée pour l'aide à la sélection, à la classification, à la lecture et à l'indexation des jugements; la présente section traite de l'implantation réalisée du modèle. Les aspects suivants sont touchés : la tâche à informatiser; la motivation du choix de la technologie des systèmes experts; l'incertitude reliée à cette entreprise; l'arrimage du système expert avec l'analyse de textes par ordinateur; le design de la chaîne de traitement des documents; les aménagements apportés au traitement standard de l'incertitude; la réalisation de la base de règles par apprentissage et les problèmes laissés en suspens.

2.1 La tâche à informatiser

Comme on a pu le constater dans la section précédente, les tâches à informatiser présentent un haut niveau de complexité et sont accomplies dans un contexte de production. Rappelons que les publications de SOQUIJ connaissent des échéances et sont assujetties aux lois du marché. Ces tâches sont dites cognitives en ce que leur accomplissement requiert la mise en oeuvre particulière et discrétionnaire de connaissances et de stratégies générales accumulées durant l'exercice répété et supervisé des tâches mêmes. Pour leur réalisation, de nombreuses informations de source et de valeur diverses et parfois contradictoires doivent être recueillies et synthétisées. Par conséquent, une méthode mixte de modélisation a été déployée; il s'agit de faire converger les résultats d'une enquête cognitive auprès des conseillers juridiques, d'un traitement statistique de la distribution des indices et d'une analyse de texte plus qualitative.

La stratégie retenue est de recourir au plus grand nombre de sources de connaissances, identifiées lors de la modélisation, pour lesquelles des indices sont repérables dans les jugements. Par source de connaissance nous entendons, par exemple, la longueur du jugement, les lois qui y sont mentionnées, le tribunal qui a rendu le jugement, etc. Dans la mesure où ces sources de connaissances s'avèrent distinctes les unes des autres, il est possible de fonctionner avec un principe de convergence. Ainsi, on est d'autant plus certain qu'un jugement pointe vers le domaine «pénal» que son numéro de greffe comporte en deuxième section, l'une des combinaisons suivantes {01, 03, 10, 27 ou 36}, que ce jugement a été rendu dans la «Chambre criminelle et pénale»; que «La Reine» est une des parties impliquées et que le *Code criminel* y est mentionné, etc. Cette stratégie présente l'avantage de fonctionner, la plupart du temps de façon satisfaisante, dans des conditions de bruit. Le bruit étant essentiellement causé ici par les indices qui pointent vers plus d'un domaine.

2.2 Motivation du choix de la technologie des systèmes experts

Pour réaliser une implantation informatique du modèle cognitif obtenu, nous avons retenu la technologie des systèmes experts (SE) pour plusieurs raisons. L'implantation d'algorithmes incomplets et/ou sujets à de fréquentes révisions est possible car les règles d'inférences qui tiennent lieu des instructions d'un programme conventionnel sont indépendantes les unes des autres et leur enchaînement est assuré par un mécanisme général appelé moteur d'inférences. Il n'est donc pas nécessaire de prévoir à l'avance le déroulement complet de la solution définitive du problème : l'implantation peut être modulaire et évolutive. La réalisation d'un prototype se trouve à jouer un rôle heuristique en permettant d'achever la conception par des boucles de tests/ajustements en situation. La structure des règles d'inférences autorise une implantation quasi directe du modèle cognitif qui a été développé : un ou plusieurs indices détectés dans le texte du jugement (la prémisse) sont mis en relation avec une rubrique du plan de classification (la conclusion). De plus, la certitude de ces relations peut être qualifiée au moyen d'un coefficient numérique qui sera cumulé tout au long de la consultation. Ce «cumul» d'une part atténue la valeur des validations subséquentes lorsqu'une validation est affectée d'un coefficient incertain et, d'autre part, renforce la valeur d'une validation qui a déjà été réalisée. Le chaînage avant des règles permet enfin d'obtenir toutes les «réponses» valides et non une seule; un jugement peut donc être classifié dans plus d'un domaine avec une certitude différente pour chacun. Le découpage en règles d'inférence facilite la génération en contexte d'un rapport qui permet de valider les associations indices/rubrique du plan de classification, de localiser précisément les dysfonctionnements et finalement d'entraîner des conseillers juridiques novices.

2.3 Les incertitudes reliées à cette entreprise

Une fois la technologie des SE retenue en raison de caractéristiques qui apparaissaient souhaitables étant donné le projet, plusieurs incertitudes demeuraient; certaines ont été résolues lors de la réalisation du prototype et les solutions retenues

feront l'objet des prochaines sections. Une incertitude provenait de ce que les indices nécessaires pour la classification des jugements sont essentiellement de nature textuelle. Ainsi, contrairement aux situations habituelles de développement des SE, les indices ne sont pas fournis directement au système par l'utilisateur ou des senseurs. Cet état de fait implique le partage du traitement entre le SE pour interpréter les indices et un logiciel d'analyse de texte par ordinateur (ATO) pour les dépister dans le texte des jugements. De plus, ces indices dépistés par le logiciel d'ATO doivent être transformés pour être admissibles au SE. Une autre incertitude consistait à développer et implanter une chaîne de traitements qui soit conforme au traitement accompli par les conseillers juridiques, notamment en dépistant les mêmes types d'indices. La difficulté est double : le dépistage en lui-même et le regroupement des indices de nature différente. Une autre incertitude enfin était liée à l'utilisation des coefficients de certitude pour rendre compte du fait que les indices sont rarement totalement fiables.

En effet, un indice peut pointer vers plus d'un domaine du droit ou encore la présence d'un indice peut être considérée comme accidentelle et constituer en quelque sorte du «bruit». De plus, le mode de cumul des coefficients présente certains problèmes qui sont documentés. Certaines autres incertitudes sont toutefois demeurées, principalement parce que des recherches plus fondamentales dont l'envergure dépassait le mandat s'avèrent nécessaires; celles-ci sont présentées dans la dernière section.

2.4 L'arrimage du système expert avec l'analyse de textes par ordinateur

Le générateur de système expert (GSE) utilisé est l'Atelier Cognitif et TExtuel (ACTE) développé au Centre ATO.CI. Le développement de ACTE a démarré en février 1988, sur la commande d'un consortium de ministères et organismes québécois appelé DELTA; il s'agit d'une intégration logicielle de SATO et d'une version optimisée du D_expert (GSE en LISP) (Paquin et al. 1989). Cette intégration permet de faire du diagnostic textuel, c'est-à-dire de ne plus modéliser comme tel le contenu des textes, mais bien les opérations cognitives de

lecture et de compréhension, opérations qui sont en jeu pour la classification des jugements (Paquin 1992). La séquence qui a été retenue consiste à effectuer en lot une série de fichiers de commande SATO qui dépistent et identifient, principalement à l'aide de concordances, les différents types d'indices. Voici, par exemple, un extrait d'un tel fichier de concordances identifiant certaines requêtes qui, lorsque le jugement n'est pas motivé, entraînent le rejet :

```
Concordance stricte pension alimentaire
Concordance ordonnée rectification registre état civil
```

L'interface entre les traitements effectués par SATO et le SE se fait par la consignation dans un fichier du résultat - succès ou échec - de chacune des concordances. Ce fichier est alors traité pour ne conserver que les résultats positifs qui sont identifiés à l'aide d'une table, ce qui permet de normaliser le segment dépisté. Si l'on poursuit l'exemple précédent et que la deuxième concordance est réussie, peu importe la formulation exacte du segment dépisté, l'appellation normalisée sera transmise au SE.

2.5 Le design de la chaîne de traitement des documents

La modélisation cognitive de la tâche a été transformée en une suite séquentielle de traitement dont une schématisation est jointe en annexe.

La première étape consiste en un prétraitement qui est requis pour rendre les jugements admissibles à SATO. Reçus en format «WordPerfect», ils sont d'abord convertis en ASCII sans perdre les codes indiquant les attributs graphiques (gras, souligné, etc.) à l'aide d'un fichier de configuration d'imprimante élaboré à cet effet. Les codes de début et de fin deviennent des valeurs de la propriété **typo**:

```
(...) l'arrêt Laurentide Motels Ltd. c. Ville de Beauport, (...)
(...) l'arrêt *typo=+soul Laurentide Motels Ltd*typo=-soul. c.
*typo=+soul Ville de Beauport *typo=-soul (...)
```

Puis, un programme en ICON[3] procède à la désambiguïsation des marques de phrase et de paragraphe (Griswold et Griswold 1990). Le point marque habituellement la fin des phrases, mais il est aussi utilisé dans la notation de nombres décimaux, dans des sigles et il indique une abréviation. Surtout dans les domaines législatifs et administratifs, la mise en page d'une énumération ne se distingue que difficilement d'une suite de paragraphes. Enfin, certaines commandes nécessaires pour que le programme SATOGEN transforme le texte en matrice admissible à SATOINT sont ajoutées : l'alphabet, les séparateurs, les valeurs de la propriété **typo**.

Dans une deuxième étape, les indices textuels relatifs à chacune des sources de connaissances, sont dépistés, principalement par l'exécution de fichiers de commande SATO renfermant des concordances, à l'exception du numéro de greffe qui est dépisté par un programme *ad hoc* en ICON.

La troisième étape est celle de la mise en relation des indices dépistés avec les domaines du droit pertinents à l'aide du SE. Ce faisant, un rapport de la consultation est produit où sont consignés, pour chacune des sources de connaissance, les indices dépistés ainsi que les associations qui sont faites avec des domaines; une justification en contexte est, à l'occasion, fournie; un exemple de rapport est joint en annexe.

La quatrième étape, appelée assistance à la lecture, est optionnelle. Elle consiste en l'affichage des indices dépistés pour effectuer la tâche de classification ou encore d'autres indices. La distinction entre les types d'indices est produite par l'utilisation de couleurs différentes. Cette visualisation peut être effectuée selon un ou plusieurs profils. Les profils offerts actuellement ont été mentionnés plus haut, il s'agit des intervenants [Juge, cour, parties, etc.]; des sources du droit [lois, articles, jurisprudence, etc.] et des outils documentaires

[3] Rappelons qu'ICON est un langage du domaine public conçu principalement pour le traitement de chaînes de caractères. Il est supporté à l'Université d'Arizona.

[thésaurus et plan de classification]. Voici, à titre d'illustration un extrait de jugement :

Il s'agit d'une procédure assez exceptionnelle puisque, le requérant allègue certaines erreurs de droit du juge de paix. (...) Il y a évidemment ici la gravité objective de l'accusation. ° C'est une des plus sérieuses, une des plus graves que le Code criminel contient -- plutôt que la Loi des stupéfiants contient (...)

2.6 Les aménagements apportés au traitement standard de l'incertitude

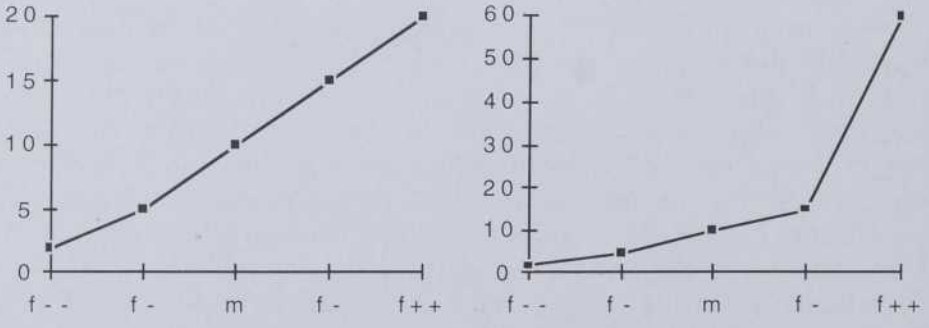
Chacun des indices peut pointer vers plusieurs domaines, avons-nous dit précédemment. De plus, une même confiance n'est pas accordée à toutes les relations établies entre les indices et les domaines. Les SE offrent la possibilité d'implanter des structures conditionnelles pondérées par des coefficients numériques, de même qu'une fonctionnalité pour leur cumul. Le cadre théorique le plus souvent utilisé est celui des coefficients de certitude développé pour le système Mycin (Buchanan et Shortliffe 1984). Rappelons que le principe du cumul des coefficients est le renforcement; en voici un exposé simplifié : si on arrive à une même conclusion à partir de deux sources de connaissances distinctes, on attribue à cette conclusion un coefficient supérieur au coefficient le plus élevé. Ce principe permet donc de discriminer les différents domaines du droit vers lesquels l'ensemble des indices repérés pointe.

Lorsqu'appliqué à notre système, ce cadre théorique pose toutefois deux ordres de problèmes[4]. Il a été démontré d'une part qu'il était très difficile pour des experts d'exprimer leur confiance dans les relations qu'ils établissent entre des faits vérifiés ou tenus pour vrais et des conclusions sous la forme d'un coefficient numérique. D'autre part, un calibrage des coefficients doit être effectué en fonction du nombre de renforcements qui sont susceptibles de se produire pour que l'effet discriminant soit optimal. En effet, si beaucoup de renforcements ont lieu alors que les coefficients sont élevés, plus le

[4] Nous tenons à remercier M. Claude Boivin, ministère du Revenu (Québec) pour le support théorique fourni dans ce développement.

résultat tend vers 100, plus il perd de la valeur discriminante; la valeur des coefficients ne doit pas être élevée. Par ailleurs, si les renforcements ne sont pas nombreux et que les coefficients sont très bas, les résultats ne seront pas convaincants. Pour remédier à ces deux problèmes, une approche modulaire a été développée. L'expression de la confiance quant à la relation entre les indices et les domaines du droit est séparée de l'algorithme de cumul par renforcement qui intervient lors d'une consultation.

Cette confiance est exprimée par des coefficients symboliques distribués sur une échelle bipolarisée qui comporte cinq valeurs : forte [f++], moyenne-forte [f+], moyenne [m], moyenne-faible [f-] et faible [f--]. Une conversion de ces coefficients «symboliques» en des coefficients numériques admissibles à l'algorithme de cumul. L'échelle numérique des coefficients est de 1 à 100. Cette fonction a deux rôles : exprimer l'écart entre les coefficients symboliques et ajuster leur valeur en fonction du nombre potentiel de renforcements. L'écart entre les coefficients détermine leur aspect discriminant. La figure de gauche montre une discrimination plutôt constante, celle qui est présentement implantée, la figure de droite montre une forte discrimination; l'utilisation du coefficient le plus élevé indique une relation prépondérante :



Le calibrage de la valeur numérique attribuée à chacun des coefficients symboliques en fonction du nombre potentiel de renforcements, se fait par essai-erreur. Des recherches supplémentaires sont requises pour déterminer une méthode exacte.

2.7 La réalisation de la base de règles par apprentissage

La technologie des SE ne présente comme tel aucun mode d'apprentissage, les règles d'inférences doivent être écrites et modifiées de la même manière : une à une à l'aide d'un éditeur spécialisé.

Afin de pallier cette carence, deux solutions ont été combinées : une approche tabulaire et une étude de corpus. Comme elles expriment des relations simples, les règles d'inférences peuvent s'exprimer sous forme de tableau, en trois colonnes : l'indice, le domaine et le coefficient de confiance, géré par une base de données ou un tableur; ainsi par exemple un extrait du tableau des numéros de greffe :

03	pena	f++	36	pena	f++	12	fami	f++
27	pena	f++	53	drli	m	04	fami	f++
01	pena	f++	06	proc	f+	43	fami	f++
10	pena	f++	41	fami	f++			

Le passage aux règles d'inférences est le fait d'un programme en ICON qui constitue la prémisse à partir de la première colonne, la conclusion à partir de la deuxième et opère le passage du coefficient symbolique en un coefficient numérique :

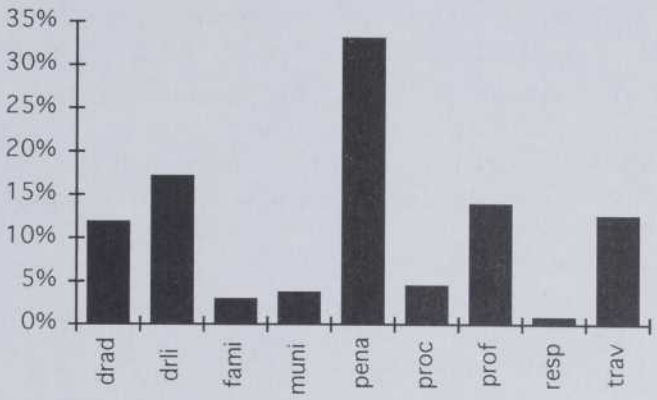
```
Connaissance Règle Définir 201 **
 Note "TRI -> 1ère page -> no de greffe : 03" **
 Auteur automatik **
 Création 1904-01-01 00-00-00 **
Si **
 Base TRI **
 Granule "Indices de première page" **
  ( Trait "section du no de greffe" =   Chaîne "03" ) **
Alors **
 Base TRI **
  Granule Document **
  ( Trait domaine =   Chaîne "Pénal" Coef 20 ) **
CanalEcrire ( Canal rapport **
  Message " Ce no. de greffe pointe vers le domaine " **
    Base TRI **
     Granule Document **
```

```
      Trait domaine **
  Message " avec une confiance forte" **
  Message " ; cumulatif de : " Coefficient **
  Message "%"  Retour **
)
```

L'étude d'un corpus constitué de jugements déjà classifiés a permis de littéralement découvrir des indices pour la plupart des sources de connaissances. À titre d'illustration, le cas des lois citées sera décrit. Pour chacun des domaines du droit, un sous-texte a été constitué de tous les passages en italique à l'aide de SATO. Ces sous-textes ont été épurés de façon à ne contenir que les lois citées dans les jugements et ont été constitués en tableaux avec le domaine attribué :

Loi de l' aménagement	municipal
Loi de l' assistance publique	municipal
Loi de l' évaluation foncière	municipal
Loi de la qualité de l' environnement	municipal
Loi de police	municipal

Ensuite, ces tableaux ont été fusionnés et triés, de façon à regrouper pour chacune des lois tous les domaines pointés. Cette distribution, suite à une validation par les conseillers juridiques pour éliminer les aberrations, a guidé l'attribution des coefficients de confiance. La règle suivie est que si la distribution est égale, un coefficient faible est attribué, sinon la force du coefficient est proportionnel à la distribution. Ainsi, par exemple, la *Charte canadienne des droits et libertés* apparaît dans les domaines suivants avec cette distribution :



À partir de cette distribution, les coefficients suivants ont été attribués :

Pénal (pena)	f++
Droits et libertés (drli)	f+
Droit administratif (drad)	m
Professions (prof)	m
Travail (trav)	m
Famille (fami)	f--
Municipal (muni)	f--
Procédure civile (proc)	f--
Responsabilité (resp)	f--

À la suite de ces opérations, on obtient un tableau en trois colonnes qui permet de générer les règles d'inférences. Les indices qui pointent vers plusieurs domaines sont regroupés dans une même règle.

2.8 Les problèmes laissés en suspens

Malgré tous les efforts déployés, des difficultés ont été laissées en suspens, parce qu'elles demandent des recherches dont l'envergure dépassait le mandat, mais dont l'intérêt apparaît évident étant donné le succès du prototype, par exemple :

- l'intégration du coefficient de confiance dans le résultat obtenu par l'algorithme d'analyse discriminante utilisé au cumul des coefficients de certitude des règles d'inférences;
- l'intégration des occurrences différentes des termes lemmatisés du plan de classification et du thésaurus dans le cumul des coefficients de certitude et la prise en compte de leur fréquence. Est-ce que plusieurs termes différents pointant vers un même domaine valent plus cher que des fréquences élevées de quelques termes ?
- la modification du modèle de cumul des coefficients par renforcement qui ne permet que l'accroissement linéaire, pour prendre en compte des indices invalidant un ou plusieurs domaines, ce qui serait plus conforme au fonctionnement cognitif des conseillers juridiques.

Conclusion

L'expérience que nous venons d'exposer a permis à l'équipe de recherche de vérifier : 1) qu'il est possible de modéliser les opérations cognitives des experts dans diverses situations de lecture à partir d'une enquête cognitive et d'une analyse sémio-statistique des textes analysés et des résultats de plusieurs types d'analyses; 2) qu'il est possible d'implanter un système expert s'appuyant sur des stratégies dépistant dans les textes certains des indices détectés par les humains, à différents niveaux d'organisation des textes (éditorial, morpho-syntaxique, intra-phrastique, intra- et inter-textuel, sémantique, pragmatique, etc.). Il a également été constaté que plusieurs de ces indices sont utilisables pour faciliter la lecture selon les objectifs poursuivis par chacune des opérations de sélection, de tri-classification et d'indexation. Pour cela, il faut disposer d'un logiciel qui, non seulement autorise le marquage des unités textuelles et lexicales selon autant de caractéristiques que les hypothèses le suggèrent, mais aussi facilite auparavant la découverte de ces propriétés puis leur manipulation au même titre que la manipulation des chaînes de caractères. Nous avons également appris que la performance de nos méthodes de modélisation/formalisation était optimale lorsque l'information requise par les tâches cognitives se

trouvait dans les textes sous la forme d'indices repérables. Ainsi, même lorsque la prise de connaissance du contenu semble superficielle (par exemple pour la sélection), si la prise de décision fait appel à des connaissances autres que celles de la langue et du cadre textuel, la tâche échappe en grande partie à nos méthodes. Par conséquent, comme les tâches requièrent pour la plupart de telles connaissances, les experts doivent toujours garder le contrôle des systèmes. En ce qui concerne l'implantation de la chaîne de traitement, nous avons constaté que le succès reposait moins sur la complexité de la technologie ou des formules mathématiques que sur la maîtrise une à une de chacune des sources de connaissances requises et des indices qui les désignent dans les textes.

Ajoutons qu'un des bénéfices importants de la recherche a consisté dans le portrait des politiques et procédures d'analyse suivies par la dizaine de conseillers juridiques telles que révélées par l'analyse des données et l'enquête cognitive, dans la constatation de quelques divergences dont certaines devaient être corrigées pour accroître la prédictibilité des index, et enfin, dans la production d'un thésaurus considérablement enrichi et l'amorce d'un meilleur arrimage entre thésaurus et plan de classification.

Remerciements

Le projet a été soutenu financièrement par les institutions suivantes : Centre francophone de recherche en informatisation des organisations (CEFRIO), Société québécoise d'information juridique (SOQUIJ), Ministère des Communications du Québec, École de bibliothéconomie et des sciences de l'information, Université de Montréal, Centre de recherche en cognition et information ATO.CI, Université du Québec à Montréal.

Plusieurs personnes ont été impliquées à diverses étapes du projet : Jean-Guy Meunier, directeur du Centre ATO.CI; Sylvie Michaud, bibliothécaire professionnelle; Myriam Desclos-Lalaude, stagiaire de l'I.E.P. (Cycle supérieur de spécialisation en information et documentation, Institut d'Études Politiques), Paris; Luc Dupuy, agent de recherche, centre ATO.CI, Yves Khawam, professeur adjoint, ÉBSI et plusieurs étudiants en

bibliothéconomie et sciences de l'information ainsi qu'en linguistique.

Bibliographie

Beacco, J.-C. et M. Darot, 1984, *Analyse de discours; lecture et expression*, Paris, Hachette / Larousse.

Beghtol, C., 1986, Bibliographic classification theory and text linguistics : aboutness analysis, intertextuality and the cognitive act of classifying documents, *Journal of Documentation*, 42(2), pp. 84-113.

Bertrand, A., 1993, *Compréhension et catégorisation dans une activité complexe : l'indexation de documents scientifiques,* Université de Toulouse-Le Mirail, Équipe de psychologie du travail ER 15- CNRS. (Thèse de doctorat).

Bertrand-Gastaldy, S., 1993, Analyse documentaire et intertextualité, *Les Sciences du texte juridique: Le droit saisi par l'ordinateur*, sous la direction de Claude Thomasset, René Côté et Danièle Bourcier, Cowansville, Les Éditions Yvon Blais, pp. 139-173.

Bertrand-Gastaldy, S., F. Daoust, G. Pagola et L.-C. Paquin, 1993, *Conception d'un prototype de système expert d'aide à l'analyse des jugements : rapport final présenté à SOQUIJ, vol. 1, synthèse des travaux*, [Montréal], Université de Montréal, École de bibliothéconomie et des sciences de l'information / Université du Québec à Montréal, Centre de recherche en information et cognition ATO.CI.

Bertrand-Gastaldy, S., L. Giroux, D. Lanteigne et C. David, 1994, Les produits et processus cognitifs de l'indexation humaine, *ICO Québec*, 6(1-2), pp. 29-40.

Blosseville, M.J., G. Hébrail, M.G. Monteil et N. Pénot, 1992, Automatic Document Classification : Natural Language Processing, Statistical Analysis and Expert System Techniques Used Together, dans *SIGIR 92, Proceedings of the Fifteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Copenhagen,* pp. 51-57.

Buchanan, B. G. et E.H. Shortliffe, 1984, *Rule-Based Expert Systems: The Mycin Experiments of the Stanford Heuristic Programming Project,* Reading, MA.

Chaumier, J. et M. Dejean, 1990, L'indexation documentaire : de l'analyse conceptuelle humaine à l'analyse automatique morpho-syntaxique, *Documentaliste* , 27(6), pp. 275-279.

David, C., 1990, *Élaboration d'une méthodologie d'analyse des processus cognitifs dans l'indexation documentaire,* Montréal, Université de Montréal, Département de communication. (Mémoire de maîtrise).

Doszkocs, T., 1986, Natural language processing in information retrieval, *Journal of the American Society for Information Science,* 37(4), pp. 191-196.

Endres-Niggemeyer, B., 1990, A procedural model of abstracting, and some ideas for its implementation, dans *TKE'90; Terminology and Knowledge Engineering*, Frankfurt, Indeks Verlag, pp. 230-243.

Farrow, J., 1991, A cognitive process model of indexing document, *Journal of documentation,* 47 (2), pp. 149-166.

Griswold, R. et M. Griswold, 1990, *The Icon Programming Language*, New York, Prentice Hall.

Meunier, J.-G., S. Bertrand-Gastaldy et H. Lebel, 1987, A call for enhanced representation of content as a means of improving on-line full-text retrieval, *International Classification*; 14(1), pp. 2-10.

Meunier, J.-G., 1992, SATO : un philologue électronique, *Documentation et bibliothèques*, 38(2), pp. 65-69.

Meunier, J.-G., S.Bertrand-Gastaldy et L.-C. Paquin, 1994, La gestion et l'analyse des textes par ordinateur : leur spécificité dans le traitement de l'information, *ICO Québec*, 6(1-2), pp. 19-28.

Paquin, L.-C., L. Dupuy et F. Daoust, 1989, ACTE : a workbench for knowledge engineering and textual data analysis in the social sciences, in *Proceedings of the Fourth International Conference on Symbolic and Logical Computing (ICEBOL4)*, Madison, Dakota State University, pp. 122-136.

Paquin, L.-C., 1992, La Lecture experte, *Technologie, idéologie et pratique*, (10) 2-4, pp. 209-222.

Van Dijk, T. A., 1977, Perspective paper : complex semantic information processing, dans D.E. H. Karlgren, H. et M. Kay, *Natural Language in Information Science; Perspectives and Directions for Research,* Stockholm, Skriptor, pp.127-163.

Chaîne de traitement combinant analyse de texte et système expert

SATO

Prétraitement

- conversion WordPerfect -> ASCII
- marquage
  . des paragraphes
  . des phrases
- SATOGEN

Description

- longueur
- le type de requête
- le no. de greffe
- le tribunal
- les parties
- l'intitulé
- lois mentionnées
- art. du C.c. mentionnés
- par les descripteurs

système expert

SÉLECTION

- longueur > 40 lignes ou > 400 mots
- type de requête

CLASSIFICATION Indices de 1ère page

- le no. de greffe
- le tribunal
- les parties
- l'intitulé

CLASSIFICATION Sources mentionnées

- lois
- articles du Code civil

CLASSIFICATION Distribution

- statistique lexicale
- documentaire

Assistance à la lecture

- dépistage d'information selon des profils
- surlignage

EXEMPLE DE RAPPORT FOURNI PAR LE SYSTEME EXPERT

SYSTÈME EXPERT POUR SÉLECTIONNER ET CLASSIFIER LES JUGEMENTS

prototype pour SOQUIJ, le CEFRIO et le MCQ
Suzanne Bertrand-Gastaldy resp., Gracia Pagola
EBSI : École de bibliothéconomie et des sciences de l'information, UdeM
François Daoust et Louis-Claude Paquin
Centre ATO-CI : Centre de recherche en cognition et information à l'UQAM

Le système expert a pour tâche principale de sélectionner les jugements
et de désigner les rubriques de classification les plus probables.

ÉTAPE DE LA SÉLECTION
Les jugements sont retenus pour traitement selon les critères suivants : la longueur dont le seuil est de 40 lignes ou 400 mots le type de requête.
Ce jugement compte 179 lignes et 2002 mots.
Le jugement est sélectionné. il a la longueur suffisante; il n'est pas répertorié parmi les requêtes rejetées.

ÉTAPE DU TRI-CLASSIFICATION
Cette étape comporte quatre analyses :
- les indices de la première page;
- les lois et articles du code civil mentionnés;
- la discrimination lexicale;
- la discrimination par les outils documentaires.

A. L'analyse des indices de la première page du jugement touche les indices suivants :
1) le numéro de greffe : 500-05-001562-899
- ne pointe vers aucun domaine
2) le tribunal : COUR SUPÉRIEURE
- ne pointe vers aucun domaine
3) Le nom des parties
- «La Reine» pointe vers le domaine Pénal avec une confiance forte ; cumulatif de : 20%
4) L'intitulé du jugement

```
 - ne pointe vers aucun domaine

 B Lois et articles du Code civil mentionnés
 «code criminel»
 - pointe vers le domaine Municipal avec une confiance faible ;
cumulatif de : 2%
 - pointe vers le domaine Pénal avec une confiance forte ; cumulatif
de : 36%
 - pointe vers le domaine Professions avec une confiance faible ;
cumulatif de : 2%
 «L'article 614.3 C.C.»
 - pointe vers le domaine Famille avec une confiance forte ;
cumulatif de : 20%

 C Analyse de la discrimination documentaire

 Cette analyse est effectuée par la projection des termes appartenant
au thésaurus et au plan de classification.

1 descripteur(s) pointe(nt) vers le domaine Droits et libertés
1 descripteur(s) pointe(nt) vers le domaine Sûretés
1 descripteur(s) pointe(nt) vers le domaine Travail
2 descripteur(s) pointe(nt) vers le domaine Responsabilité
3 descripteur(s) pointe(nt) vers le domaine Procédure civile
24 descripteur(s) pointe(nt) vers le domaine Pénal

__________ F i n _ d u _ t r a i t e m e n t __________
```

Liste des auteurs

SUZANNE BERTRAND-GASTALDY est professeur titulaire à l'École de bibliothéconomie et des sciences de l'information de l'Université de Montréal où elle est responsable du profil «Analyse de l'information et bases de données». Ses recherches portent essentiellement sur les méthodes d'indexation et d'élaboration de vocabulaires de domaine assistées par ordinateur ainsi que sur l'évaluation des thésaurus à la lumière des théories sémio-cognitives.

DENIS BOUCHARD est professeur au Département de linguistique de l'Université du Québec à Montréal. Ses principaux thèmes de recherche sont la sous-détermination de la syntaxe étant donné l'apport de la sémantique et la sous-détermination de la sémantique étant donné l'apport du contexte cognitif.

LORNE H. BOUCHARD est professeur au Département d'informatique de l'Université du Québec à Montréal. Son domaine de recherche est le traitement automatique de la langue naturelle en vue de la construction de systèmes à base de connaissances. Il est co-responsable du Groupe Interdisciplinaire de Recherche et d'Études en Informatique Linguistique (GIREIL).

HENRIETTA CEDERGREN est professeur au Département de linguistique de l'Université du Québec à Montréal. Elle œuvre dans les domaines de recherche suivants : la sociolinguistique et la relation entre la phonologie et la phonétique.

LAURENCE DANLOS est professeur de linguistique informatique à l'Université de Paris 7 et directrice de l'équipe de recherche TALANA

(Traitement Automatique du LAngage NAturel). Ses domaines de recherche sont principalement la génération de textes, la traduction automatique et la sémantique lexicale.

FRANÇOIS DAOUST est informaticien et professeur associé au Département de linguistique de l'Université du Québec à Montréal. Il travaille au sein du Service ATO (analyse de texte par ordinateur). Il est le concepteur du logiciel d'analyse de texte SATO.

LYNE DA SYLVA est étudiante au doctorat au Département de linguistique et de traduction de l'Université de Montréal et assistante de recherche au Département de linguistique de l'Université du Québec à Montréal.

ANNE-MARIE DI SCIULLO est professeur au Département de linguistique de l'Université du Québec à Montréal. Ses principaux thèmes de recherche sont la théorie linguistique, la syntaxe, la sémantique, la morphologie, la linguistique computationnelle et les analyseurs morphologiques pour les langues naturelles.

ANDRÉ DUGAS est professeur au Département de linguistique de l'Université du Québec à Montréal. Ses recherches sont reliées aux industries de la langue en vue notamment de la construction de dictionnaires électroniques.

FERNANDE DUPUIS est professeur associée au Département de linguistique de l'Université du Québec à Montréal et poursuit des recherches avec les membres du Service ATO. Ses travaux portent sur l'évolution de la syntaxe, en particulier sur les changements dans la syntaxe de l'ancien et du moyen français. Elle s'intéresse également aux problèmes de codification et d'analyse de corpus à l'aide d'outils informatiques.

MARC DYMETMAN, détenteur d'un doctorat d'état en informatique, a été chercheur en linguistique informatique au CITI (Centre d'innovation en technologies de l'information) entre 1987 et 1994. Il occupe présentement un poste de chercheur au Rank Xerox Research Center de Grenoble.

LOUISETTE EMIRKANIAN est professeur au Département de linguistique de l'Université du Québec à Montréal. Elle travaille dans le domaine de la linguistique informatique et s'intéresse plus particulièrement à la syntaxe et aux grammaires d'unification.

MASSIMO FASCIANO est étudiant au doctorat en informatique à l'Université de Montréal. Il travaille en génération de texte dans le groupe Scriptum du laboratoire Incognito. Son sujet de thèse porte sur la combinaison harmonieuse de textes et de graphiques pour des

rapports statistiques et sur la génération automatique de ces deux aspects.

GEORGE FOSTER, détenteur d'un M.Sc. en informatique, est chercheur en linguistique informatique au CITI depuis 1992. Il est en outre inscrit au programme de Ph.D. en informatique de l'Université McGill.

CHRISTOPHE FOUQUERÉ, professeur d'informatique à l'Université de Paris-Nord, s'est d'abord intéressé aux problèmes de correction de texte écrit. Depuis quelques années, il a orienté sa recherche vers la validation de grammaires du Langage Naturel, parallèlement à l'étude des logiques nécessaires à la représentation des connaissances. Dans ce cadre, il utilise la Logique Linéaire pour représenter la connaissance tant de sens commun que celle liée au traitement syntaxique de la langue.

PIERRE ISABELLE, détenteur d'un doctorat en linguistique, œuvre dans le domaine de la recherche en traduction automatisée depuis plusieurs années. Entre 1975 et 1981, il fut membre puis directeur scientifique du groupe de recherches en traduction automatique de l'Université de Montréal (TAUM). Depuis 1985, il est Responsable du programme de traduction assistée par ordinateur du CITI (Industrie Canada).

JEAN-MARC JUTRAS, détenteur d'un M.Sc. en linguistique, est linguiste-informaticien au CITI depuis 1990.

BETSY KLIPPLE a participé au projet de Traitement linguistique parallèle en tant que professionnelle de recherche au Département de linguistique de l'Université du Québec à Montréal ; son domaine de spécialité est la morphologie.

GUY LAPALME est professeur au Département d'informatique et de recherche opérationnelle à l'Université de Montréal. Dans le cadre du groupe Scriptum, il travaille depuis une dizaine d'années dans le domaine de la génération de texte. Il a dirigé plusieurs maîtrises et doctorats dans ce domaine. Il s'intéresse également à la programmation fonctionnelle et aux applications dans le domaine de la bio-informatique.

CHRISTINE LEONHART, diplômée en traduction de l'Université Laurentienne à Sudbury (Ontario), a occupé un poste de terminologue au Bureau de la traduction du gouvernement fédéral pendant cinq ans. En 1984, elle s'est jointe au groupe de travail responsable du développement de TERMIUM® III, la banque de données linguistiques du gouvernement du Canada. Par la suite, elle a participé à la mise en œuvre de la banque de données, à la formation des utilisateurs et à la création de la version sur CD-ROM de la base de données. Ses efforts

sont maintenant centrés sur le développement de TERMIUM® et d'un poste de travail pour les terminologues du Bureau de la Traduction.

FRANÇOIS LÉVEILLÉ, étudiant à la maîtrise en informatique à l'Université du Québec à Montréal, a collaboré au projet de Traitement linguistique parallèle en qualité de professionnel de recherche au Département d'informatique de l'Université du Québec à Montréal.

ELLIOT MACKLOVITCH œuvre dans le domaine de la traduction assistée par ordinateur depuis plus de quinze ans. Il était Chargé de projets en TA au Bureau de la traduction du Secrétariat d'État, où il a également travaillé comme traducteur vers l'anglais. Ancien membre du groupe TAUM de l'Université de Montréal, il a déjà été chercheur invité au Groupe d'Études pour la Traduction Automatique (GETA). Il est présentement chercheur, chargé de la coordination des projets au sein du groupe TAO au CITI.

GRACIA PAGOLA est détentrice d'une maîtrise en bibliothéconomie et sciences de l'information et a fait des études universitaires en linguistique et informatique. Elle est chargée de cours à l'EBSI et au Cégep Maisonneuve. Agente de recherche pour plusieurs projets d'indexation et de contrôle de vocabulaire assistés par ordinateur, elle est coauteur de publications sur ce sujet et a travaillé à la constitution de plusieurs bases de données bibliographiques et en texte intégral.

LOUIS-CLAUDE PAQUIN est professeur au Département des communications de l'Université du Québec à Montréal. Il couvre le domaine dit des nouvelles technologies de l'information et de la communication (NTIC). Sa recherche et sa production portent sur le multimédia interactif, autant sur CD-ROM que sur Internet. Il est directeur d'un laboratoire consacré aux technologies interactives. Auparavant, il a fait partie d'un centre de recherche en analyse de textes par ordinateur où il a participé au développement de systèmes experts en analyse de textes.

FRANÇOIS PERRAULT, détenteur d'un M.Sc. en informatique, a travaillé au CITI comme informaticien-linguiste entre 1988 et 1995. Il est présentement à l'emploi de la société Bunyip.

HÉLÈNE PERREAULT est détentrice d'un diplôme en linguistique de l'Université du Québec à Montréal. Elle a travaillé auprès d'Henrietta Cedergren sur le français parlé à Montréal en tant que chercheur. La prosodie est son domaine de recherche et l'organisation de la durée, sa spécialité.

XIAOBO REN, détentrice d'un doctorat en linguistique de la Sorbonne, a travaillé au CITI comme linguiste-informaticienne entre 1991 et 1995.

CÉLINE ROBITAILLE, en tant que professionnelle de recherche au Département de linguistique de l'Université du Québec à Montréal, a contribué à l'élaboration du module lexique pour le prototype de Traitement linguistique parallèle.

MICHEL SIMARD a obtenu son M.Sc. en informatique en 1991. Depuis cette date il est informaticien-linguiste au CITI.

JAN VAN VOORST a participé au projet de Traitement linguistique parallèle en tant que professionnel de recherche au Département de linguistique de l'Université du Québec à Montréal ; son domaine de spécialité est la sémantique.

Achevé d'imprimer en mai 1996 chez
VEILLEUX
IMPRESSION À DEMANDE INC.
à Boucherville, Québec

Données de catalogage avant publication (Canada)

Vedette principale au titre :

Traitement automatique du français écrit : développements théoriques et applications

(Les cahiers scientifiques de l'Acfas ; 86)
Textes présentés lors d'un colloque tenu à l'Université du Québec à Montréal le 17 mai 1994.
Comprend des réf. bibliogr.

ISBN 2-89245-141-8

1. Français (Langue) - Français écrit - Informatique - Congrès. 2. Français (Langue) - Analyse du discours - Informatique - Congrès. 3. Français (Langue) - Grammaire - Informatique - Congrès. 4. Traitement automatique des langues naturelles - Congrès. 5. Linguistique - Informatique - Congrès. 6. Traduction - Informatique - Congrès. I. Emirkanian, Louisette. II. Bouchard, Lorne H. III. Association canadienne-française pour l'avancement des sciences. IV. Collection: Les Cahiers de l'Acfas ; 86.

PC2074.5.T72 1996 448'.00285 C96-940436-0

Le traitement automatique de la langue écrite est un domaine de recherche multidisciplinaire qui prend sa source essentiellement dans les recherches en linguistique et en informatique.

Compte tenu du rôle central de la faculté langagière parmi les processus cognitifs d'ordre supérieur, l'analyse et la génération de l'écrit constituent des domaines de recherche fondamentaux pour le développement des systèmes à base de connaissances. L'étude de la problématique du traitement automatique de l'écrit montre que l'évolution des connaissances dans ce domaine a un impact qualitatif important sur la nature de ce qui peut être automatisé et sur le partage effectif des efforts entre personne et machine. Les articles de ce recueil font état de travaux sur le traitement automatique du français écrit.

9 782892 451412

ISBN 2-89245-141-8