



Modélisation différentielle du texte, de la linguistique aux algorithmes

Nadine Lucas

► To cite this version:

Nadine Lucas. Modélisation différentielle du texte, de la linguistique aux algorithmes. Traitement du texte et du document. Université de Caen, 2009. tel-01073406

HAL Id: tel-01073406

<https://theses.hal.science/tel-01073406>

Submitted on 9 Oct 2014

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Modélisation différentielle du texte

de la linguistique aux algorithmes

Nadine Lucas

23 juin 2009

mémoire d'habilitation à diriger des recherches

Université de Caen Basse-Normandie

Jury :

Ioannis Kanellos, Professeur d'informatique, Telecom Bretagne, rapporteur

Anne Nicolle, Professeure d'informatique émérite, Université de Caen, rapporteur

Violaine Prince, Professeur d'informatique, Université de Montpellier 2, rapporteur

Jean-Gabriel Ganascia, Professeur d'informatique, Lip6

Sophie Moirand, Professeure de linguistique, Université Paris 3 Sorbonne Nouvelle

Laurence Rosier, Professeur de linguistique, Université libre de Bruxelles

REMERCIEMENTS

Ce travail est le fruit d'un long parcours et de multiples influences auxquelles je ne saurais rendre pleinement justice. Je voudrais évoquer d'abord avec reconnaissance la mémoire de Jean-Jacques Origas, dont l'influence fut décisive dans ma formation à l'Institut national des langues et civilisations orientales.

Je voudrais exprimer ma reconnaissance aux nombreux acteurs qui ont bien voulu jouer le rôle d'informants, des auteurs d'articles aux utilisateurs de logiciels, ainsi qu'aux étudiants sagaces et aux collaborateurs qui m'ont posé des questions judicieuses.

Les personnes qui ont échangé leurs connaissances au cours des années sont nombreuses et je voudrais leur témoigner ma gratitude ; pour leurs discussions fructueuses dans le domaine de la linguistique japonaise, en particulier Catherine Garnier à l'INALCO, puis à l'EHESS, Irène Tamba, Martine Prost, Patrik Le Nestour, Satô Naomiki. Dans le domaine de la linguistique, j'ai eu la chance de rencontrer sur mon chemin Lone Takeuchi, Nishina Kikuko, Sophie Moirand, Marie-Paule Péry-Woodley, Georges Vignaux, Tae Suzuki, Jacqueline Authier-Revuz, Dominique Kinsler, Laurence Rosier, Mai Ho-Dac et d'autres encore qui m'ont encouragée et souvent confortée au-delà des différences d'approche.

Mes premiers pas en informatique linguistique ont été guidés par Patrice Pognan, Michel Fanton, Jean-Paul Horn et Jacques Vergne, puis par Tanaka Hozumi et son équipe. Par la suite en informatique, j'ai bénéficié des connaissances de nombreux collègues dont Violaine Prince, Jean-Marc Le Carpentier, Arnaud Soulet et Bruno Crémilleux. Hervé Déjean reste pour moi une référence en informatique linguistique, nous avons partagé de passionnantes discussions sur les automates. Quant à Dominique Dutoit, il a toujours des sujets de dissension à me proposer pour des disputes roboratives.

Pour l'épistémologie, la philosophie et d'autres questions fondamentales, je remercie Sean O'Nuallain, ainsi que Davy Spillane, Jean-Baptiste Berthelin, Asari Makoto, et Hervé Le Crosnier.

Je salue pour son inlassable curiosité Nathalie Cousin, dont le soutien jamais démenti m'est précieux.

René Fréreux, Pascal Buléon et les membres du groupe RIAS à l'université de Caen proposent de salutaires et nourrissantes réflexions sur l'activité scientifique. Les équipes de recherche pluridisciplinaires GeoSem, Bingo et Calico sont également des terreaux de discussion que j'apprécie. Les réalisations concrètes qui en découlent sont riches d'enseignement. J'ai bénéficié et bénéficie d'un cadre de travail stimulant au GREYC, que tous en soient remerciés. Ali Akhavi, Bruno Zanuttini et Frédérique Loew ont répondu avec entrain à mes nombreuses questions. Jurek Karczmarcuk en a suscité de nombreuses.

Je remercie tout particulièrement Anne Nicolle pour son accueil chaleureux au GREYC et ses questionnements au sein de l'équipe ISLand. Elle a montré la voie par son engagement dans le rôle de « passeur » de connaissances à la frontière des disciplines. Je lui dois aussi son soutien actif pour cette habilitation. Je remercie Bernard Morand pour son aide en génie logiciel et sa détermination à aborder les questions les plus difficiles. Eric Bruillard m'a donné l'occasion de mettre en œuvre certaines de mes idées et de les confronter à un public critique.

Je remercie spécialement Jacques Vergne pour son écoute depuis mes débuts en recherche, qui coïncident avec les siens, sa capacité à trouver des solutions en reformulant les problèmes différemment, et son impaviderité face au tumulte du monde. Je suis également très redevable à Emmanuel Giguet, ne serait-ce que pour son talent à retenir l'essentiel de nos discussions et à le traduire en langage informatique. Plus que le temps record de ses implémentations, c'est l'élégance dans la conception d'algorithme qui le rend indispensable. Il a été mon meilleur étudiant et mon meilleur enseignant, nous avons partagé un travail considérable, des échecs et des succès.

Enfin, merci aux amis et proches qui ont patiemment supporté le temps de rédaction et m'ont apporté leur soutien, tandis que je leur faisais défaut.

Maria-Caterina Manes-Gallo, Daniel Luzzati et Roland Hausser ont proposé de relire mon manuscrit. Bernard Morand et Dominique Dutoit m'ont fait des remarques dont je leur suis reconnaissante. Benoît Habert a posé des questions pertinentes à toutes les pages.

Je remercie les rapporteurs qui ont bien voulu lire attentivement ce mémoire. Je remercie enfin les membres du jury d'avoir accepté de juger d'un travail qui n'est ni de la linguistique, ni de l'informatique et de se pencher sur un parcours atypique.

TABLE DES MATIERES

Chapitre 1 Introduction : les enjeux de l'étude des textes	7
1. Contexte et positionnement	7
2. Objectifs	12
3. Choix et définitions	14
4. Contributions	17
Chapitre 2 Le contexte et les choix épistémologiques	23
1. Des définitions conflictuelles	23
2. Types et unités d'analyse de texte	27
3. Les interprétations	37
4. Usages d'un modèle pour les applications informatiques	40
5. Conclusion partielle	49
Chapitre 3 Multilinguisme et analyse automatique de textes	51
1. L'approche à base de patrons ancrés	52
2. La détection du discours rapporté dans la presse	52
3. Détection de la structure thématique dans les articles de presse	65
Chapitre 4 La gestion de l'échelle dans les applications Internet	75
1. L'approche inductive en extraction d'information sur la toile	75
2. La fouille de collections de textes	79
3. L'analyse des forums	86
Chapitre 5 Perspectives et questions ouvertes	93
1. Les acquis scientifiques	94
2. Evaluation	95
3. Les questions à approfondir	97
Chapitre 6 Perspectives en enseignement et recherche	99
1. Méthodologie	100
2. Angles d'attaque	101
3. Retombées applicatives attendues	103
Références bibliographiques	109

Liste des figures et tableaux

Figure 2.1 Schématisation de l'organisation du discours	p. 31
Figure 2.2. Les fonctions du langage selon Jakobson LL p. 71, disposition sur la page imprimée	p. 32
Figure 2.3. Les fonctions du langage selon Jakobson reprises par Sciuto	p. 32
Figure 2.4. Les fonctions du langage selon Jakobson et l'interprétation de la dominante en trois dimensions par Lucas	p. 32
Figure 2.5 Analyse de Mann & Thompson (1992) pour un chapeau d'article	p. 35
Figure 2.6. Segmentation de Grosz & Sidner (1986) pour un essai court	p. 35
Figure 2.7. La macro-structure de Van Dijk pour le genre journalistique	p. 38
Tableau 2-1 Quelques courants d'analyse en linguistique	p. 40
Tableau 2-2 Modèles repris en informatique linguistique	p. 40
Figure 2.8. <i>Schemata</i> des constructions canoniques	p. 44
Figure 2.9. Constructions canoniques et points de vérification théoriques	p. 49
Figure 3.1. Coloriage d'une dépêche en fonction du type de discours rapporté	p. 55
Figure 3.2. Coloriage d'une dépêche avec DR et absence de commentaire	p. 55
Figure 3.3. Détection de citations et de chaîne de citation coréférentielle	p. 59
Figure 3.4. Position des maillons de chaînes de citation référant à un informant unique	p. 61
Figure 3.5. Détection des citations avec un informant et rattachement à une chaîne	p. 61
Figure 3.6. Détection de chaînes de citation et de l'informant pour chaque chaîne dans un article à plusieurs informants	p. 62
Tableau 3-3 Détection de citations dans des dépêches financières par EDDAP2	p. 63
Figure 3.7. Détection réussie sur un article mono locuteur avec une chaîne	p. 64
Figure 3.8. Détection incertaine sur un article avec trois chaînes et plusieurs informants	p. 64
Tableau 3-4 Mesure d'efficacité sur la détection et la qualification des citations	p. 65
Tableau 3-5 Comparaison de la qualification des locuteurs des citations et de la qualification des informants des chaînes	p. 65
Figure 3.9. Diagramme UML structure des données du modèle	p. 68
Figure 3.10. Segmentation et hiérarchisation d'un article de vulgarisation par THEMA	p. 68
Figure 3.11. Exemple d'exposé en français hiérarchisé par Unithem	p. 69
Figure 3.12. Exemple d'explication en français hiérarchisé par Unithem	p. 70
Figure 3.13. Exemple d'exposé en italien hiérarchisé par Unithem	p. 70
Figure 3.14. Exemple d'exposé en russe hiérarchisé par Unithem	p. 71
Figure 3.15. Exemple d'exposé en arabe hiérarchisé par Unithem	p. 71
Figure 3.16. Exemple d'exposé en japonais hiérarchisé par Unithem	p. 72
Figure 4.1. Interface de <i>Cinéphile</i> montrant la synthèse de l'entité film	p. 76
Figure 4.2. Interface allemande, interrogation sur un acteur « Berry »	p. 77
Tableau 4-1 Evaluation de la pertinence de la base <i>Cinéphile</i>	p. 78
Figure 4.3. Exemple d'analyse d'un paragraphe au niveau du paragraphe	p. 80
Figure 4.4 Exemple d'analyse d'un paragraphe au niveau de la partie	p. 81
Figure 4.5. Exemple d'analyse d'un paragraphe au niveau du virgule	p. 81
Tableau 4-2 Echantillon de descripteurs multi-niveaux	p. 81
Figure 4.6. Petit forum d'éducation, analysé par ThemAgora zoom sur la hiérarchie des thèmes	p. 86
Figure 4.7. Comparaison d'un forum associé à une tâche (OS Project) et d'un forum associé à un cours (OS Concepts)	p. 87
Figure 4.8. OS Concepts, zoom sur le dernier « moment » de discussion et sur un débat analysé par Agora	p. 88
Figure 6.1. Structuration thématique d'un article de clinique, avec ses tableaux	p. 103

Chapitre 1

Introduction

Les enjeux de l'étude des textes

1. CONTEXTE ET POSITIONNEMENT

En 1978, date à laquelle j'ai soutenu une thèse ancien régime, à l'Institut National des Langues et Civilisations Orientales (INALCO), je m'intéressais au langage spécialisé de l'océanologie. De là, j'ai suivi un parcours qui m'a menée de la traduction à la documentation puis à la recherche, du privé au public et de la linguistique à l'informatique. Mon entrée au Cnrs¹ date de 1982. Mes travaux à l'INALCO, puis au Centre de recherche en linguistique sur l'Asie orientale de l'EHESS², ont porté sur l'étude des textes, principalement la macro syntaxe, sur le japonais et sur la comparaison des langues — la linguistique contrastive. Travailler sur des textes était encore une option à risques dans les années 1980. Quoique les travaux pionniers aient déjà été publiés, ils restaient confidentiels. On tenait alors majoritairement qu'au-delà de la syntaxe de phrase, il n'existe plus la moindre organisation, puisque le locuteur est libre de son expression et n'obéit à aucune règle (cf p. 18). L'étude du discours s'est fort heureusement répandue depuis le tournant du XXI^e siècle, mais reste lieu de controverses.

Outre les références théoriques sur lesquelles je reviendrai, mes travaux sont proches de ceux de contemporains, par exemple Sophie Moirand, qui propose dès 1990 une grammaire du texte et des dialogues, à partir de textes simples et de l'oral, en particulier le discours didactique à l'école (Moirand, 1990). Avec son équipe du Centre d'étude des discours ordinaires et spécialisés, le

¹ Centre national de la recherche scientifique

² Ecole des hautes études en sciences sociales

Cediscor, elle s'oriente vers la vulgarisation scientifique dans la presse (Moirand, 2007). Toujours soucieuse des méthodes, elle s'intéresse à l'énonciation ou s'attache à cerner l'explication ou l'argumentation (Moirand *et al.*, 1993 ; Moirand 2003 ; Doury & Moirand, 2005).

L'approche rhétorique est renouvelée par l'étude des figures de pensée, reliant les figures de style et le raisonnement, en rupture avec l'éclairage esthétique dominant. Par leur orientation multilingue, mes travaux sont également proches de ceux de Kjersti Fløttum menés à la même époque (mais alors publiés en suédois) sur un corpus littéraire puis académique (Fløttum, 2001 ; Fløttum & Rastier, 2003) ; ils s'en distinguent par la modélisation. Dans une autre direction, Fahnestock se penche sur la rhétorique à l'œuvre dans les écrits scientifiques (Fahnestock, 1999).

Des dix années passées à étudier la structure relationnelle des textes scientifiques en japonais et en français, je retiendrai l'enseignement du corpus et l'intérêt de la méthode distributionnelle, points qui seront développés dans les chapitres suivants. Par enseignement du corpus, j'entends l'étude manuelle du matériau texte (voir p. 27). Elle a été menée sur une collection de plusieurs centaines d'articles, académiques et de vulgarisation, collection qui a depuis largement dépassé le millier (sans compter les traitements automatiques). La méthode distributionnelle, préconisée par Bloomfield (1933) et ses successeurs Harris (1951, 1952) et Hockett (1958), met l'accent sur le relevé systématique de la forme et de la position, notées séparément. Cette étude a abouti à une description des articles académiques. Mais la description n'est pas suffisante. Pour démêler ce qui relève de la langue, du genre, du style individuel, deux théories, celle de Jakobson et celle de Yamada, m'ont servi d'appui, dans le cadre de la thèse d'état que je préparais alors³.

La théorie de Yamada est issue de la comparaison du japonais, du chinois et de l'allemand (Yamada, 1908, 1936). Elle est fortement influencée par la psychologie et les travaux allemands sur l'aperception. Il faut signaler que cet auteur a été quasiment mis à l'index après la 2^{ème} guerre mondiale, du fait de sa récupération par, ou sa coopération avec, le régime militaire, et de ce fait n'a pas été traduit en anglais. Au Japon même il sort peu à peu du purgatoire, la célébration du centenaire de sa grammaire a été marquée en 2008 dans sa province d'origine.

La théorie de Jakobson est plus connue. Ce pionnier, animateur du Cercle de Moscou puis de l'école de Prague couvre tous les champs de la linguistique, fonde des domaines nouveaux, la phonologie et la poétique. Il œuvre avec Troubetskoy dans les années 1920 jusqu'à la 2^{ème} guerre mondiale. Il a ensuite émigré en Europe de l'Ouest et aux Etats-Unis. Son œuvre, qui enjambe littérature et linguistique, est considérable et difficile. Elle a marqué profondément le structuralisme. Elle a été traduite et compilée en anglais dans des œuvres choisies en 8 volumes (Jakobson, 1960-1988), et traduite tôt en français par Todorov (Jakobson, 1963, 1973a, 1973b).

³ Cette thèse n'a pas été soutenue, pour plusieurs raisons, dont l'absence de traduction de Yamada pour raisons politiques ; la thèse d'état a disparu en tant qu'exercice académique tandis que je me confrontais au domaine de l'informatique linguistique.

Sans doute Yamada et Jakobson ont en commun d'établir explicitement leur méthode sur plusieurs niveaux d'analyse. Ces deux auteurs ont aussi en commun de présenter les principes de leur méthode, en relation avec leurs résultats. Ce point est essentiel, puisque j'étais davantage préoccupée par les opérations que par les étiquettes et déjà motivée par la transposition des acquis méthodologiques de la linguistique vers l'informatique.

En effet, ma première exposition à l'informatique linguistique n'a pas été celle de la plupart de mes collègues. A l'INALCO en 1982-1985, j'ai été formée au CERTAL⁴ dans la lignée de l'école de Prague, telle qu'illustrée par Sgall en informatique. Cette approche dite structurale était très peu connue alors en dehors du monde slavisant, les ténors écrivant en tchèque à de rares exceptions près (Sgall, 1982 ; Hajicova, 1983). Le troisième Cercle de Prague, comme on l'appelle aujourd'hui, met l'accent sur les opérations (les relations), contrairement à l'école américaine, qui met l'accent sur les opérands (les mots) (Peregrin, 1995).

Si les écrits d'informatique linguistique d'Europe de l'Est⁵ étaient peu accessibles dans les années 80, en revanche, j'étais familière des travaux du premier Cercle de Prague en linguistique (la classique Ecole de Prague). Les travaux de référence de ce qu'on a coutume d'appeler le second Cercle de Prague dans les années 60 ont commencé à être publiés en anglais ou en français dans les *Travaux du cercle linguistique de Prague* (Trnka, 1964 ; Daneš, 1966). Ceux du troisième Cercle de Prague, dominé par les informaticiens linguistes à partir des années 90, ont été largement publiés en anglais par Hajičová et ses collaborateurs (Hajicova *et al.*, 1990 entre autres). Certains travaux anciens ont aussi été traduits en anglais dans les *Travaux du cercle linguistique de Prague nouvelle série* depuis 1995. La diffusion internationale des travaux fondateurs n'a cependant commencé qu'avec les réimpressions par Benjamins, depuis 2006, des publications traduites du tchèque (Sgall, 1990 ; Sgall, 1992 ; Sgall, 1995 entre autres) ainsi que des *Travaux* qui étaient peu accessibles.

Ce contexte de formation décalé par rapport aux préoccupations contemporaines et aux matériaux de référence accessibles explique une position très minoritaire, sinon marginalisée en linguistique informatique. Je la partage avec Jacques Vergne qui a également reçu la formation du CERTAL à la programmation dite « sans dictionnaire », appuyée sur le calcul des *relations* morpho-syntaxiques. Cependant, depuis longtemps les pratiques pragoises se sont écartées de la morpho-syntaxe calculatoire pure et dure (Hajicova *et al.*, 1992). Après avoir exploré nombre de formalismes de calcul, les néo-structuralistes se sont rapprochés du modèle dominant en traitement

⁴ Centre d'étude et de recherche sur le traitement automatique des langues, dirigé par le Professeur Patrice Pognan à l'INALCO.

⁵ pour les plus jeunes, on disait alors « derrière le rideau de fer ». La Tchécoslovaquie a été envahie par l'URSS en 1968. La révolution de velours en 1990 lui a redonné son autonomie. La partition de la République tchèque et de la Slovaquie a eu lieu fin 1992.

automatique des langues, celui de la mémorisation de données avec le *Prague Dependency TreeBank* (Sgall *et al.*, 2003)⁶.

Ces quelques notes sur mon parcours marquent l’empreinte forte du structuralisme pragoïse sur la direction de mon travail : le multilinguisme, d’emblée, l’intérêt pour le discours et le texte, les méthodes informatiques à très forte composante calcul et très faibles ressources mémoire. En même temps, la perspective historique montre qu’entre le Cercle de Prague « classique » plus ou moins porté au rang d’icône aujourd’hui, et la linguistique informatique de Prague plus ou moins idéalisée, il y a un fossé. Enfin, entre le passé reconstruit et la réalité contemporaine il y a bien des chemins qui ne se croisent pas.

Le but de cette présentation n’est pas de justifier d’une ou de ma théorie du texte, mais plutôt de montrer comment j’en suis venue à dissocier la modélisation linguistique de la modélisation informatique. Ces deux activités ne sont pas symétriques et encore moins transitives.

En effet, contrairement à une option admise, je considère, au terme de mes longues années d’apprentissage, au demeurant passionnantes, qu’il y a une rupture irréductible entre ce qui est utile à un apprenant et ce qui est utile en programmation. Les méthodes que j’ai proposées pour l’analyse manuelle des textes font l’objet d’un manuscrit que je projette de rendre accessible à un plus large public sous forme d’ouvrage. L’analyse automatique se fonde sur d’autres principes. Je défends donc une approche du traitement automatique des langues (ou TAL) à l’opposé de la vision anthropomorphe de l’intelligence artificielle (IA), qui insiste sur la simulation, le fait que l’IA a pour but d’avoir toutes les apparences de l’intelligence (humaine ou rationnelle), voire sur le fait que le fonctionnement interne du système IA doit ressembler également à celui de l’être humain ou être rationnel. Au fur et à mesure que l’on découvre l’outil complexe qu’est un ordinateur, l’idée d’un fonctionnement compréhensible en tout point s’efface au profit d’un résultat compréhensible.

J’ai pourtant conservé le terme de « différentielle » par lequel je qualifie ma méthode. Ce terme est utilisé dans l’équipe ISLanD, et porte l’influence de Coursil (1992, 2000). Il a aussi été proposé par Rastier (2001) pour remplacer « structurale », adjectif galvaudé ou mal compris. Il met en avant la tradition saussurienne, selon laquelle, en linguistique, tout est différence. Toutefois, l’orientation macro syntaxique, celle que j’ai choisie, a longtemps été récusée par Rastier et bien d’autres. Ma préoccupation était de définir des relations, et des opérations indépendantes du style personnel et de la langue du texte.

Il me fallait schématiser et systématiser. La notion de schématisation appliquée au texte est exprimée par Grize en ces termes (1990 :73)

Sauf à de rares exceptions comme « un train peut en cacher un autre », une schématisation n’est pas faite que d’un seul énoncé. Elle ne l’est pas non plus d’une simple succession d’énoncés. C’est un

6 voir le nouveau site du Cercle de Prague <http://www.cerclideprague.org>.

système, une structure diront certains, dont les éléments soutiennent entre eux des relations multiples. Ainsi les énoncés sont organisés en configurations de dimensions variables, lesquelles configurations à leur tour se composent pour constituer un tout. Le tout peut être de la taille de *A la recherche du temps perdu* ou « Bains interdits. Eau polluée ».

Je n'avais mesuré à mes débuts l'aspect minoritaire de l'approche multilingue dans le monde qu'à travers les difficultés techniques pour traiter le japonais en France et le français au Japon : c'était avant l'adoption de la norme ISO-IEC 10646 qu'on appelle communément Unicode. En 1993, j'ai rejoint le Laboratoire de Mécanique et d'Informatique pour les Sciences de l'Ingénieur (LIMSI) et proposé une schématisation des textes avec des relations multiples. Cette proposition était trop éloignée de l'approche lexicale et monolingue en vigueur pour être acceptable par ce laboratoire.

En effet, l'approche de la syntaxe en « IA langage »⁷ s'appuyait sur l'héritage américain : l'étiquetage de chaque mot dans les phrases, à partir de dictionnaires et de réseaux sémantiques, et non sur la *projection* de relations grammaticales (Sabah, 1988, 1989). Les relations étaient formulées comme patrons de succession (séquence d'étiquettes), et non pas internalisées comme chronologie d'opérations, une étrangeté venue de l'Est, que je trouvais pour ma part parfaitement raisonnable⁸. Ma modélisation était jugée irréprésentable informatiquement. Dans le chapitre 2, je reviendrai brièvement sur les écueils rencontrés, et je commenterai ici et là dans les chapitres suivants les différences entre la vision descriptive ou « énumérative » dominante en TAL, en France et dans le monde, et la vision « projective » que je défends. La différence majeure est que dans la vision dominante on a d'abord modélisé l'objet (la langue ou le corpus), puis l'utilisateur, tandis que je modélise la méthode.

En 1998 j'ai rejoint le Groupe de recherche en Informatique, Image, Instrumentation de Caen (GREYC). J'ai proposé une orientation multilingue, rendue possible par l'approche de l'analyse syntaxique sans dictionnaire défendue par Jacques Vergne (Vergne, 1989). Techniquement, les choses sont devenues plus simples après l'avènement d'Unicode (Giguet & Lucas, 2002 ; Lucas, 2002). Sans abandonner les aspects théoriques et linguistiques, j'ai travaillé sur des applications en bonne intelligence avec l'équipe I3 rassemblant les recherches sur le langage et l'informatique. Celle-ci ayant été remodelée en trois équipes en 2003, j'ai poursuivi mes travaux dans l'équipe ISLanD (Interaction sémiotique, langues, diagrammes) dirigée par Anne Nicolle puis Jacques Vergne depuis 2007. Cette équipe met l'accent sur le point de vue de l'utilisateur dans l'interaction homme/machine, ainsi que sur la pluridisciplinarité, l'expérimentation, et les aspects

⁷ Intelligence artificielle appliquée au langage naturel

⁸ Cette perception était d'autant plus ancrée chez moi que j'avais séjourné au Japon dans le laboratoire Tanaka à l'Université de technologie de Tokyo, où certains travaux slaves étaient connus et l'approche « informée par les connaissances linguistiques » était bien reçue.

épistémologiques — entre autres les méthodes à ressources restreintes — davantage que sur une spécialisation par la tâche, ou le « corps de métier » modelé sur le savoir faire des entreprises, comme la recherche d'information, le résumé, l'indexation ou la traduction.

Au cours des dix ans passés au GREYC, mon activité principale a été de définir des algorithmes : évaluer des objectifs et des voies d'approche pour les traitements informatiques linguistiques. La grammaire est souvent présentée comme une réécriture à partir des mots et de leurs étiquettes, une vision très chomskienne, qui n'est pas celle de l'école de Prague telle que véhiculée via Sgall et Pognan (Pognan, 1975, 2007 ; Kuboň *et al.*, 2007). Mon activité a pourtant évolué, au fur et à mesure de la formation reçue et de la pratique en informatique. Je me suis écartée de la morpho-syntaxe et appliquée à transformer les règles en opérations, une activité qui est celle du génie logiciel et en particulier de la conception et des schémas d'activité (Morand, 2006). Je parle de définition d'algorithmes, au minimum dans le sens de « mode de résolution de problème » et ordinairement dans le sens de conception d'algorithme informatique. Il m'arrive aussi de programmer des logiciels d'étude⁹, pour tester rapidement des hypothèses.

La présentation qui suit ne forme ni la synthèse ni le catalogue de mes activités scientifiques¹⁰, mais est destinée à éclairer mon engagement depuis une quinzaine d'années, à l'interface de la linguistique et de l'informatique, disciplines entretenant des rapports parfois difficiles. La conviction qui m'a soutenue est que les recherches théoriques sur le discours proposent des notions indispensables pour revoir la façon de procéder habituelle, au bénéfice des utilisateurs. La modélisation en linguistique est néanmoins très différente de la modélisation en informatique, car les processus en jeu n'ont rien à voir (Nicolle, 2003).

En attendant un hypothétique renversement des priorités dans l'enseignement de l'informatique et du TAL (Morand, 2009), une méthodologie plus adaptée au matériau traité et aux usages ne peut que profiter à l'informatique linguistique. J'illustrerai l'application de mes méthodes à la recherche d'information avec par exemple la détection des citations dans les dépêches de presse, l'analyse thématique d'articles ou d'ouvrages ; la caractérisation de collections d'articles, le résumé, le suivi des forums.

2. OBJECTIFS

Mon objectif initial était de montrer la pertinence des notions issues de la théorie linguistique, spécialement de l'analyse de discours et de la linguistique textuelle, dans un milieu où le souci

⁹ selon Anne Nicolle, logiciel destiné à éprouver une méthode, tester une hypothèse, sans prétention à la perfection et sans lourde batterie d'évaluation des résultats (Nicolle, 1996).

¹⁰ voir en annexe la liste de mes publications, qui couvre des champs plus classiques, comme la macro syntaxe ou l'énonciation du côté linguistique et l'approche par tâche du côté informatique. La conception de modèles est le fil rouge et le lien entre ces activités entrant dans des disciplines reconnues.

applicatif est dominant et occulte souvent la méthodologie. J'ai changé sur ce point. J'adopte ici une approche factuelle, orientée par les besoins, car sans un contexte d'usage, il n'y a point de solution.

Avec l'explosion d'Internet, le besoin de constituer des unités de sens à différentes échelles se fait sentir. Les articles de presse forment des flux dont on souhaite connaître la teneur, les articles de recherche forment des collections que l'on souhaite consulter rapidement. Les ouvrages électroniques et les bibliothèques numériques se multiplient. Proposer un ouvrage en réponse à une requête en recherche documentaire ne répond pas à l'attente des utilisateurs, on parle alors d'extraits ou microdocuments (Heinonen, 1998). Depuis quelques années, les congrès et ateliers sur le texte se sont multipliés, avec des spécificités applicatives, les dépêches boursières, les textes médicaux notamment. Les traitements à gros grain sont devenus indispensables. La problématique du résumé de texte reste à l'ordre du jour, mais celle de la synthèse de collections de documents prend de l'importance.

Autre problématique, gérer le multilinguisme de la toile. La diversité linguistique rend les traitements sous-jacents aux moteurs de recherche (indexation) de moins en moins gérables, elle nécessite la traduction des mots-clés dans un nombre croissant de langues, sans parler des sociolectes (par exemple, les nouvelles formes employées en messagerie électronique). Les limites du traitement mot à mot sont bien connues, pourtant, les propositions offrant une alternative au traitement à fondement lexical sont rares.

2.1. Les enjeux économiques

Les travaux présentés ont pour trait commun des méthodes peu coûteuses car elles ne prennent que le texte en entrée, elles sont dites « endogènes » (par opposition avec les méthodes qui s'appuient sur des dictionnaires externes) ou « robustes ». On les appelle aussi « légères » ou « parcimonieuses ». Le coût d'acquisition et de maintenance des ressources externes, principalement les dictionnaires et grammaires, représente un handicap, surtout pour les petites entreprises qui forment le tissu économique local et national.

Les travaux présentés concernent des applications recherchées par des professionnels. La presse, les articles académiques, les forums sont autant d'objets d'étude pour lesquels il est possible de proposer des analyses peu dépendantes de la langue, mais au contraire dépendantes de l'objectif ou la visée de l'utilisateur. Il me semble utile de proposer des outils adaptés au genre¹¹, se déclinant en plusieurs langues, comme alternative aux outils en principe génériques proposés pour une langue.

¹¹ Le genre de texte regroupe des catégories associées à des usages sociaux, ordinairement rapportés à des époques et à des cultures. Les nomenclatures du genre varient selon les auteurs et les besoins, par exemple, le genre journalistique, le genre académique font référence à des situations contemporaines connues (Moirand, 2003), tandis que le genre épistolaire ou le genre épique sont des catégories littéraires davantage caractérisées par le style (voir infra p. 15).

Le traitement de grandes collections d'articles nécessite également de rechercher des moyens d'annotation rapides et fiables, pour filtrer les documents répondant à une requête sans faire appel à la consultation de ressources externes.

Le rapport qualité prix est un avantage des méthodes dites « robustes » ou « parcimonieuses » que nous avons développées, elles sont accessibles à des utilisateurs ne possédant pas de gros moyens informatiques et financiers. Cela oblige à mieux définir les besoins réels des utilisateurs, à abstraire davantage les propriétés du texte et les traitements qui permettront de répondre aux attentes.

2.2. Les enjeux politiques et culturels

Les raisons pour développer des approches « factorisées » du traitement des textes ou des documents multimedia sont particulièrement justifiées en Europe, entité politique multilingue et sensible à l'équilibre entre les diverses expressions linguistiques (site Europa). La factorisation réfère à l'usage d'un programme pour des textes dans diverses langues ayant des traits communs, par exemple du point de vue de la syntaxe ou encore de la source émettrice. Souvent en effet, les informations recherchées sont les mêmes. Le coût de traitement des langues officielles traitées une à une, à partir du lexique, est une très lourde charge. Le problème est démultiplié dans la perspective de la traduction, lorsque chaque langue source analysée individuellement doit être mise en relation avec une langue cible. Les procédures actuelles ne permettent que des couplages deux à deux. Elles défavorisent les langues pour lesquelles des ressources électroniques n'ont pas été constituées de longue date. Si les méthodes statistiques permettent des traitements moins coûteux, cela se fait trop souvent au prix de la qualité. En pratique, l'usage de l'anglais se banalise comme *lingua franca*.

Il est donc tout à fait justifié de développer des méthodes portant sur des ensembles de langues, qui renversent la perspective, en exploitant la similitude du résultat escompté et en relativisant les différences d'expression, ce qui permet une meilleure économie des traitements informatiques. Ces méthodes doivent inclure des tâches de diagnostic non seulement de graphie, de langue, mais aussi de famille de langue et de style. Dans le cas de la traduction, les traitements partiraient d'une langue vers un ensemble d'autres. Les recherches incluent des processus de recherche de *résolution optimale* pour analyser le texte de départ dans le cas de la traduction, et pour proposer des documents répondant à une requête dans le cas de la recherche d'information multilingue.

3. CHOIX ET DEFINITIONS

Modéliser *la méthode* (les opérations, processus, opérateurs...), plutôt que l'objet, est mon souci constant. Ce n'est pas le rôle du linguiste dans la vision dominante, car il empiète sur les prérogatives de l'informaticien responsable de l'architecture du système et de l'implémentation des programmes. Je souligne donc à nouveau l'ouverture d'esprit de mon équipe, qui m'a permis de m'impliquer dans la définition des algorithmes dès 1999.

Je définis quelques termes que j'emploie mais qui sont souvent employés ailleurs dans des sens différents. Certains des termes de cette liste et quelques autres seront discutés plus avant dans le chapitre suivant. J'y reviendrai dans les deux derniers chapitres prospectifs de ce mémoire.

Algorithme : mode de résolution de problème, stratégie adoptée pour résoudre un problème à l'aide d'un ordinateur. J'appelle souvent « algorithme maître » l'approche ou angle d'attaque du problème et « processus » le déroulement des opérations incluant une chronologie. Une « solution technologique » provient par exemple des techniques d'algorithmique du texte, solutions éprouvées pour une sous partie du problème. Ainsi l'algorithme de Boyer-Moore, qui traite de la similarité de chaînes de caractères quelconques, est une technique dans la « reconnaissance des contours de segments de texte », l'idée maîtresse, issue de la méthode distributionnelle, laquelle est traduite informatiquement par le processus itératif dit « d'exploration des quatre coins ». J'ai été influencée par des présentations de mes collègues du GREYC, en TAL mais aussi en algorithmique et en image, il arrive certainement que je rebaptise métaphoriquement certains processus que j'emploie. Il arrive aussi que je les détourne.

Corpus : collection de textes à traiter par ordinateur. Le corpus d'étude est un échantillon réputé représentatif, prélevé et étudié à la main pour proposer un traitement adéquat pour les objectifs, en fonction des constantes et des mesures observées. L'échantillon ajouté pour évaluer l'adéquation du traitement proposé *a posteriori* s'appelle corpus de test.

Genre : ensemble de textes similaires en vertu de la situation de communication, pris en synchronie, dans le sens de Moirand (2003) ou de Marcoccia (2003).

D'où une définition toujours provisoire mais un peu plus précise du genre, qu'on considère comme une représentation socio-cognitive intériorisée que l'on a de la composition et du déroulement d'une classe d'unités discursives, auxquelles on a été « exposé » dans la vie quotidienne, la vie professionnelle et les différents mondes que l'on a traversés, une sorte de patron permettant à chacun de construire, de planifier et d'interpréter les activités verbales ou non verbales à l'intérieur d'une situation de communication, d'un lieu, d'une communauté langagière, d'un monde social, d'une société... (Moirand, 2003)

Décrire les genres dans une perspective socio-cognitive, c'est considérer que les traits définitoires d'un genre sont les suivants (voir Berkenkotter & Huckin 1995) :

- *Les genres sont dynamiques* : les genres sont des rhétoriques dynamiques qui sont mises en œuvre par des acteurs comme des réponses appropriées à des situations récurrentes et qui servent à stabiliser l'expérience individuelle et à lui donner une cohérence et un sens.
- *Les genres sont situés* : notre connaissance des genres est liée à notre participation à des activités communicatives situées. La connaissance d'un genre est, de ce point de vue, une forme de « cognition située ».

- *Les genres sont décomposables en forme et contenu* : la connaissance d'un genre, c'est la capacité à savoir quel contenu est adapté à quelle forme.
- *Les genres sont « dialectiques »* : le genre auquel appartient une activité communicative est à la fois construit par l'activité et le cadre cette activité.
- *Les genres sont des indices de communautés* : L'adoption d'un genre est un indice des normes, de l'idéologie, de l'épistémologie d'une communauté de paroles [...]. (Marcocchia, 2003)

Kanellos (2008) rappelle les visions dominantes du genre et fait du document numérique (DN) un genre à part entière :

Suivant la première [vision dominante], le genre serait le produit d'une analyse logique et sa détermination relèverait d'une activité classificatoire. Suivant la seconde, il serait de la nature d'un prototype et sa détermination procéderait de l'explicitation d'une structure fondée sur la similitude. À ces deux, on proposerait volontiers une vision du genre du DN en termes de projet d'interprétation : le genre serait alors une donnée permettant de déclencher des lectures efficaces, respectant les normes d'une communauté. (Kanellos, 2008)

Le *répertoire des genres* (Orlikowski & Yates, 1994) sert à préciser les intitulés des catégories employées. Dans les applications, ces intitulés sont idéalement définis en accord avec les utilisateurs.

Grain : la notion de granularité est employé depuis longtemps par Prince entre autres. Le grain correspond pour moi à un champ opératoire et un traitement, il compte donc plus d'une fenêtre d'observation, plus d'une unité typographique ou dispositionnelle en entrée, et éventuellement plus d'une unité logique en sortie. Par exemple, le grain phrase dominant en TAL est un grain qui tient compte des phrases (traitées une par une) et des mots, mais aussi de certaines lettres (terminaisons) et tend à rejoindre une analyse logique par *proposition*.

Mise en forme matérielle ou MFM : terme de Virbel (1985) pour définir les unités typographiques et délimitées par la disposition sur la page, perceptibles à l'œil. En machine, représentation intermédiaire fournie par un segmenteur informatisé, prenant en entrée le document brut et donnant en sortie sa hiérarchie et ses séquences (structure logique) comme entrée d'un système d'interprétation.

Morphologie : désigne tout ce qui a rapport à la forme par opposition à la position. Ordinairement observée au niveau de la chaîne de caractères.

Ordre de grandeur : correspond à un grain de traitement, avec des unités de mesures simples (longueur, profondeur) ou complexes (densité, débit, accélération) en rapport les unes avec les autres, ce qui permet des déductions et des calculs. Dans la gestion de l'échelle, il faut s'assurer que les ordres de grandeur restent en relation d'équivalence pour rendre compte d'un phénomène.

Processus : Les processus informatiques impliquent un déroulement dans le temps (Nicolle, 2003 ; 2006).

Référentiel : par analogie avec la physique, arrière-plan méthodologique ; ensemble d'opérations captant des rapports expliquant un phénomène, et ensemble de mesures associées à un ordre de grandeur qui permettent de calculer des valeurs pour des instances de ce phénomène. Un espace cartésien est un référentiel, une théorie ou une *dominante* dans le sens de Jakobson¹² est un référentiel.

Résolution : par analogie avec la résolution numérique et la résolution métrologique. Le détour par la **mise au point** présente la question des ordres de grandeur. Le problème de la mise au point a été posé par l'Ecole de Prague dans les années 60, en rapport avec une vision systémique de la linguistique à différents niveaux (notamment phonétique et morphologie). Différents termes ont été employés « netteté et flou », « centre et périphérie » ou encore « prototype et allomorphes ». Voici ce qu'en dit Daneš en 1966, citant Hockett, 1963.

Mais l'essence de [la relation périphérie centre à travers] toutes ses manifestations [linguistiques à différents niveaux] peut être subsumée par le fait qu'elle infirme une conception commune de la langue, celle qui veut que son organisation soit exprimée par des patrons fixes, symétriques, réguliers, et un système uniforme d'unités (à différents rangs). C'est exactement ce présupposé fallacieux qui a conduit de nombreux linguistes, [...], à des impasses. Ils ont découvert empiriquement au bout du compte et à deux extrêmes opposés, que cette conception aboutit à de fausses solutions : ou bien ils déniaient toute systématisme à la langue, ou bien, au contraire, ils cherchent à s'adapter aux données établies, en les simplifiant au point de les dénaturer, pour les faire entrer coûte que coûte dans les patrons prévus. [...] Ces deux solutions sont également a-scientifiques et on ne peut qu'être d'accord avec Hockett, quand il dit [...] « La plupart des systèmes sont en quelque sorte manipulés pour apparaître nets et symétriques par vertu des procédés semi-magiques de la logistique propre aux analystes. Cette manipulation est [une nécessité] heuristique — elle met en valeur des relations à l'intérieur d'un système, qui seraient restées obscures sinon. Mais les asymétries, même si on les repousse aux confins du système, en font partie intégrante ». (Daneš, 1966, pp. 9-10)

La résolution en métrologie suppose un instrument de mesure ou d'enregistrement : la résolution est le plus petit écart entre deux valeurs, tel que l'appareil en donne une mesure différente. En informatique, on parle de la *résolution* d'une image numérique : elle est donnée en nombre de pixels par unité de longueur (centimètre ou pouce). La résolution d'une image numérique s'exprime en PPI (Pixel Per Inch) la résolution d'impression d'une imprimante se détermine en DPI (Dot Per Inch). Un intéressant résumé de la problématique de la perception et de la manipulation informatique

¹² Pour une glose de la notion de dominante ([1960] 1963 : 209-248) voir Louis Hébert (2006).

d'images montrera sans doute quelle analogie je construis entre la résolution d'image et la résolution de texte (en chaînes de caractères) (Karczmarszuk, 2007).

Style : Je définis le style à la manière de la rhétorique comme moyen d'expression des idées (Fahnestock, 1999). Dans cette acception, il n'y a pas d'opposition entre le fond et la forme, et le style n'est pas d'emblée défini comme collectif ou individuel. Un résumé plus succinct est proposé ici à partir du site *Silva rhetoricae*.

style	<i>elocutio</i>	λεξις
-------	-----------------	-------

One of the five canons of rhetoric. Style concerns the artful expression of ideas. If invention addresses what is to be said; style addresses how this will be said. From a rhetorical perspective style is not incidental, superficial, or supplementary: style names how ideas are embodied in language and customized to communicative contexts. (<http://humanities.byu.edu/rhetoric/silva.htm>)

Texte : le texte entier, l'ouvrage, l'article correspondant à un acte de communication en un temps et un lieu. Correspond à document et non à texte opposé à image, ou texte opposé à phrase.

4. CONTRIBUTIONS

Mes contributions portent sur la façon de relier des connaissances établies de longue date en rhétorique et en linguistique à des connaissances récentes en informatique ; en particulier je pose le problème de modélisation dans le traitement informatique des textes dans l'optique « robuste » et « différentielle ».

4.1. Modéliser le modèle

Dans la modélisation des relations, qu'elles soient syntaxiques ou macro syntaxiques, j'ai développé un travail sur les opérations et non sur les opérandes, ce qui permet de sortir d'un mode d'approche trop rigide, hérité de la pratique américaine descriptive. L'approche dominante considère qu'il est nécessaire de modéliser *l'objet* langue ou idiome, à travers son lexique.

Le type de modélisation que je préconise vise à *traiter* des textes variés par une procédure uniforme, sous certaines conditions : par exemple que l'objectif du traitement soit bien ciblé et que les textes traités soient comparables. Le genre de texte est une des manières de regrouper des textes comparables, par leurs conditions de création et de réception, ou encore le style.

On emploie couramment les termes de *déduction* et d'*induction* en informatique. Suivant une remarque de Bernard Morand, il serait sans doute préférable d'employer le terme de *calcul* lorsqu'il s'agit d'une opération en machine ou d'un algorithme. Ce qui reviendrait à garder *déduction* (ou *induction*) pour l'activité humaine de raisonnement. Le mérite de cette distinction est de pouvoir ensuite poser la question de la relation homme/machine, la machine effectuant la partie calcul de la déduction, à charge pour l'homme de fournir les prémisses et d'interpréter la conclusion. Je n'ai pas toujours respecté cette discipline dans ce mémoire, tant elle demande d'efforts au lecteur. Mais il est

vrai qu'elle permettrait de bien séparer ce qui est prévu par le programmeur, ce qui est fait automatiquement et ce qui est perçu par l'utilisateur *in fine*.

Quoique depuis des années j'aie défendu l'idée de modéliser des relations et des relations de relations, cette idée n'a pas été très audible. Elle a pourtant été défendue et illustrée par Ganascia et son équipe, notamment en apprentissage, en ce qui concerne le retour d'expérience et l'exploitation de notions comme la *récurrence* qui caractérise un *comportement* (Ganascia, 2001). L'idée de calculer sur des calculs ré-apparaît, sous d'autres vocables et dans d'autres domaines, par exemple dans celui de la robotique et des systèmes dits autonomes, sous la problématique de l'adaptation et de l'auto-contrôle (en anglais *self-**). Ailleurs en fouille d'Internet, elle a surgi sous les termes d'adaptation semi-contrôlée, de reconfigurabilité, d'auto-diagnostic ou d'auto-correction du système.

4.2. Des méthodes différentielles

La première différence importante entre les travaux dominants en TAL, qui exploitent des formes — stockées en mémoire — exclusivement ou prioritairement, et mon approche, est que la *position relative* des indices exploités est pour moi indispensable, prioritaire. La forme ne vaut qu'en co-occurrence, comme propriété ou attribut d'un *tagmème* (ce point sera développé pp. 40 à 49). Cela entraîne une représentation de la position, à travers l'usage de tableaux ou de graphes dans nos réalisations. La plupart du temps, la question n'est pas posée en termes de *trouver quelque chose de défini* n'importe où, mais bien de trouver *ce qu'il y a* (indéfini) *en un lieu défini*. Ces méthodes nécessitent de gérer les espaces de recherche, et par conséquent de bien reconnaître la structure typo-dispositionnelle du document (Déjean, 1998b ; Déjean & Meunier, 2007 ; Giguët *et al.*, 2008).

Deuxième différence, les informations sur *l'absence* d'une forme recherchée à une position donnée sont exploitées autant que la présence de cette forme à cette position (voir p. 43, 48 sq. et 56, 75). Autrement dit j'utilise les traces d'une recherche. Les indices morphologiques recherchés sont des indices fiables, ils sont peu nombreux. Les méthodes de déduction sont fortement contextualisées, ce qui permet d'améliorer les résultats. Une marque zéro, dans le métalangage des linguistes, est une marque, à condition bien entendu qu'elle soit « oppositive ». Dans le métalangage des informaticiens, ceci sera exprimé par la notion de « décision sous incertitude ».

La troisième différence est que les objets informatiques manipulés sont des relations, constitutives de structures ou de constructions : en particulier les relations oppositives et complémentaires, et non les formes elles-mêmes. Dans la vision dominante en TAL, les formes sont réputées préexister et avoir des « étiquettes » indépendamment de la visée de l'application. Je considère pour ma part que c'est au contraire *le modèle qui fait voir*, qui permet d'interpréter ce que l'on voit dans un cadre. Les indices exploités sont souvent des *résultats de tests* sur des formes

quelconques, typiquement présence ou absence d'une chaîne répétée, présence ou absence d'une chaîne partiellement semblable à la même position, tous problèmes dits de « comportement » qui sont gérés par l'algorithmique du texte sans qu'il soit besoin de mémoires de formes attendues.

Le modèle relationnel est une grille de lecture, il est donc choisi en fonction de l'objectif des utilisateurs. Je projette un modèle relationnel sur un corpus, le modèle ayant une cohérence interne, fournissant un système d'interprétation. Les réalisations de ce système — les instances — sont diverses et données, à travers les cas traités. Il faut savoir les relativiser : les marques sont inconstantes (on peut même dire arbitraires), mais les relations sont constantes et les comportements plus encore. Je cherche à créer les conditions d'interprétation optimales pour un utilisateur. Autrement dit, le problème est de gérer l'extrapolation fiable de valeur d'une marque à partir d'un contexte qui doit être bien délimité et suffisamment balisé.

4.3. Les méthodes robustes appliquées au texte

Dans la modélisation des relations macro syntaxiques, j'ai développé un travail sur les opérations transverses, dites multigrains, c'est-à-dire sur des espaces de recherche et d'interprétation qui soient raisonnables dans un « ordre de grandeur ». Comme l'a formulé Greimas, « il faut garder le jeu à l'échelle du joueur ». Ainsi, les mesures ne sont pas les mêmes pour des dépêches ou pour des ouvrages, pour des textes ou des collections de textes : on utilise en première instance les unités typo-dispositionnelles, du paragraphe au chapitre ou au volume, pour aller vite (Virbel, 1985). Pour y arriver informatiquement de manière autonome, on a recours à la reconfiguration dynamique ou au paramétrage dynamique des unités de travail, ou recherche de l'ordre de grandeur.

J'ai abordé le style du document dans une perspective ancienne, comme lien entre l'auteur, le public et le sujet traité (site Silva Rhetoricae). Ce qui permet de traiter des textes variés par une procédure uniforme. On passe ainsi à un problème d'adaptation, permettant de projeter des attentes concernant le public de l'analyseur (les utilisateurs) qui ne sont que partiellement connus.

Ce problème est appelé « validation ou vérification des propriétés du système », on peut le résoudre si l'on connaît les caractéristiques de l'analyseur et celles du corpus traité. Une question qui se pose dans les adaptations dynamiques est qu'à un moment le logiciel ne fait plus ce qu'il était censé faire, ou ne fait plus quelque chose que l'utilisateur reconnaît comme porteur de sens. Il faut donc doter ces systèmes de moyens de vérification de cohérence sur leurs propres résultats, une question appelée « capacité réflexive » ou encore « self-contrôle » ou « auto-correction ».

4.4. Le paramétrage du grain et l'auto-contrôle

Dans les travaux présentés, le grain d'analyse est pris en compte, ce qui n'est pas courant. Le paragraphe est de plus en plus souvent représenté comme unité d'analyse du texte. J'ai introduit des unités de plus haut niveau, comme la section ou le chapitre. De plus, plusieurs grains d'analyse sont

pris en compte dans une démarche « descendante » dans le sens du tout vers les parties, du texte vers ses constituants. Par exemple, un article scientifique est divisé en zones, en sections et en paragraphes. Enfin, le grain d'analyse optimal pour une tâche sur une collection de documents est calculé. Par exemple dans le genre journalistique, pour détecter des citations, une unité d'analyse « groupe de phrases » est appropriée en français, et constitue une unité construite pour présenter les résultats.

Les paramètres de variation que sont la longueur et la densité du texte relèvent du style. On peut les envisager du point de vue du style individuel, ou, ce qui est la solution majoritaire, dans le cadre des sous-genres, ou du style collectif. Ainsi une dissertation est plus longue qu'un exposé. On sait aussi qu'un article de sciences humaines est plus long qu'un article de sciences exactes. Pour gérer ces paramètres de variation, il est souhaitable de faire un profilage de textes (Habert *et al.*, 2000) et même un diagnostic (proactif) du texte entrant. Je reviendrai plus loin (pp. 97-98 et 101-102) sur ces termes. La disposition du texte est très révélatrice du sous-genre, la ponctuation également. Ces paramètres sont traités automatiquement sous le terme générique de « stylométrie » proposé par Karlgren (2000). J'ai repris en partie ces calculs mais en les filtrant par des contraintes fortes en fonction de la tâche visée (donc de la grille de lecture) et du corpus traité, comme nous le verrons à partir d'exemples d'applications dans les chapitres 3 et 4.

Le paramétrage du grain d'analyse, c'est-à-dire, du point de vue informatique, le passage en paramètres des grains traités, est une condition nécessaire pour permettre l'analyse robuste de textes longs, de forums ou d'ouvrages : il est important en effet que l'arbre ne cache pas la forêt. On se donne ainsi un degré de liberté plus grand pour traiter de documents plus ou moins volumineux jusqu'à un niveau de détail plus ou moins fin.

4.5. Les traitements bilingues ou multilingues

On peut appeler les traitements endogènes développés dans l'équipe ISLanD « a-lingues », c'est-à-dire indépendants de la langue (Vergne, 2005 ; Lardilleux & Lepage, 2008). Le terme « a-lingue » que j'ai proposé en 2001 a été retenu par Jacques Vergne et ensuite par Yves Lepage, mais il peut induire en erreur, si on le comprend comme « refus de la linguistique ». En réalité, c'est le contraire. Ce sont des traitements qui exploitent les *propriétés* des modèles de linguistes, les contraintes logiques. Généralement les constantes de genre sont plus exploitables en pratique que les traitements fondés sur des descriptions par langue (j'utilise aussi le terme *idiome* préféré par les linguistes pour éviter certaines confusions, entre « langue officielle » et « dialecte » par exemple).

Le choix de travailler sans ressources externes élimine une incertitude sur les incomplétudes du dictionnaire. Il permet de réduire les coûts de traitement. Sur le plan scientifique, ce choix permet de reformuler des hypothèses linguistiques en relation avec le problème posé. Mais il faut ici entendre « hypothèses linguistiques » au sens fort « hypothèses de la linguistique » ou « hypothèse

d'un linguiste ». Les points suivants marquent une nette coupure avec le paradigme dominant, qui oppose les traitements par langue / idiome, dits symboliques, et les traitements statistiques.

a) Ordinairement les travaux de TAL s'appuient sur le lexique, la partie la moins stable et la plus fine de la description des langues (géopolitiques). Toutefois, il est admis sans discussion que les traitements doivent s'appuyer sur des descriptions mémorisées de formes spécifiques, telles les annotations de dictionnaire associant un mot et une catégorie grammaticale (ou plusieurs), ou encore, dans l'analyse de textes, l'association de passages de texte et des valeurs attribuées manuellement par l'annotateur humain à ces passages. Cela suppose un travail coûteux, à refaire chaque fois que l'on change de langue ou de domaine d'application. Les limites de ces méthodes sont connues : le souci d'exhaustivité dans la description s'oppose au souci de généralisation de la procédure d'affectation de valeur à une nouvelle occurrence. Mon modèle définit des *relations*, leur distribution et leur arité et non les termes de la relation.

b) Une méthode qui ne part pas de la description mémorisée du corpus, ni de l'idiome, mais de la théorie linguistique, exploite en fait la cohérence de la théorie choisie, les capacités réductrices d'une théorie, et aussi les capacités limitées de l'interprétant ou de l'utilisateur *in fine* (Morand, 2004). Elle permet d'unifier sous un même métalangage des phénomènes perçus comme similaires. Typiquement, rechercher des citations dans des articles de presse est une tâche qui implique d'établir une relation entre locuteur et citation : les moyens de marquer syntaxiquement une citation sont peu nombreux, et caractérisent des familles de langue. Les variantes tant typographiques que syntaxiques et morphologiques peuvent faire l'objet de classes d'équivalence. Les techniques mobilisées sont plus proches de l'algorithmique du texte ou de la géométrie que des approches n-grammes ou lexicales du type « sacs de mots ».

c) Les outils statistiques ou les métriques utilisés (ex: lois de Zipf) sont des moyens technologiques correspondant à la fois à des observables discrétisables qui peuvent être comptés automatiquement (les unités typo-dispositionnelles) et à des grilles d'interprétation. Ils sont plus ou moins élaborés, par exemple ils tiennent compte des propriétés des *hapax*, de la caractérisation début-milieu-fin des segments selon les principes distributionnels, des propriétés des séries comme l'*incrementum*.

d) La vision de l'écrit comme linéaire ou unidimensionnelle me semble un obstacle à dépasser. Ce qu'il faut créer est au contraire un espace d'interprétation plus riche qu'un espace à une seule dimension, à supposer que l'écriture soit une linéarisation. Pour atteindre un degré de généralisation suffisant pour gérer des textes dans le cadre des traitements multilingues ou multigrains (tenant compte de l'échelle), j'ai développé un travail sur les relations. Ce travail permet d'approcher la notion d'ordre de grandeur dans un calcul.

Concrètement, les relations (thème et rhème, locuteur et citation, harmonisation des participants à une discussion et débat proprement dit, etc.) seront exemplifiées dans les chapitres 3 et 4.

Chapitre 2

Le contexte et les enjeux de l'étude des textes

Ce chapitre a pour but de montrer les profondes différences d'approche du texte entre linguistique et TAL dominant. Il propose un éclairage succinct sur différentes notions, à commencer par celle de discours ou de texte, et quelques repères dans les approches actuelles pour situer mes choix méthodologiques. Il explique peut-être pourquoi ce n'est pas simple de travailler à l'interface de cultures qui se méconnaissent. La linguistique textuelle ou analyse de discours que je préconise traite de différents grains, tandis qu'en TAL l'analyse par déplacement d'une petite fenêtre d'observation reste ancrée sur le mot comme atome et ne dépasse que rarement la phrase comme cadre d'analyse.

1. DES DEFINITIONS CONFLICTUELLES

Le texte ou discours écrit est l'objet de mon travail. Toutefois, les termes de *document*, de *texte* et d'autres encore sont couramment employés et discutés (voir la réflexion de Pédaque, 2006). Le premier mot à définir est celui de corpus. C'est une collection d'objets d'étude, des phrases, des textes ou des documents. En linguistique, le corpus est construit dans une visée, comme échantillon représentatif du phénomène observé. En informatique, c'est plus simplement le matériau empirique qui sert de donnée d'entrée (Habert *et al.*, 1997 ; Rastier, 2002).

1.1. Définitions en sciences humaines

Dans le champ des sciences humaines, de vastes discussions ont opposé les linguistes sur la légitimité du discours comme objet d'étude. Pour mémoire, ces prises de position à propos de la macro-syntaxe ou « grammaire » du texte.

La structure syntaxique d'une phrase constitue un système de verrouillage aussi définitif que la structure morphologique du mot. Sans doute a-t-on quelque latitude d'élargissement interne (procédures d'enchâssement diverses, par nominalisation, subordination, coordination, incise etc.), mais aucune d'enchâssement externe. La frontière de mot comme celle de phrase est infranchissable dans le domaine grammatical, morpho-syntaxique. Tamba-Mecz, *La sémantique*, 1988 p. 119.

Les mots les phrases et les textes sont encore dans les faits l'objet de disciplines distinctes que séparent des frontières académiques plutôt que scientifiques. [...] Si la linguistique restreinte, centrée sur la morphosyntaxe, domine encore, nous entendons prouver le mouvement en marchant, montrer que le texte est irréductible à une suite de phrases ; mieux, qu'il constitue non seulement l'objet empirique, mais l'objet réel de la linguistique. Rastier. *Sens et textualité*, 1989 p. 5.

Depuis les années 1980, d'autres discussions ont opposé les tenants des études textuelles et ceux de l'analyse du discours, avant d'aboutir à un consensus. Le terme de discours est plus générique, il englobe la forme écrite et la forme orale. Le texte est inclus "dans le champ plus vaste de pratiques discursives qui doivent elles-mêmes être pensées dans la diversité des genres qu'elles autorisent et dans leur historicité" (Adam 1999 : 39).

Les définitions mettent l'accent sur des propriétés du discours ou du texte. L'analyse du discours, sous le nom de rhétorique, remonte à l'Antiquité grecque. Toutefois, il reste peu de choses de cette tradition, comme le soulignait Barthes (Barthes, 1966) et à sa suite de nombreux contributeurs du *Dictionnaire d'analyse de discours* dont voici quelques extraits (Charaudeau & Maingueneau, 2002).

discours : Notion qui était déjà en usage dans la philosophie classique où, à la connaissance *discursive*, par enchaînement de raisons, on opposait la connaissance *intuitive*. Sa valeur était alors assez proche de celle du *logos* grec. En linguistique, cette notion, mise en avant par G. Guillaume, a connu un essor fulgurant, avec le déclin du structuralisme et la montée des courants pragmatiques.

« Discours » entre dans une série d'oppositions classiques. En particulier : discours vs phrase discours vs langue, discours vs texte, discours vs énoncé. (Maingueneau, in Charaudeau & Maingueneau, 2002 : 185 sq.)

rhétorique : La rhétorique est la science théorique et appliquée de l'exercice public de la parole, prononcée face à un auditoire dubitatif, en présence d'un contradicteur. Par son discours, l'orateur s'efforce d'imposer ses représentations, ses formulations, et d'orienter une action. La rhétorique a été définie par les théoriciens de l'Antiquité et portée jusqu'à l'époque contemporaine par un paradigme de recherche autonome. [...]

En France, la rhétorique a disparu officiellement du cursus de l'Université républicaine au tournant du siècle dernier (Douay, 1999). La question de la renaissance de la rhétorique est un topos ; l'effacement du mot « rhétorique » est peut-être nécessaire à sa survie dans l'analyse de discours. (Plantin in Charaudeau & Maingueneau, 2002: 505-508).

Cette dernière définition met au premier plan le discours oral et la relation entre un orateur et son public. Les nouveaux moyens de communication tendent à effacer la distinction de medium, par la possibilité d'archivage du son et de l'image animée. Le discours collectif est également favorisé, à travers l'écriture polyphonique, les forums et encyclopédies en ligne comme Wikipedia. Nous mettons en regard une définition du discours et du texte de Rastier (2001) et celle plus large du texte de McKenzie (1991, cité par Bazin, 1996).

discours : ensemble d'usages linguistiques codifiés attaché à un type de pratique sociale. Ex. : discours juridique, médical, religieux. (Rastier, 2001 : 298).

texte : suite linguistique autonome (orale ou écrite) constituant une unité empirique, et produite par un ou plusieurs énonciateurs dans une pratique sociale attestée. Les textes sont l'objet de la linguistique. (Rastier, 2001 : 302).

texte : L'étymologie même du mot « texte » confirme qu'il est nécessaire d'étendre son acception usuelle à d'autres formes que le manuscrit ou l'imprimé. Le mot dérive, bien entendu, du latin « *texere* », qui signifie « tisser » et fait donc référence, non pas à un matériau particulier, mais à un processus de fabrication et à la qualité propre ou à la texture qui résulte de cette technique (...) sous le terme « texte », j'entends inclure toutes les informations verbales, visuelles, orales et numériques, (...) tout ce qui va de l'épigraphie aux techniques les plus avancées de discographie. (Mc Kenzie 1991 cité par Bazin 1996).

J'ai employé d'abord le terme de texte, référant à l'écrit sous sa forme classique, mais beaucoup d'auteurs emploient « texte » pour « extrait de texte » ; j'ai donc recouru au terme de discours, selon l'usage dominant en linguistique, dans le sens de tout cohérent. J'emploie aussi le terme de discours écrit, ce qui peut paraître étrange. Dans le cas des forums, par exemple, le discours collectif est produit et archivé sous forme écrite, mais il est conçu comme document pour l'action plutôt que comme texte en tant qu'achèvement en soi (Zacklad, 2004) ; alors que, dans la rédaction collective d'encyclopédie en ligne, *texte* garde sa connotation de complétude et d'autorité, sinon de pérennité.

1.2. Définitions en informatique

L'usage dominant en informatique est d'employer *document* pour évoquer un tout sémantique, éventuellement multimédia. On parle de document numérique ou de document électronique, et non de texte électronique. Souvent en effet, le mot texte renvoie à des extraits de texte ou à la partie en format texte s'opposant à l'image (ou aux autres formats). Dans le cadre de l'informatique linguistique, un autre terme technique est employé dans les programmes et à l'oral, mais il n'apparaît guère à l'écrit, c'est celui de *fichier*. Ce terme se décline suivant le format en fichier

texte, fichier bimodal, fichier image etc. Il pourrait être défini comme objet technique qui généralement contient un texte ou un document électronique. Il existe bien sûr des documents divisés en plusieurs fichiers, ou des fichiers qui contiennent plusieurs textes. Cela est vrai de tout support, le livre matériel en est le premier exemple. Nous devons tenir compte de cet aspect dans les traitements informatiques, soit pour diviser le fichier texte entrant, soit au contraire pour éviter de prendre une partie pour un tout.

Ces quelques remarques à propos des définitions montrent le besoin de constituer une unité virtuelle en rapport avec le sens, appelée discours. Cette unité abstraite coïncide ordinairement mais non nécessairement avec une unité tangible ou manipulable (un livre, un fichier électronique).

Les définitions acceptées en informatique pour le texte et son traitement sont très marquées par l'approche dominante qui part du dictionnaire et conserve l'orientation de l'intelligence artificielle américaine.

Texte (traitement du). Pour Walter Kintsch et Teun van Dijk comme pour Charles Perfetti, la compréhension en audition comme en lecture s'effectue par un traitement mot à mot, phrase à phrase (et éventuellement paragraphe à paragraphe), qui vise à l'élaboration d'une représentation mentale intégrée dans laquelle les informations littérales sont interprétées à la lumière des connaissances préalables de l'auditeur ou du lecteur. [...] Elle est élaborée à partir d'informations explicites de nature lexicale organisées en phrases, phrases qui sont elles-mêmes agencées en textes aux structures plus ou moins contraignantes (récits, articles scientifiques, modes d'emploi). (Fayol in Houdé *et al.* 1998 : 141).

Texte (algorithmique du). Algorithmique spécialisée en traitement du texte. Les bases techniques concernent la recherche de motifs, la comparaison de mots et l'alignement de séquences. Elles sont utilisées dans les domaines de la recherche documentaire, de l'indexation pour les moteurs de recherche et des logiciels système (édition, traitement et compression du texte). Les méthodes décrites trouvent leurs applications dans les questions de traitement de la langue naturelle, d'analyse des séquences génétiques et de bases de données textuelles. (Crochemore *et al.*, 2001, 4^e de couverture et présentation du cours).

Text parsing

[...] I have discussed two different notions of parsing that appear in the literature on natural language processing. The first, which I call grammar parsing, is the well-defined parsing problem for formal grammars, familiar from both computer science and computational linguistics; the second, which I call text parsing, is the more open-ended problem of parsing unrestricted text in natural language, which I define as follows:

Given a text $T = (x_1, \dots, x_n)$ in language L , derive the correct analysis for every sentence $x_i \in T$. (Nivre, 2006 : 440)

Pourquoi l'unité x est-elle nécessairement la phrase ? Certes le consensus existe, mais il n'est ni étayé ni argumenté.

On trouvera une approche, par bien des aspects originale, mais illustrant bien la coalescence entre texte et mots écrits dans Dutoit (2000).

2. TYPES ET UNITES D'ANALYSE DE TEXTE

2.1. Origines de la diversité

Les origines de la diversité des analyses linguistiques sont imputées d'une part aux observables, c'est-à-dire à ce qu'il est facile d'observer dans une langue donnée ou dans une famille de langues (les langues latines par exemple) ; d'autre part, à des choix historiques ou épistémologiques, des préférences culturelles, qui fondent les grandes familles de pensée linguistique (Swiggers, 1997 ; Auroux, 1989, 1992, 2000).

Les moyens techniques disponibles à une époque changent les conditions d'observation, par exemple la conservation de traces de l'oral sous la forme d'enregistrements des sons a bouleversé la perception de la langue au début du XXe siècle. La représentation électronique de l'écrit et les moyens de calcul associés forment également un terreau technologique favorisant une nouvelle perception (Vergne, 1998, 2001 ; Nicolle, 2002 ; Habert, 2004). Dans le même temps, la perte des repères habituels rend la tâche difficile.

Je m'efforcerai de montrer que les outils informatiques construits par analogie avec le travail humain et nommés de même (« analyseurs » par exemple) se sont considérablement écartés de la « simulation » anthropomorphe. On continue à appeler analyseurs des outils qui sont des transducteurs ou des calculateurs de chemins dans des graphes et qui n'utilisent aucun des repères que le lecteur voit comme significatifs. Le résultat est obtenu par des voies propres au calculateur, et sa qualité n'a plus aucun rapport avec la plausibilité de la simulation. Ainsi, on peut définir suivant les principes de Gelstat un « fond » au lieu de définir une « forme » saillante, si les moyens de capter la forme saillante sont peu fiables.

2.2. Modélisations du discours

Les analyses de discours en sciences humaines sont en grande partie disjointes des traitements informatiques des langues et des textes, les domaines de savoir étant restés cloisonnés. Le métalangage est donc différent, comme on l'a vu avec *discours* et *document*. Il en va de même avec les notions clés de *structure*, de *modèle* et modélisation, de *relations*.

2.2.1. Modélisation en linguistique

Le terme de *théorie* s'applique en linguistique à une vision d'ensemble, à une œuvre, *modèle* renvoie à une modélisation ou schématisation de certains phénomènes, en général dans une langue donnée et avec des exemples illustratifs : on parle ainsi du modèle de la proposition de Vaugelas (repris dans la grammaire scolaire du français).

En sciences humaines, les analyses de discours portent sur des textes au sens d'écrit cohérent. Pour faciliter l'approche, je présente quelques types ou familles de modèles de manière très schématique, avant de revenir sur des exemples plus précis. La première distinction importante pour notre propos est la fenêtre d'observation. Une famille d'auteurs traite du discours en tenant compte de son élasticité, une autre choisit un cadre d'observation fixe, souvent de quelques paragraphes.

La première famille est celle des théoriciens qui s'intéressent au discours comme un tout, et admet l'élasticité, ou si l'on veut la taille variable d'un discours comme propriété. Deux écoles principales peuvent être reconnues, l'école slave, dont Jakobson est le maître incontesté, et l'école américaine avec Harris. L'Europe occidentale est au confluent de ces deux courants (Firth, [1957] 1968 ; Van Dijk, 1972). L'école slave a donné des repères et des méthodes pour aborder le discours, à travers les contes, les épopées ou la poésie. Le premier Cercle de Prague est le foyer de cette recherche dite structurale, parce qu'elle définit des *rapports de forme* ou de structures à travers des contrastes et des continuités. Troubetskoy et Jakobson y élaborent les principes de la phonologie. Un des apports majeurs de Jakobson est la prise en compte de la relation entre auteur et lecteur, dans la théorie dite de l'énonciation (Jakobson, 1960-1988 ; en français 1963, 1973). Dans cette tradition, les opérations (complexes) ont plus d'importance que les opérandes.

L'école américaine, avec Harris et Pike entre autres, définit aussi des principes généraux, mais met l'accent sur des procédures, des outils méthodologiques. Harris a défini le concept de sélection, une position remarquable que peut occuper une forme remarquable (Harris, 1951, 1952). Le concept de tagmème recouvre une unité théorique qui relie une position, et la forme qui occupe cette position (Pike, 1958, 1967).

Cependant, au Japon et bien plus tôt, dès 1908, Yamada avait proposé une analyse de ce type pour le japonais, en précisant qu'il s'agit d'une option méthodologique (Yamada, 1908). Ce qui importe est de comprendre comment les opérations de mise en texte (ou en phrase) sont interprétables et objectivables¹³.

La deuxième famille comprend des successeurs des pionniers, le second Cercle de Prague pour l'école slave, tardivement traduits (Firbas, 1964 ; Vachek, 1964 ; Daneš, 1994), et dans une seconde vague, van Dijk (1972), Adam (1977) influencés par cette école. Halliday (1976) se réclame de Firth, pour l'école européenne; enfin Longacre (1976) se réclame de Pike pour l'école américaine. Ces auteurs ont en général tenté de formaliser les études de discours en leur donnant un cadre, une fenêtre d'observation fixe, généralement petite (de l'ordre du paragraphe). Le souci est de relier les théories du texte à des connaissances préexistantes sur la syntaxe de phrase (Halliday, 1985).

¹³ Ce point, qui fait écho aux mises en garde d'autres grands linguistes, est souvent ignoré par les successeurs qui pensent disposer d'une description d'une langue, laquelle devient rapidement normative.

Dans leurs travaux ultérieurs, à la suite de van Dijk, Longacre et Halliday ont abordé le texte de manière plus globale, en relation avec le genre (van Dijk, 1973, 2005 ; Longacre, 1983 ; Halliday & Martin, 1993). Harris s'attaque à une théorie mathématique de l'information et l'applique au discours scientifique, à l'échelle de l'article (Harris, 1991, 2002). Au Japon, la recherche passe très vite du paragraphe au texte et au genre (Sakuma, 1983, 1989). Le même mouvement s'observe plus tardivement en Europe, avec plusieurs ramifications (Adam, 1999 ; Adam *et al.*, 2004 ; Rastier, 2001).

Les analyses de texte macro syntaxiques sont par conséquent ascendantes en général: elles cherchent à faire la jonction entre des connaissances établies en syntaxe de phrase et l'approche du texte. Elles reposent sur des modèles binaires (deux fonctions) ou ternaires (trois fonctions principales), à l'instar des grammaires de phrases (Bülher, 1934). Les modèles les plus connus sont compositionnels. Ils supposent que chaque unité atomique a un rôle, et que la composition des unités du plus bas niveau suffit à justifier le niveau suivant.

Au contraire, les travaux de Jakobson et ceux de Yamada, auxquels je me réfère, ne supposent pas le principe de composition. Ils mettent en avant les principes de *perception différentielle*, et l'irréductibilité du tout à la somme des parties. Ce qui se voit à une échelle, à un degré de résolution, n'est pas ce que l'on voit quand on accommode à une autre résolution. De manière imagée, une carte de France n'est pas la somme des cartes à grande échelle, celle des communes, par exemple : les hameaux disparaissent, d'autres contours apparaissent.

L'exposition est un mode d'organisation du discours par expansions successives, qui s'interprètent en thème et rhème (une relation constitutive, deux segments fonctionnels principaux), tandis que l'argumentation en est une autre, fondée sur l'agencement des propositions et des jugements. Les propositions sont construites en respectant l'équilibre ou la relative symétrie axiale des constituants (trois segments fonctionnels principaux, sujet, verbe, compléments). Les *propriétés* mathématiques des modèles ont été discutées par exemple par Bühler et Šaumjan sur le plan épistémologique (Bühler, 1934; Benveniste *et al.*, 1966). De même, Jakobson s'intéresse aux visions géométriques de la grammaire. La géométrie et l'arité des relations sont les données de base des algorithmes que j'utilise.

Pour illustrer ces modèles, je propose deux textes courts. Ce sont à dessein des textes quelconques et non des cas d'école. L'exemple 1 s'interprète plus facilement en exposé, avec un thème (le titre) et un rhème (le corps de texte). Dans le corps de texte, le premier paragraphe expose de manière cataphorique ce qui est développé dans le sous-rhème (le reste du texte). Ce modèle est dit asymétrique, car les segments fonctionnels sont de taille très inégale. Souvent même, les unités de mesure typographiques ne sont pas du même ordre. Ce n'est pas le cas ici, dans une dépêche de presse très découpée, où la mesure paragraphe est utilisable.

Exemple 1. FDJ t20 Ricine mesure paragraphe¹⁴

Des traces de ricine dans des flacons gare de Lyon à Paris

PARIS (AFP), le 21-03-2003

Des traces de ricine, un poison violent, ont été relevées lundi dans deux flacons découverts dans une consigne de la gare de Lyon à Paris, selon le ministère de l'Intérieur.

Le 17 mars, "suite à un appel de la SNCF, la police a saisi dans un casier de la consigne de la gare de Lyon deux flacons contenant de la poudre, une bouteille contenant un liquide et deux autres flacons contenant eux aussi un produit liquide", indique le ministère.

"Les analyses effectuées ont permis de constater que les deux derniers flacons contenaient des traces de ricine dans un mélange qui s'est révélé être un poison très toxique", ajoute-t-il, précisant que "les analyses se poursuivent".

En vertu du plan Vigipirate, tous les casiers de consignes doivent être ouverts tous les trois jours en moyenne. A la présence de ces flacons, découverts par des vigiles, la SNCF a fait appel à la police.

Celle-ci a dans un premier temps pensé à du charbon (anthrax), avant que les produits ne soient confiés à un laboratoire spécialisé en Essonne, indique une source proche de l'enquête.

De source judiciaire, on indiquait que ces analyses étaient jeudi "très avancées", mais que des résultats définitifs étaient attendus dans les prochaines heures, tous les contenus des récipients n'ayant pas encore été identifiés.

C'est la première fois qu'est rendue publique en France la découverte d'une trace de ricine. Un tel produit avait déjà été retrouvé à Londres chez des islamistes soupçonnés de préparer des attentats.

L'enquête a été confiée à la section antiterroriste de la Brigade criminelle, chargée de découvrir notamment l'origine et l'usage de ces produits. La section antiterroriste du parquet a été saisie.

Dans l'immédiat, les enquêteurs restaient très prudents et, de source proche de l'enquête, on relevait que "rien n'est exclu".

Les syndicats de police, dont le SNOP, majoritaire chez les officiers de police, ont d'ores et déjà, indique-t-on de source syndicale, demandé au ministère de l'Intérieur, que les policiers soient "dotés d'équipements adéquates" face à de telles situations.

La police antiterroriste britannique avait interpellé plusieurs Nord-Africains les 5 et 7 janvier après la découverte de traces de ricine, dans un mini-laboratoire situé dans un appartement de Wood Green, un quartier du nord de Londres non loin de la mosquée de Finsbury Park, lieu de rendez-vous des islamistes.

Il n'est pas exclu, précise-t-on de source proche de l'enquête, que cette découverte ait un lien avec ces réseaux.

Reste à déterminer quand ce produit aurait pu être déposé dans cette consigne avant la date de la découverte, pourquoi, pour qui. Il faut le démontrer et l'enquête sera sans doute longue et difficile", indique-t-on encore de même dans l'attente des résultats définitifs des analyses.

Un lien avec le déclenchement des hostilités en Irak "n'est pas non plus avéré en l'état des investigations", ajoute-t-on, mais cela est qualifié de "troublant".

La ricine est la toxine végétale la plus toxique. Son étude comme arme biologique remonte à la Première guerre mondiale aux Etats-Unis. Elle a été rendue célèbre par les services secrets bulgares et leur fameux parapluié.

L'exemple 2 ci-dessous s'interprète plus facilement en proposition tripartite. Les phrases sont numérotées en gris. Ce modèle présente une symétrie axiale. Le pivot F4-F5 est à la fois anaphorique et cataphorique.

Exemple 2. FSV4 Anthonome mesure phrase

Ferme la bouche quand tu manges!

1 L'anthonome du cotonnier, parasite du coton, fait du bruit en mangeant. 2 C'est ce qui a permis à Robert Hickling, ingénieur en acoustique de l'université du Mississippi, de mettre au point une nouvelle technique pour détecter l'intrus : les capsules du cotonnier sont plongées dans une boîte remplie d'eau chaude, insonorisée et équipée de micros. 3 On peut ainsi écouter les larves de l'anthonome festoyer et les localiser. 4 Jusqu'ici, les fermiers découpaient la capsule pour en inspecter l'intérieur. 5 Or, les larves à peine écloses sont invisibles à l'œil nu, et cette opération est longue et coûteuse. 6 La petite boîte pour écouter aux portes du coton devrait donc, malgré son prix (850 dollars), trouver des acquéreurs.

¹⁴ Les références des exemples sont données après la bibliographie p.118.

Si l'on se donne une représentation plane du discours, comme dans la figure 1, on peut se représenter le texte comme une structure géométrique. L'exposition, à gauche, correspond peu ou prou au « cadrage », la proposition, à droite, correspond au « centrage » (Péry-Woodley, 2000).

Dans la figure 1, la relation symbolisée par un arc de cercle est « enveloppante », elle est inhabituelle, mais je l'ai élaborée pour exprimer certaines propriétés du discours, la notion de clôture de l'exposition notamment ; elle est généralement mieux reçue en Asie. Pourquoi le schéma de gauche (exposition) est-il constitué en fait de trois constituants successifs représentés par des rectangles, alors que le modèle est dit binaire ? Parce que nous nous intéressons aux opérations. Le texte est subdivisé une première fois en thème et rhème. Il est possible de réitérer l'opération dans le rhème une deuxième fois, et ainsi de suite. Dans la partie droite, il y a deux relations et trois segments fonctionnels constituant une sorte de macro proposition sur le modèle sujet, verbe, complément. Dans des textes plus longs, ces trois segments s'interprètent classiquement comme thèse, antithèse et synthèse.

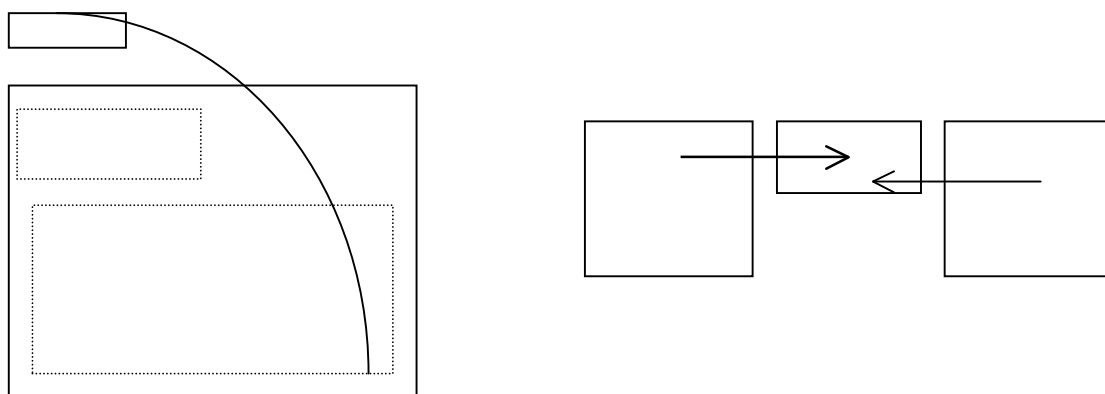


Figure 2.1. Schématisation de l'organisation du discours : exposition à gauche et proposition à droite

En rhétorique classique, le discours est divisé en quatre ou six parties, mais elles sont inégales, et elles sont ramenées à deux ou à trois grands segments fonctionnels, selon les auteurs ou les traditions. Dans une division ternaire, préférée dans la tradition française, le discours se ramène à trois segments, souvent appelés simplement Introduction ; Développement ; Conclusion. Une subdivision plus fine en six parties s'ordonne comme suit : Exorde, proposition ; narration, preuve, réfutation ; péroration. La division n'est pas faite à parts égales et la structure n'est pas fractale.

Existe-t-il d'autres modèles ? La réponse est oui, puisque les linguistes se sont penchés sur une grande variété de textes, sur une grande variété de besoins, et en ont tiré une grande variété d'opérations aboutissant à des structures plus ou moins complexes. Il existe des modèles qui peuvent se représenter en géométrie plane, en cercles concentriques, en ellipses ou en spirale, et ceux qui utilisent une troisième dimension, voire plus, notamment ceux de Jakobson, qui font appel à la notion de profondeur (premier plan et arrière-plan) et aussi de *dominante*. Pour une glose de ces

notions, voir Hébert (2006). Les illustrations suivantes montrent des modélisations graphiques des fonctions du langage de Jakobson, telles qu'elles sont présentées dans *Language & Literature* : 71 (Fig 2) et telles qu'elles sont reprises par Sciuto (Fig. 3) ou expliquées par moi (Fig. 4).

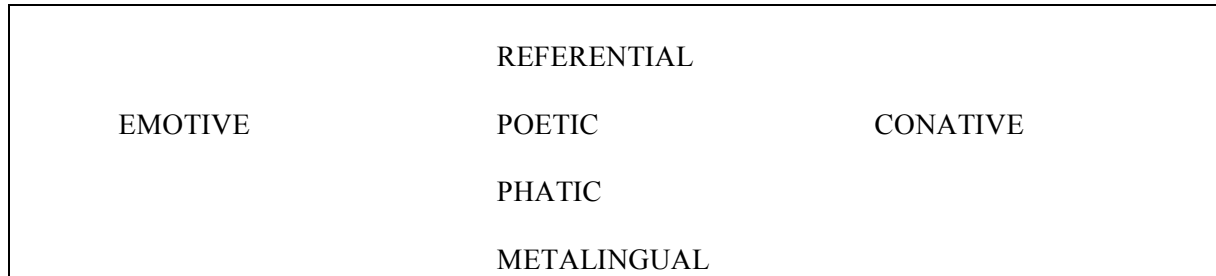


Figure 2.2. Les fonctions du langage selon Jakobson, LL p. 71 disposition sur la page imprimée

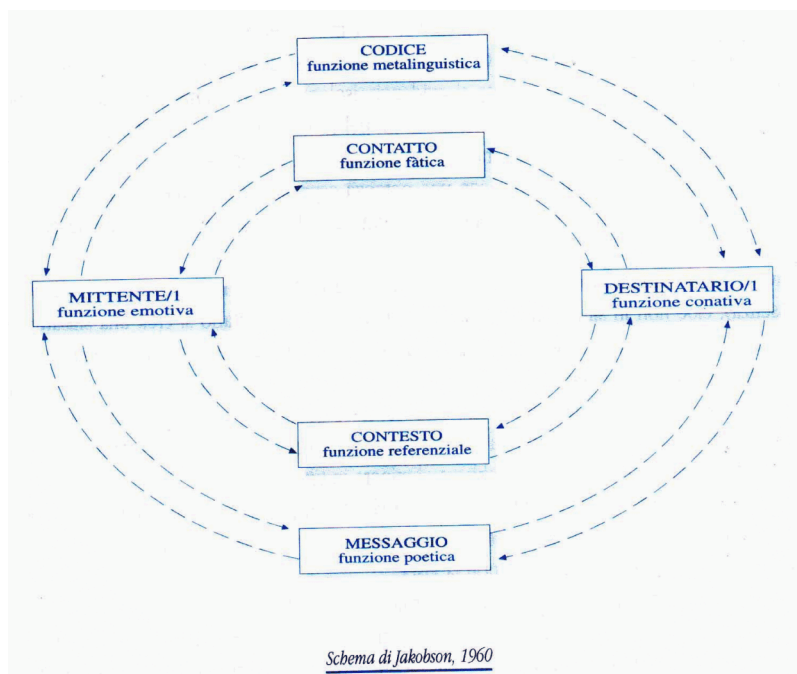


Figure 2.3. Les fonctions du langage selon Jakobson et leur interprétation par Sciuto

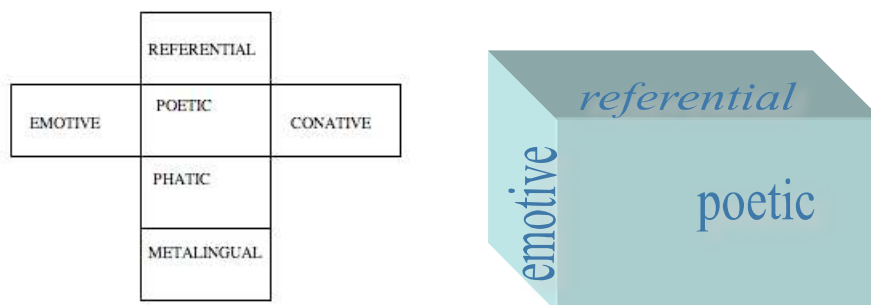


Figure 2.4. Les fonctions du langage selon Jakobson et l'interprétation de la dominante en trois dimensions par Lucas

Un même texte peut-il être représenté par deux modèles ou plus ? Cela arrive souvent, puisqu'un modèle met en valeur certaines propriétés du texte. L'exercice a été proposé maintes fois, ainsi *Les chats* de Baudelaire a été analysé de divers points de vue sur le texte, après l'étude de Jakobson et Levi-Strauss (Delcourt & Geerts, 1999) ou plus didactiquement un texte pour enfants *The Piggott family* a été analysé par les auteurs anglais selon diverses approches, lexicale, co-référentielle, structurelle ou macro-syntaxique (binaire et ternaire) (Hoey, 2001).

Enfin, les modes d'organisation du texte ne sont pas exclusifs, si l'on considère différentes fenêtres d'observation et des enchaînements d'opérations. Des passages de texte relevant de l'argumentation peuvent être reconnus dans un rhème, et un argument peut aussi être organisé localement par une construction en thème et rhème. L'explication est une construction qui relève un peu de l'exposé et un peu de l'argumentation, comme on le verra plus loin. Je pense pour ma part que l'unicité du modèle, réclamé par les pédagogues, mais aussi par les informaticiens, est un obstacle majeur pour l'analyse automatique. Les attentes des utilisateurs sont variées, un seul modèle d'interprétation ne peut les satisfaire.

D'autre part, à ma connaissance et selon les remarques de Jakobson et de l'école de Prague classique, aucun texte ne peut être analysé récursivement à tous les étages avec le même modèle, et avec les mêmes unités (Trnka, 1964) ; à l'exception de quelques constructions littéraires exploitant un procédé unique par jeu, typiquement par l'Oulipo (Kaeser, 1997).

2.2.2. Modélisation en informatique linguistique

Les théories linguistiques sont conçues pour servir de guide à ceux qui s'en servent, jusqu'à récemment des lecteurs, des pédagogues, d'autres chercheurs en sciences humaines. La mécanisation de l'analyse linguistique pose de nouveaux problèmes d'objectivation, de reformulation des besoins et des moyens. La modélisation est ainsi une activité à part entière, qui mobilise la communauté des chercheurs en informatique linguistique.

Toutefois, la modélisation dominante en informatique linguistique part des instances, des mots et des phrases, et non des théories. Il est important de souligner l'importance accordée aux travaux américains, voire leur domination exclusive sur la scène académique. Les deux traits frappants sont la dépendance aux dictionnaires et son corollaire le monolinguisme, ainsi que la dépendance aux grammaires scolaires et son corollaire, l'adoption de fenêtres d'observation très petites. Les auteurs s'intéressant au texte et à l'informatique linguistique (Webber, 1988 ; Kando, 1997 ; Teufel, 1999) finissent invariablement par traiter des relations inter-phrases ou de l'annotation de phrases (Prasad *et al.*, 2006 ; Seki *et al.*, 2006 ; Abdalla & Teufel, 2006).

Le discours ou le texte est une unité naturelle de sens. Le courant de l'intelligence artificielle a affirmé dans les années 1970-80 son ambition de le « comprendre » par des moyens mécaniques. Dans la phraséologie de l'époque, c'est l'ordinateur qui comprend ; ni le programmeur à l'origine

du traitement, ni l'utilisateur qui en bénéficie ne sont situés. De cette ambition, il reste des objectifs toujours d'actualité : la traduction de qualité, le résumé de qualité ou l'indexation personnalisée. Ces objectifs sont désormais reliés à l'utilisateur, en particulier dans les applications.

Dans l'école américaine, les postures épistémologiques caractérisant le paradigme scientifique dominant en traitement automatique des langues (TAL) ont évolué. La croyance que l'ordinateur donne un accès illimité à des connaissances sur le monde a été forte. L'idée que ces connaissances doivent être archivées en mémoire va de pair avec le nominalisme, qui attribue le sens au mot (O'Nuallain, 2002). Suite aux désillusions de la compréhension par l'intelligence artificielle, la tendance est à la description et à l'archivage sur le mode de l'énumération.

Il est intéressant de noter que les modélisations informatiques exigent habituellement des unités directement catégorisées, alors que la reconnaissance d'une structure en linguistique est justifiée par des liens croisés qui définissent les éléments fonctionnels. Les principes de base du traitement classique du langage naturel, tels qu'ils sont présentés par exemple dans (Enjalbert, 2005) sont la *compositionnalité* et la présence de marques (concrètement de mots) permettant d'établir des patrons entendus comme séquences linéaires fixes. On notera aussi le souci de constituer des ensembles homogènes et de taille réduite pour représenter le contenu d'un texte.

La majorité des travaux de TAL s'appuient sur l'exploitation de corpus depuis les années 1990. Si le retour à l'observable est sain, ces travaux « épousent les idiosyncrasies des données sans grand discernement » (Prince, 2008). D'autre part, sémantique reste synonyme de sémantique lexicale. Les analyses automatisées du texte portent sur des sous-problèmes reconnus comme verrous à faire sauter pour donner accès au sens : la reconnaissance d'unités dites thématiques ou de chaînes de co-référence, étudiés dans une fenêtre d'observation de quelques phrases. Les concours organisés entre systèmes restent massivement jugés par des informaticiens. Les compétitions américaines sont orientées par des applications, comme TREC (*Text Retrieval Conference*), DUC (*Document Understanding Conference*) et TAC (*Text Analysis Conference*) qui désormais fédère les deux précédentes. L'évaluation des systèmes consiste à mesurer automatiquement l'adéquation des réponses fournies (automatiquement) à des questions posées par les organisateurs et reformulées à l'aide de mots-clés. L'analyse du texte et le résumé reviendraient à manipuler des chaînes extraites de textes, perçus comme sac de phrases, *in fine* sac de mots.

Les modèles linguistiques connus par les informaticiens américains traitent du texte au sens d'extrait ou passage de texte, de la taille du paragraphe en général. Les plus cités sont la RST *Rhetorical structure theory* dit aussi noyau et satellites (Mann, Matthiessen & Thompson, 1989 ; Mann & Thompson, 1992)¹⁵ ; et le modèle des intentions, dit aussi de focalisation (*focus theory*)

¹⁵ <http://www.sfu.ca/rst> pour le site général et <http://www.sfu.ca/rst/07french/index.html> pour la version française, due à Péry-Woodley et collaborateurs.

(Grosz & Sidner, 1986, 1990). Ce dernier a évolué vers celui du centrage concernant des phrases ou des couples de phrases (*centering theory*) (Grosz & Sidner, 1995).

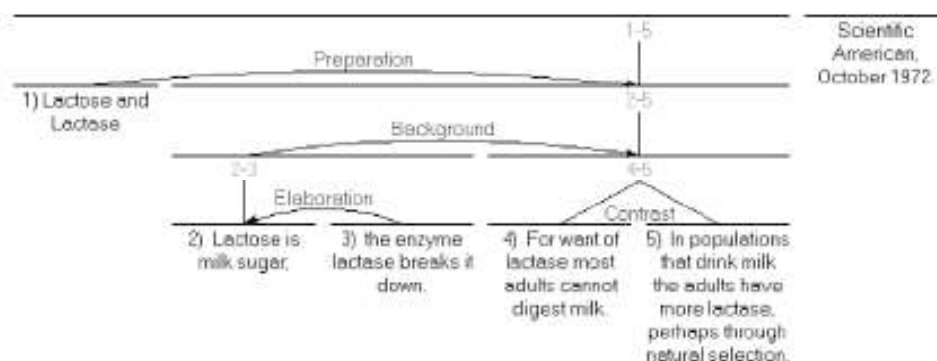


Figure 2.5 Analyse de Mann & Thompson (1992) pour un chapeau d'article

Le modèle de la RST se présente comme l'héritier d'une longue lignée de travaux anglophones sur le discours, notamment Longacre et Halliday (années 80), ainsi que sur la compréhension en TAL (Péry-Woodley, 2000). Les arcs sont annotés dans la figure 5. On comprend bien que traiter d'un chapeau d'article de vulgarisation, d'un paragraphe entier, est à l'époque révolutionnaire, dans le domaine du TAL, qui parvient difficilement à traiter des phrases complexes.

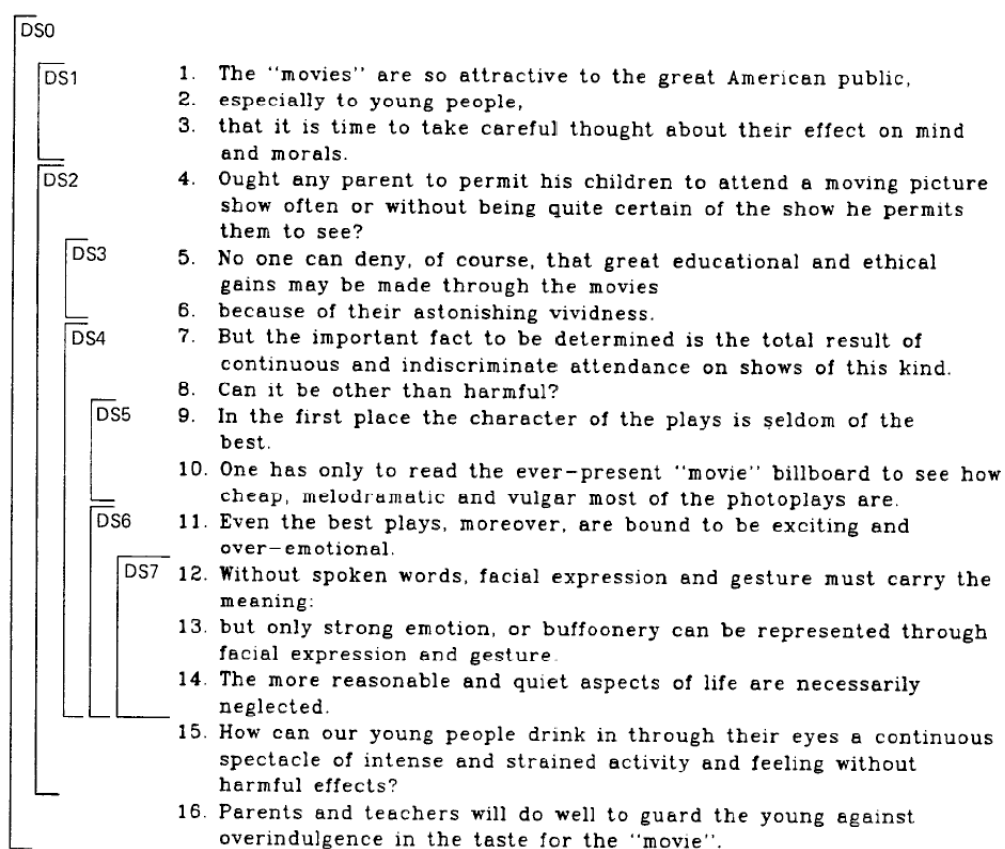


Figure 2.6. Segmentation de Grosz & Sidner (1986) pour un essai court

On remarque dans la figure 6 une propriété recherchée et valorisée par les informaticiens, c'est une hiérarchie strictement inclusive, directement appliquée sur l'instance, sur cinq niveaux de profondeur. Autrement dit, le chevauchement des relations est banni : on montre un résultat « propre » et non des opérations. Le modèle est asymétrique, et récursif. Il ressemble de ce point de vue à celui de l'exposition. Le thème est plus petit que le rhème et seul le rhème est subdivisé.

Les segments sont hétérogènes, en revanche, du point de vue de la mesure typographique. Tantôt des phrases comme en 5, tantôt des espaces ponctués par une virgule comme en 2, ont été découpés. Le titre n'est pas pris en compte. Plusieurs de ces segments regroupés assument une fonction. Les « conclusions » ont un statut particulier, le petit segment 16 est à la fin du niveau DS0 et pourrait apparaître (du point de vue du rapport de forme) comme un thème déplacé « postposé » ; cependant le segment 15 clôt la structure DS2 qui a bien une tête thématique en 4. On a donc bien une recherche de parenthésage et de centrage.

La RST comme le modèle de focalisation sont du type binaire du point de vue opératoire, car les auteurs pensent que c'est le seul qui puisse être formalisé informatiquement. L'analyse de textes et la résolution d'anaphore à partir de la RST a été tentée par la *Veins Theory* (Cristea *et al.*, 1998). L'analyse des textes en vue du résumé a également été implémentée par Marcu suivant le modèle de Mann et Thompson (Marcu, 1997, 2000a). Le traitement de la co-référence a été mis en œuvre d'après Grosz & Sidner (McCoy & Strube, 1999). Toutes ces approches sont de type ascendant (*bottom-up*), des briques (segments marqués) vers la construction, et dirigées par les marques de surface.

Les deux modèles traitant de groupes de phrases, RST et focalisation, ont été intégrés dans un traitement automatique qui propose des formalisations arborescentes concurrentes ou alternatives (Marcu, 2000b). Les phrases sont les unités d'entrée, les propositions sont les unités de traitement reliées entre elles en structures, annotées en sortie. Marcu propose de voir le modèle de la RST comme un jeu de relations entre segments discrétisés adjacents, et le modèle de Grosz et Sidner comme un jeu de relations récursives entre segments discrétisés.

Une réflexion intéressante sur la façon ou plutôt les façons de modéliser en science cognitive a été menée précocement en France (Ganascia, 1998). Plus récemment en Suède, deux stratégies en informatique linguistique, l'une dirigée par le modèle, l'autre dirigée par les données sont explorées pédagogiquement (Nivre, 2006). Dans la stratégie dirigée par le modèle, plusieurs moyens d'informatiser une grammaire (de phrase), en l'occurrence une grammaire de dépendance, à partir des travaux de Prague, a été proposée par Nivre et ses collègues (Nilsson *et al.*, 2006 ; Hall *et al.*, 2007 ; Nivre, 2008). On voit par là que la réflexion épistémologique avance en sciences de l'ingénieur.

2.3. Nouveaux défis

En informatique, les systèmes dits de recherche documentaire sont fondés sur la recherche de mots-clés fréquents, avec le présupposé qu'un terme fréquent représente bien le contenu d'un document. Il s'agit davantage de statistique lexicale sur de très grandes collections de documents que de traitements linguistiques.

Quelques systèmes destinés à l'analyse de textes pris individuellement sont fondés sur des considérations issues de la linguistique (ce qui est souvent traduit par « traitements cognitifs »). Dans ce cas, l'unité de traitement est soit le texte entier, soit le texte fractionné au préalable. On peut citer par exemple les récits (Polanyi, 1985), les articles académiques (Kircz, 1991 ; Kando, 1997). La démarche est alors descendante (*top-down*) ou plutôt divisive du tout vers les parties.

La représentation de données très nombreuses ont un temps suscité des espoirs lorsque les machines vectorielles ont fait leur apparition : si davantage de données sont représentées sur des vecteurs pour représenter le discours, alors on doit avoir davantage de latitude pour le calcul du sens (Lafourcade & Prince, 2001 ; Reitter, 2003). Par ailleurs, les systèmes hybrides entre symbolisme et statistique (Marcu, 2000) ou stylistique-linguistique et statistique (Karlgrén, 2000) ont vu le jour.

Toutefois, les résultats sont assez décevants. Beaucoup d'auteurs constatent que le déluge des calculs n'est pas la réponse au déluge informationnel (Labadié & Prince, 2008a). On constate aussi que la valeur des indices, notamment l'acception des vocables, n'est pas la même suivant le contexte, ou réciproquement, que la taille des échantillons importe (Habert *et al.* 2005 ; Lamprier *et al.*, 2008 ; Labadié & Prince, 2008b). Le calcul par lui-même ne donne pas de sens aux données textuelles, tout dépend de la façon dont les données sont préparées et interprétées (Kanellos & Mauceri, 2008).

3. LES INTERPRETATIONS

Je présente ici succinctement deux courants d'analyse reliés à des objectifs en recherche d'information. Un point doit être précisé. Les auteurs, quand ils sont linguistes, formalisent le discours soit comme une construction à géométrie variable, soit comme une succession de vues obtenues en déplaçant une grille d'analyse — correspondant à une fenêtre d'observation — tout au long du texte. C'est cette dernière vision, linéaire, qui est reprise en informatique. La première vision, celle d'un discours élastique ou « adimensionnel » est minoritaire, mais elle se donne un programme de grande envergure.

Nous retiendrons ici du propos saussurien des principes, qui sont autant de critères de caractérisation épistémologique : prééminence des relations sur les unités, détermination du global sur le local, lien entre description grammaticale et études textuelles, autonomie du langage à l'égard de tout critère référentiel, méthodologie différentielle, synthétisant les pratiques de la linguistique historique et comparée, inscription de la linguistique au sein d'une sémiotique générale. Synthétisant

ces acquis de la linguistique structurale, la sémantique des textes développe aujourd'hui une théorie des formes sémantiques et expressives et s'oriente vers la sémantique de corpus multimedia. (Rastier, 2006 *La structure en question*, p. 1)

Parmi les travaux qui illustrent cette approche et qui sont en relation avec les applications que je traite plus loin, l'étude du discours rapporté à travers les époques et les langues a fait l'objet d'un véritable renouveau (Rosier, 1998a, 1998b, 2009). De très nombreux chercheurs se retrouvent dans cette nouvelle école européenne (Lopez-Muñoz *et al.*, 2004, 2006).

3.1. À la recherche des thèmes du discours

Le maître mot des études de discours, le thème ou topique (*topic*) est un terme dont la définition, même approximative, est fort laborieuse et discutée. Elle a été portée en particulier par le nouveau Cercle de Prague (Hajicova *et al.*, 1992). D'une part, le thème est assimilé à un syntagme nominal, et relié à un champ lexical et conceptuel (c'est une sorte de mot-clé documentaire). D'autre part, le thème est également assimilé à une ouverture, à un élément qui se trouve en tête de discours ou de fragment de discours. L'analogie entre le thème de discours et le thème de phrase (formellement, un syntagme nominal en tête) n'est discutée que rarement (Van Dijk, 1977).

Le modèle de van Dijk tel qu'illustré pour la presse écrite est un modèle de division binaire (van Dijk, 1985, 1986, 1988). Cette division est du type thème et rhème (ou exposition). Elle est exprimée sous forme arborescente et la première partition du texte sous les termes *Summary* et *News story* est inégale. Le modèle permet une analyse récursive avec certaines contraintes. Il est plus sophistiqué qu'il n'y paraît à première vue.

Le schéma suivant représente le modèle abstrait ou macro-structure de Van Dijk pour le genre journalistique.

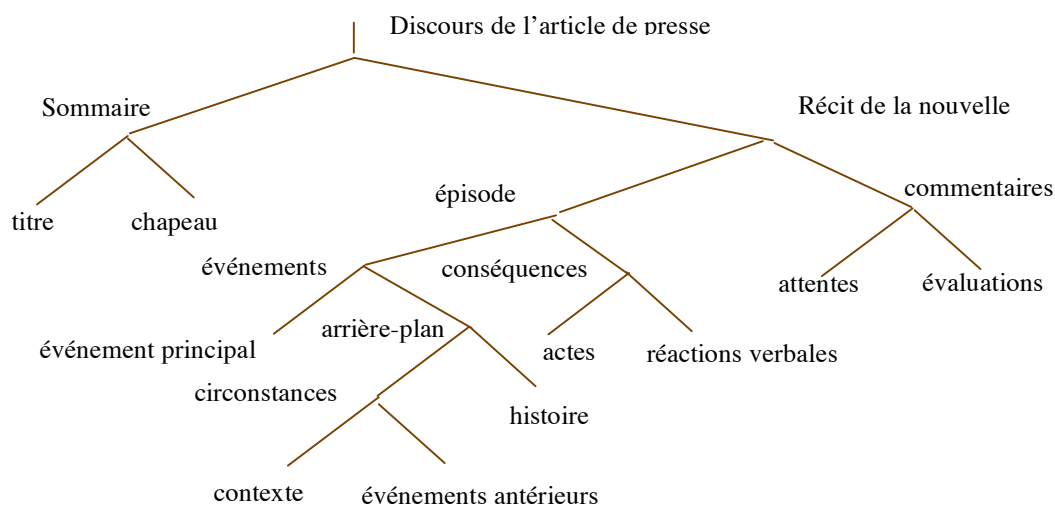


Figure 2.7. La macro-structure de Van Dijk pour le genre journalistique

La vision théorique chez van Dijk est soigneusement dissociée des instances en langues particulières (idiomes). Ainsi les *structures* sont des grilles d'interprétation abstraites, tandis que les *schemata* ou schémas qui les instancient sont dédiés à un idiome (l'anglais, le russe...). Les *macro-structures*, reflétant les catégories sémantiques utiles à l'interprétation dans un genre donné, sont organisées *in fine* par une *super-structure* captant de manière abstraite les schémas de discours, par exemple le schéma narratif, que l'on peut voir à l'œuvre dans la presse mais aussi dans des textes littéraires, donc dans plusieurs genres. Il y a d'autres superstructures, pour le schéma argumentatif par exemple (van Dijk, 1992).

Cette modélisation, concernant l'ensemble du texte journalistique, est élaborée à partir de constructions proposées entre autres par l'Ecole de Prague. Dans le contexte informatique, les « thèmes » d'un texte journalistique n'ont rien à voir avec une construction, ils sont appréhendés par l'intermédiaire de la distribution lexicale. La segmentation du texte est faite par des règles statistiques en fonction de la distribution des termes dans le texte (Hearst, 1994 ; Lamprier *et al.*, 2008).

3.2. À la recherche de l'argumentation

Les approches de l'argumentation sont multiples (Doury & Moirand, 2005). Les modèles théoriques proposés par l'école anglaise et américaine sont fondés sur une analogie de structure entre la proposition et la macro-proposition (modèle ternaire). Van Dijk a mis en avant la structure propositionnelle des textes, notamment dans ses travaux avec Kintsch, qui ont connu un grand succès (Van Dijk & Kintsch, 1983).

Halliday a été l'un des pionniers des études de discours en Europe et celui qui a eu la plus forte influence dans le monde anglophone. Il a remis à l'honneur les relations interphrastiques (anaphore) et a mis en exergue la notion de cohésion dans le paragraphe ou son équivalent à l'oral (Halliday & Hasan, 1976). Halliday s'appuie ensuite à la suite de Firth sur une tradition occidentale d'analyse des textes qui favorise l'argumentation (Firth, 1957, ed. Palmer, 1968). Le système élaboré par Halliday est connu sous le nom de *Systemic Functional Linguistics* ou SFL (Halliday, 1985). La structure macro syntaxique de la SFL s'applique à de petits passages de texte, vus en tant que groupes de propositions. Ces unités s'appuient sur la grammaire traditionnelle et l'étude du lexique (d'où l'appellation *lexicogrammar*). Cette théorie est reprise par Martin (Martin, 1992).

Enfin, Landauer a proposé une métaphore informatique du traitement en mémoire du discours, venant de la psycholinguistique (Foltz *et al.*, 1998). Il s'appuie fortement sur les mots en mémoire immédiate. Un modèle dit *Latent Semantic Analysis* a été formalisé (Dumais *et al.*, 1998). Etant donnée la forte dépendance au lexique du TAL, ce modèle a été bien reçu, repris pour l'oral et au-delà transposé au traitement de l'écrit et de l'image (Asher & Lascarides, 2003 ; Datta *et al.*, 2008).

Les tableaux suivants constituent une tentative, nullement exhaustive, pour schématiser les différentes tendances évoquées et donner quelques mots-clés. Le tableau 1 replace les théories et les modèles auxquels je me réfère parmi d'autres théories linguistiques connues, principalement européennes. Le tableau 2 récapitule les modèles cités en informatique linguistique. On remarque que les théories sont reprises à travers des successeurs et que les théories des anglophones dominent largement. Les noms en gras renvoient aux implémentations.

Tableau 2-1. Quelques courants d'analyse en linguistique

approches	mots-clés	auteurs
théorie générale réflexion sur les modèles	dominante, plans, énonciation, embrayeurs	Jakobson
	superstructure, macro structures, schemata	Van Dijk
	tagmémique	Pike, Longacre
	aperception	Yamada
théorie, modélisation des constructions	énonciation	Jakobson
	énonciation, isotopie	Rastier
	énonciation, polyphonie	Fløttum, Nolke
	exposition	Yamada, Mizutani
	exposition	Jones, Hoey
	argumentation	Longacre
	typologie, séquences	Adam
	SFL systemic functional linguistics	Halliday, Martin
	information theory	Harris

Tableau 2-2. Modèles repris en informatique linguistique

approches	mots-clés	cadres d'observation et auteurs	
		dans fenêtre paragraphe	dans le texte entier
exposition	<i>news discourse</i>	Liddy	Van Dijk
	saillance	Firbas, Hajicova	Trnka
	RST	Mann & Thompson, Marcu	
	focus, intentions	Grosz & Sidner, Marcu	
	cadres de discours	Charolles, Bilhaut	
	exposition	Sakuma, Kando	Kando
argumentation	macro proposition	Van Dijk & Kintsch	Van Dijk
	SFL	Argamon	Halliday ; Martin
			Teufel
énonciation		Teufel	
	LSA	Dumais et al. Asher & Lascarides	
	<i>information theory</i>	Grishman, Friedman	Harris

4. USAGE D'UN MODELE POUR LES APPLICATIONS INFORMATIQUES

Comme indiqué en préambule, j'ai étudié particulièrement deux théories, dont celle de l'exposition de Yamada (cf p. 8, p. 28). Celle-ci a été présentée à travers trois modèles. Le premier, qui a été longuement élaboré dans plusieurs ouvrages entre 1908 et 1936, est appliqué à la phrase (Yamada, 1908, 1936) ; le deuxième est appliqué aux textes littéraires anciens (1913) ; le dernier, à la langue parlée (1922). L'exposition est une façon d'organiser un discours, et d'en rendre compte par expansions successives, comme suggéré dans la figure 2.1 p. 31.

Quel que soit le modèle de référence choisi, il est possible de l'adapter, à la condition d'en connaître les ressorts. Autrement dit, à la condition de séparer le modèle abstrait (référentiel) des modèles particuliers ou *schemata* selon van Dijk, associés à des marques facilement repérables, dans une fenêtre d'observation donnée et dans une langue donnée. Cette adaptation permet de comparer des réalisations différentes dans différentes langues, ou dans différents genres. L'aller retour entre les *schemata* permet de définir les propriétés du modèle, en termes d'opérations ou de relations complexes. Je reviens d'abord sur ces points, avant d'aborder le génie logiciel.

4.1. Transposition de modèle linguistique

4.1.1. Comparaison de langues

Le processus d'assimilation d'une nouvelle langue par analogie de structure a été utilisé souvent en linguistique, historiquement par exemple pour repérer en chinois une structure propositionnelle à l'échelle de la phrase, alors que les marques des langues latines d'origine font défaut (le verbe n'est pas morphologiquement marqué en chinois). Il s'agit donc de projeter une grille d'analyse, un jeu de relations syntaxiques, permettant d'interpréter sémantiquement des réalisations qui ne sont pas marquées régulièrement. On dit que la structure est *reconstructible* (Hagège, 1982).

Le japonais est une langue où le thème est marqué morphologiquement dans le cadre de la phrase, ce qui permet de reconnaître une construction syntaxique à l'échelle habituelle d'observation. La construction thème rhème est binaire, il y a deux fonctions, thème et rhème, et la construction est asymétrique, le thème est plus petit que le rhème. Il est considéré comme mis en facteur par rapport aux constituants du rhème (Garnier, 1982, 2001 ; Lucas, 1991). Ce type de construction a aussi été décrit dans les langues slaves à l'échelle de la phrase et du paragraphe (Hajičová, 1990 propose une modélisation informatique). Si l'on s'intéresse à la structure (telle que les linguistes l'entendent, un système), et non à ses réalisations spécifiques, on peut envisager de la généraliser. Ceci a été tenté en changeant de langue, par exemple en passant du tchèque à l'anglais (Hajičová, 1983 ; Hajičová & al., 1990) ou du japonais à l'anglais et vice-versa (Hinds, 1983 ; Iida, 1997) ou du japonais au français (Lucas, 1993).

4.1.2. Comparaison de fenêtres d'observation

On peut aussi transposer une construction en changeant d'échelle (en passant de la phrase au paragraphe ou au-delà en macro syntaxe). La construction thème rhème est marquée positionnellement dans le cadre de la phrase en anglais, elle est reconnaissable également dans des textes courts en anglais (Jones, 1977). La construction thématique est marquée morphologiquement en japonais dans la phrase (Yamada, 1936 ; Garnier, 2001) mais elle est également observable dans des chapitres d'ouvrages (Teramura, 1990) et dans des textes longs comme les manuels scolaires (Lucas, 2006).

4.1.3. Comparaison de genre et registre

Les différences propres au genre sont connues en informatique sous le vocable de typologie des textes. Les articles journalistiques ou les lettres dans la correspondance amicale sont plus souvent construits en thème et rhème que les romans ou les articles académiques.

Les préférences ou les normes culturelles sont plus fortes dans certaines situations, par exemple, en français, il est déconseillé d'employer les constructions thématiques à l'écrit, à l'échelle de la phrase, alors qu'à l'oral, ces constructions sont fréquentes (Morel, 1994 ; Lacheret & François, 2004). En japonais soutenu, il est conseillé d'employer les constructions thématiques à l'écrit, mais il est permis de les oublier à l'oral. On appelle *registres* les degrés allant du familier au guindé ou formel.

On voit que les paramètres de variation sont nombreux. En français, les constructions thème rhème sont normales à l'oral à l'échelle de la phrase, mais peu prisées à l'écrit. En revanche, cette construction est préférée à l'échelle du texte pour écrire certains articles, par exemple en informatique.

4.2. Transposition de modèle vers un programme

L'analyse de textes par des moyens macro syntaxiques doit être exprimée de manière compréhensible par des informaticiens pour être automatisée. Il faut donc définir des procédures en termes de génie logiciel (Morand, 2006). Les points d'appui des prises de décision dans un système informatique sont très variables, selon que l'on considère le processus d'ensemble, des sous-procédures, ou des données fixes (par exemple des valeurs sous forme d'étiquettes stockées en mémoire). En informatique comme ailleurs, l'abstraction permet de se dégager de la description des instances (Ganascia, [1993] 2007).

4.2.1. Qu'est-ce que fait un programme ?

On est passé de la vision statique univoque (bases de données) à une vision fonctionnelle (bases de données relationnelles). Cependant, il s'agit toujours dans ce cas d'un stockage mémoire que l'on utilise par interrogation et dont l'évolution se fait par mise à jour (donc avec perte de l'état

antérieur). Une des questions qui revient souvent dans les discussions d'ISLand est de trouver une notation qui aide à garder la trace des calculs et des changements d'affectation d'une variable au cours d'un processus, au lieu de perdre toute mémoire de ces transformations : l'informatique semble reproduire le « calcul sur la poussière » pratique préalable aux notations mathématiques algébriques (Herreman, 2001).

La notion de rôle a eu une grande importance, un objet X pouvant jouer un rôle R dans une situation ou un contexte spécifié (Nicolle, 2006). Cette notion est transposée dans les langages orientés objet.

En génie logiciel, les schémas UML proposent des types de liens, pour lesquels on peut stipuler le nombre d'occurrences *du lien* (1 à *n*). La formalisation des liens en contexte permet de se détacher de la description des objets eux-mêmes ou des classes d'objets qui peuvent servir soit une fois seulement, soit plusieurs fois. Pourtant, bien souvent, le génie logiciel est ignoré en TAL, comme si la description était une réponse en soi.

4.2.2. Renversement de perspective

Un modèle syntaxique ou macro syntaxique peut être représenté par des relations, ou des liens, obligatoires ou optionnels. Cependant, cette représentation n'est facile que pour les cas d'école, autrement dit, il est facile de transposer informatiquement une construction canonique, et cela a déjà été fait (*inter alia* Hausser, 2005). Comme le fait remarquer entre autres Marcu, le problème pratique en analyse robuste est de décider, non pas tant ce que l'on souhaite voir (c'est la fonction du modèle de référence), mais surtout comment relier un observable à un modèle de référence, ou comment réduire un texte quelconque au schéma idéal retenu à des fins d'interprétation.

La plupart des auteurs choisissent de modéliser un corpus (l'objet à analyser) en prévoyant des cas canoniques et des variantes pour un maximum de configurations attestées, constituant ainsi des listes de cas mémorisés. On peut aussi modéliser des opérations d'interprétation (la méthode d'analyse). Dans ce cas, la modélisation informatique concerne le méta-modèle ou la méta-heuristique (Ganascia, 1985, 1991, 1993, 2008 ; Nivre, 2008).

De plus, il me semble important de renverser la perspective par rapport à une façon de penser qui admet que l'ordinateur simule l'humain, « apprend », « comprend », remplace le linguiste ou remplace le lecteur. L'ordinateur est un artefact et il joue un rôle particulier, en ce qu'il est doté de la capacité à générer un cadre de lecture, plus exactement à calculer si un texte inconnu relève ou non d'une analyse à l'aide d'un modèle connu.

Sont représentés uniquement les processus de décision permettant de diagnostiquer si le texte entrant peut être ramené ou non au modèle linguistique de référence. Cela suppose que l'on connaisse les propriétés relationnelles du modèle projeté et ses limites, et que l'on soit capable de rejeter un texte « inadéquat » — non pas « mal formé », ce qui supposerait un jugement dépréciatif.

Un texte qui ne se prête pas à l'analyse attendue n'intéressera pas l'utilisateur, souvent parce qu'il s'agit d'un intrus dans une collection, par exemple une dépêche publicitaire dans un corpus de dépêches d'actualité.

Le choix que j'ai fait est de modéliser des opérations d'interprétation (la méthode d'analyse). Il s'agit d'un méta-modèle qui représente les *opérations* sur les relations constituant une structure complexe. Il exploite les *contraintes* inhérentes à un monde clos, celui du linguiste théoricien. J'ai donc tenté d'abord de passer par des représentations schématiques de relations ou de modèles mentaux des relations. J'appelle ce monde clos le référentiel. La projection des opérations sur un espace plan donne une géométrie des constructions esquissée plus haut (Fig. 1 p. 31).

La figure 8 représente trois géométries, chacune munie d'un jeu de relations constitutives d'une construction canonique : A l'exposition, B l'argumentation et C l'explication. Les rectangles représentent des *segments fonctionnels* (et non des supports physiques comme des pages). Les ajouts possibles sont en pointillé. Les marques sont organisées comme un parcours et les rectangles sont disposés suivant une géométrie : il s'agit donc d'un *espace mental*, organisé par des relations, et non d'une description du texte. Dans la figure 8, les cercles représentent des points de vérification théoriques pour les relations stipulées, dans un idiome donné. Il s'agit donc de schémas ou *schemata* de van Dijk. Dans une langue, un chemin est tracé par marquage régulier. En effet, l'économie perceptive entraîne l'effacement des marques redondantes. De plus, l'économie de la langue fait que l'absence de marque vaut pour une marque « en creux » dans un contexte donné. Les marques attendues (tagmèmes) sont en noir et les optionnelles en gris.

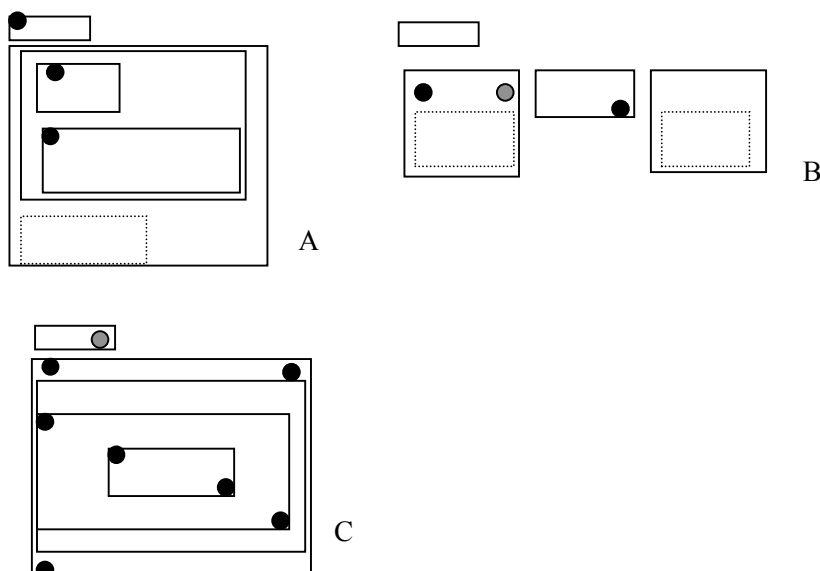


Figure 2.8. Schemata des constructions canoniques : marques obligatoires en noir, optionnelles en gris

Reprenons nos premiers exemples Ricine, Anthonome et ajoutons un autre texte, Afar pour voir à quoi correspondent les endroits marqués. Il ne faut pas s'imaginer que le schéma décrit des textes

physiques ou des exemples attestés. Nous sommes ici dans la projection d'un modèle opératoire. Il est donc question de remonter d'un résultat (le texte entrant) à un *processus analytique* et à un modèle d'interprétation ou grille de lecture. Les boîtes encadrent des segments de MFM, qu'il s'agit d'explorer pour vérifier l'adéquation du modèle d'exposition à ce texte, la compatibilité du texte avec les propriétés de ce modèle.

Le lecteur s'attend sans doute à une explication du même type que ce qui serait fait devant des étudiants, c'est-à-dire la démonstration de l'intérêt des indices retenus, s'appuyant sur le sens. Pour illustrer mon propos, humainement, il est intéressant de se fier à des indices comme *Reste à déterminer* au début du §13, pour caractériser le début de la fin de l'exposé. La forme est remarquable mais peu fréquente. Elle ne convient donc pas à une modélisation informatique. De tels indices ne sont pas « obligatoires » et de plus, devant une langue inconnue, l'analyse échouerait.

Exemple 3 (1 repris). FDV Ricine mesure paragraphe, sélection phrase

Des traces de ricine dans des flacons gare de Lyon à Paris

PARIS (AFP), le 21-03-2003

§1 Des traces de ricine, un poison violent, ont été relevées lundi dans deux flacons découverts dans une consigne de la gare de Lyon à Paris, selon le ministère de l'Intérieur.

§2 Le 17 mars, "suite à un appel de la SNCF, la police a saisi dans un casier de la consigne de la gare de Lyon deux flacons contenant de la poudre, une bouteille contenant un liquide et deux autres flacons contenant eux aussi un produit liquide", indique le ministère.

§3 "Les analyses effectuées ont permis de constater que les deux derniers flacons contenaient des traces de ricine dans un mélange qui s'est révélé être un poison très toxique", ajoute-t-il, précisant que "les analyses se poursuivent".

§4 En vertu du plan Vigipirate, tous les casiers de consignes doivent être ouverts tous les trois jours en moyenne. A la présence de ces flacons, découverts par des vigiles, la SNCF a fait appel à la police.

§5 Celle-ci a dans un premier temps pensé à du charbon (anthrax), avant que les produits ne soient confiés à un laboratoire spécialisé en Essonne, indique une source proche de l'enquête.

§6 De source judiciaire, on indiquait que ces analyses étaient jeudi "très avancées", mais que des résultats définitifs étaient attendus dans les prochaines heures, tous les contenus des récipients n'ayant pas encore été identifiés.

§7 C'est la première fois qu'est rendue publique en France la découverte d'une trace de ricine. Un tel produit avait déjà été retrouvé à Londres chez des islamistes soupçonnés de préparer des attentats.

§8 L'enquête a été confiée à la section antiterroriste de la Brigade criminelle, chargée de découvrir notamment l'origine et l'usage de ces produits. La section antiterroriste du parquet a été saisie.

§9 Dans l'immédiat, les enquêteurs restaient très prudents et, de source proche de l'enquête, on relevait que "rien n'est exclu".

§10 Les syndicats de police, dont le SNOP, majoritaire chez les officiers de police, ont d'ores et déjà, indique-t-on de source syndicale, demandé au ministère de l'Intérieur, que les policiers soient "dotés d'équipements adéquates" face à de telles situations.

§11 La police antiterroriste britannique avait interpellé plusieurs Nord-Africains les 5 et 7 janvier après la découverte de traces de ricine, dans un mini-laboratoire situé dans un appartement de Wood Green, un quartier du nord de Londres non loin de la mosquée de Finsbury Park, lieu de rendez-vous des islamistes.

§12 Il n'est pas exclu, précise-t-on de source proche de l'enquête, que cette découverte ait un lien avec ces réseaux.

§13 Reste à déterminer quand ce produit aurait pu être déposé dans cette consigne avant la date de la découverte, pourquoi, pour qui. Il faut le démontrer et l'enquête sera sans doute longue et difficile", indique-t-on encore de même dans l'attente des résultats définitifs des analyses.

§14 Un lien avec le déclenchement des hostilités en Irak "n'est pas non plus avéré en l'état des investigations", ajoute-t-on, mais cela est qualifié de "troublant".

§15 La ricine est la toxine végétale la plus toxique. Son étude comme arme biologique remonte à la Première guerre mondiale aux Etats-Unis. Elle a été rendue célèbre par les services secrets bulgares et leur fameux parapluie.

Dans la perspective d'un traitement automatique, le traitement ne part pas du sens. En effet, en informatique, les méta-modèles sont mathématiques ou géométriques. Il existe ainsi des outils qui font ce qui est nécessaire, ou tendent vers le but recherché mais qui restent inconnus de l'utilisateur potentiel linguiste.

Le processus informatique est selon moi inverse, par rapport à la mise au point d'un modèle macro syntaxique, puisque le traitement *exploite* un modèle et ses propriétés pour générer un positionnement du texte dans une grille de lecture, si toutefois il y rentre sans forcer (voir Fig. 3.11 p. 68). Ceci est très différent de l'étiquetage réputé consensuel de tous les constituants.

Le texte appartient au sous-genre journalistique des dépêches, le grain paragraphe est un grain fin, qui ne contient qu'une ou deux phrases : alors, il est impossible de supposer qu'un paragraphe contienne une construction. En revanche, les exceptions à la règle qui donne que 1§ = 1 phrase sont des « intrus » dans cette série particulière¹⁶ et probablement des bornes de construction, ainsi le §8 et le §13 qui contiennent deux phrases et le §15 qui en contient 3. Toutefois, du point de vue du traitement, il n'est pas opportun de passer chronologiquement d'un paragraphe au suivant, puisque le modèle est exprimé à travers une topologie. On va donc tester les « quatre coins » imaginaires du modèle.

L'algorithme va chercher à vérifier un certain nombre de propriétés du modèle, l'asymétrie du rapport de forme, à travers la distribution des effectifs de constituants de MFM. Il y a une seule solution possible pour la distribution stylistique observée, autrement dit une analyse possible en vertu du modèle de l'exposition.

Pour une analyse macro syntaxique rapide, permettant de produire une hiérarchie, on teste une relation de co-variance (non répétition) et une relation d'accord lexical (répétition approximative d'une chaîne de caractères). Ces caractéristiques dites morphologiques sont détectées sur des chaînes de caractères quelconques par des algorithmes classiques (Aho & Corasick, 1975 ; Boyer & Moore, 1977 ; Karp & Rabin, 1987). Les relations issues de critères hétérogènes sont recouvrantes et servent de contraintes à satisfaire : on combine en effet la mise en forme matérielle, la typographie, la position des tagmèmes par rapport à la ponctuation, les répétitions dans les sélections de début et fin pour caractériser *des comportements discursifs* (et non des segments perçus comme statiques et dotés de propriétés intrinsèques).

Concrètement on s'intéresse à la différence ou contraste entre le titre du point de vue MFM et le début et la fin du corps de texte. Certains éléments du lexique en revanche sont répétés ; ce qui veut dire que la recherche dans le titre d'une sous-chaîne de la chaîne de caractères du §1 donne un résultat positif (je souligne à nouveau que la forme en elle-même est indifférente mais que le

¹⁶ C'est habituellement le contraire, les paragraphes courts sont des paragraphes de transition. On utilise, non pas une règle fondée sur la fréquence ou la plausibilité : § court → transition, mais l'idée de Gestalt. On teste l'hypothèse : si « forme » différente du « fond » → transition ?.

résultat du test sur la répétition est important, parce qu'il vaut pour *isomorphisme*). La recherche dans le dernier § d'une sous-chaîne de la chaîne de caractères du §1 donne un résultat positif. On cherche alors le facteur commun entre le début et la fin du corps de texte, ici le trait présence de guillemets en position seconde et antépénultième, à l'exclusion de la position première et pénultième (un patron en poupée russe).

Le calcul, qui ne sera pas déroulé dans le détail, parce qu'il n'a pas de sens pour le lecteur, aboutit immédiatement à un résultat de segmentation, équivalent à l'attendu manuel, mais sans s'appuyer sur des hapax. Il permet de colorier des segments, que le lecteur interprète.

Lorsque l'on projette un modèle, on ne cherche pas à énumérer tous les cas possibles et tous les accords possibles, comme dans l'approche descriptive. On ne cherche pas non plus à caractériser positivement tous les segments, un par un. On cherche à subdiviser un tout réputé cohérent et à valider, tant que rien ne s'y oppose, que l'on reconnaît une construction. C'est la règle du *nihil obstat* (en latin). Dans l'exemple 4 sur l'anthonome, il y a empêchement à interpréter le texte comme un exposé. En effet, le titre n'est pas cataphorique.

On reconnaît en revanche les contours de trois segments, en projetant une autre grille d'analyse. En procédant de la même manière que précédemment sur le texte, découpé en phrases, on délimite des segments, qui ne sont pas basés sur le sens, mais sur la syntaxe. Deux segments sont consécutifs et accordés.

Exemple 4 (2 repris). Argumentation FDSV4 Anthonome mesure phrase, sélection virgule

Ferme la bouche quand tu manges!

F1 L'anthonome du cotonnier, parasite du coton, fait du bruit en mangeant.

F2 C'est ce qui a permis à Robert Hickling, ingénieur en acoustique de l'université du Mississippi, de mettre au point une nouvelle technique pour détecter l'intrus : les capsules du cotonnier sont plongées dans une boîte remplie d'eau chaude, insonorisée et équipée de micros.

F3 On peut ainsi écouter les larves de l'anthonome festoyer et les localiser.

F4 Jusqu'ici, les fermiers découpaient la capsule pour en inspecter l'intérieur.

F5 Or, les larves à peine écloses sont invisibles à l'œil nu, et cette opération est longue et coûteuse.

F6 La petite boîte pour écouter aux portes du coton devrait donc, malgré son prix (850 dollars), trouver des acquéreurs.

Ce qui est intéressant du point de vue de la logique de la construction est de délimiter l'antithèse, en 4 et 5 mais ni le début de la thèse, ni le début de l'antithèse ne sont marqués de façon patente. L'antithèse ne porte pas de critère morphologique simple à repérer comme la négation verbale. En revanche, la fin du premier et la fin du deuxième segment sont marquées. Insistons encore sur la transposition des marques exploitables. Humainement, on repère des segments rétroflexes. En 3, le lien pronominal et anaphorique vers 2 *On peut ainsi* est repérable, de même que la coordination disjonctive *Or*, en 5. On forme humainement ainsi des ensembles bornés 1-3 et 4-5.

Automatiquement, il est préférable de rechercher des frontières : 3 a un caractère particulier des frontières, c'est un segment non subdivisé. Ensuite on vérifie un accord (en l'occurrence en

coordination) : la présence de *et* dans la fin des deux phrases 2 et 3 ainsi que 5. On interprète le rôle de ces segments positionnellement en vertu du modèle, comme la thèse précédant l'antithèse. La seule solution pour 6 est synthèse. On y retrouve en outre des chaînes de caractères du premier segment comme *boîte* et *écouter*, qui pour l'homme sont des indices de cohésion lexicale. On peut aussi les retrouver par l'algorithmique du texte. On n'a pas utilisé de marques de début, alors que le schéma B de la Figure 5 le faisait attendre. En effet, on n'est pas ici dans le cas prototype. Mais la structuration n'échoue pas pour autant, parce qu'il est équivalent de délimiter des débuts ou des fins.

Exemple 5. Explication FSV 12 Afar

Le point chaud de l'Afar sous surveillance

F1 Près de 90% des volcans naissent en bordure des plaques tectoniques, au niveau des dorsales et des plaques de subduction. F2 Mais il existe un deuxième type de volcanisme, beaucoup moins répandu, dont l'origine ne semble pas être liée aux mouvements tectoniques : le volcanisme de point chaud. F3 « *Certains volcans apparaissent au milieu des plaques lithosphériques et résultent de la remontée rapide de matière chaude provenant des profondeurs du manteau*, explique Jean-Paul Montagner, directeur du Département de sismologie de l'Institut de physique du globe de Paris (IPGP). F4 *Ces panaches mantelliques percent la croûte terrestre et à mesure du défilement des plaques au-dessus du point chaud, se forment des chapelets d'îles volcaniques parfaitement alignées (Hawaï, La Réunion...)* ». F5 Mais comment et à quelle profondeur naissent-ils? F6 Parviennent-ils tous en surface? F7 Quelle est leur structure intime? F8 Pour répondre à ces questions, un programme d'étude géophysique coordonné par Michel Cara, directeur de l'École et observatoire des sciences de la terre (Éost) de Strasbourg a été mis en place. F9 Deux équipes de l'IPGP et de l'Éost se sont ainsi rendues au Yémen et en Ethiopie, régions où se trouve l'un des rares points chauds émergés. F10 Organisée dans le cadre du programme « Corne de l'Afrique » de l'Insu, leur mission avait pour but de densifier le réseau de sismomètres large bande afin « d'échographier » le globe en profondeur. F11 « *Au lieu d'utiliser les ultrasons, nous nous servons des ondes sismiques pour imager les points chauds*, explique Jean-Paul Montagner. F12 *Ces ondes se propagent plus lentement dans les milieux chauds.* F13 *En repérant les anomalies de vitesse, nous pouvons ainsi cartographier les panaches mantelliques en 3 dimensions.* » F14 Pendant une semaine, les chercheurs parisiens ont sillonné le Yémen à la recherche de zones épargnées par le « bruit culturel » (les vibrations produites par l'activité humaine). F15 C'est finalement au nord d'Aden qu'une nouvelle station a été mise en place, venant enrichir le dispositif de surveillance déjà installé dans l'année écoulée — une station au Yémen, et trois sur la rive éthiopienne de la Mer Rouge. F16 « *Nous attendons à présent que les données s'accumulent*, explique le chercheur. F17 *Fin 2001, nous devrions être en mesure de fournir une image détaillée du sous-sol de la corne africaine.* »

Ce texte relève de l'explication partielle. En réalité, l'original est un document illustré par une carte. Le commentaire de la carte présente une entrée en matière plus longue qu'un exposé normal avec l'exposé du problème général avant celui du problème particulier. L'énoncé du sous problème apparaît en F5-6-7. Le résultat analysé automatiquement est proposé Figure 3.12 p. 70.

Dans la figure 9, la topologie de la *macro structure* est plus abstraite que celle des *schemata*, les cercles représentent tous les points de vérification théoriques pour les relations stipulées, l'ensemble des possibles. Cette représentation me paraît nécessaire pour gérer des réalisations dans des langues différentes, qui seront exprimées à travers des *schemata* comme en figure 8. Tous les débuts et fins des constituants sont à examiner. C'est ce que j'appelle « maillage » par analogie avec la géographie et la physique. Toutefois, le schéma reste une vue de l'esprit, il ne décrit ni des pages, ni des textes attestés.

En pratique, si l'on traite un corpus homogène, tous les points de vérification ne sont pas nécessaires, il suffit que la structure soit reconstituible. Une construction est marquée suivant le principe d'économie, toute paire adjacente de points théoriques est réduite à un seul point marqué. Pour chaque langue, on peut reconstruire des *schemata* concrets en ajoutant des marques attendues.

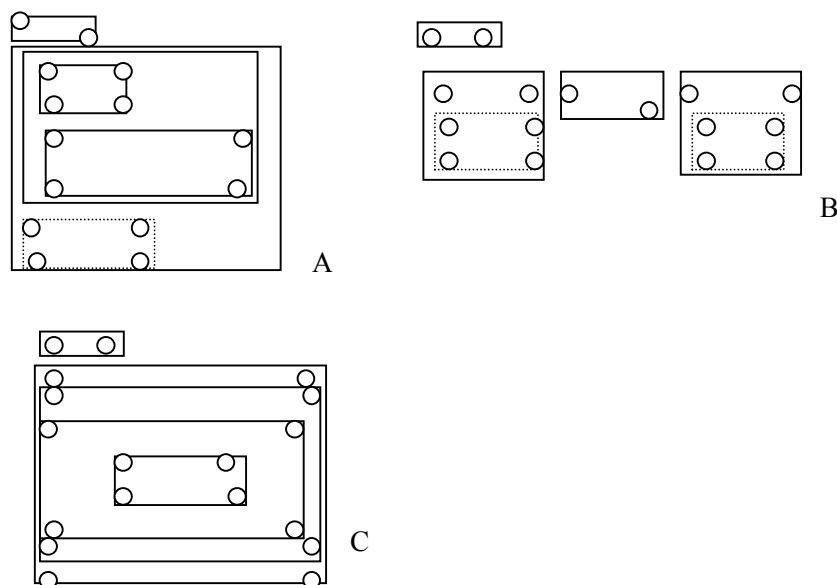


Figure 2.9. Constructions canoniques et points de vérification théoriques

La « systémique technologique » suppose une réunification de paradigme entre la logique interprétative des modèles initiaux (linguistiques dans notre cas) et de la logique opératoire de l'ingénierie de conception (Aït El-Hadj, 2006). L'algorithmique du texte est la technologie adaptée au diagnostic de corpus, pour détecter des répétitions, des régularités, les positions où se trouvent des *hapax* entre autres (Crochemore, Hancart & Lecroq, 2001 ; Crochemore & Rytter, 2002 ; Navarro & Raffinot, 2007). Encore faut-il donner à la « chaînologie » ou étude des chaînes de caractères quelques opérations supplémentaires pour traiter effectivement de textes (avec une structure hiérarchique) et non de lignes ou de plans.

Enfin, il est nécessaire de transposer concomitamment des opérations et des indices, pour passer de la reconnaissance humaine de la structure et du sens à une mécanisation de la segmentation et de l'étiquetage ou du coloriage. Les techniques d'algorithmique du texte sont souvent très coûteuses, si on les prend « au pied de la lettre », au grain caractère et sur un alphabet égal à un alphabet d'écriture. Elles sont beaucoup optimisées si les hypothèses linguistiques sous-jacentes au traitement sont explicitées comme contraintes sur les solutions valides.

5. CONCLUSION PARTIELLE

Le choix des moyens à mettre en œuvre dépend des ambitions poursuivies. Une pondération est nécessaire entre la couverture (nombre de cas différents traités), la robustesse (capacité à traiter des

variantes peu communes) et la fiabilité de la procédure. Il faut également envisager le coût en termes de temps de calcul, et de ressources en mémoire. Si l'on veut assurer des résultats fiables, à partir de marques morphologiques par exemple, il faut rester proche de la fenêtre d'observation idéale en fonction de la langue et du corpus traités. Si l'on veut s'en éloigner sans crainte d'extrapolation abusive, il faut gérer les marques absentes et spécifier les contraintes de reconstruction, ce qui est toujours difficile.

Trois idées fortes devraient néanmoins se dégager de la présentation qui précède. En premier lieu, on peut choisir de modéliser un corpus (l'objet à analyser) ou bien de modéliser des opérations (la méthode d'analyse). J'ai choisi de modéliser la méthode d'analyse. Je travaille ordinairement en binôme avec un informaticien, plus ou moins expérimenté. Souvent, au final, la modélisation est un hybride et elle n'est pas toujours explicitement commentée. On peut cependant, dans un modèle *ad hoc*, envisager de gérer les variantes d'un modèle linguistique canonique, pour tenir compte des paramètres du genre, du style ou autre et faciliter l'interprétation du texte en sortie. En analyse automatique, il est nécessaire de bien spécifier les limites de la méthode.

En second lieu, il existe un choix assez étendu de grilles d'analyse des textes/ du discours, qui correspondent à des objectifs différents, mais aussi à des traditions culturelles différentes. Le choix d'un cadre théorique correspond à un diagnostic de la tâche. Il doit tenir compte de l'application visée, des demandes des utilisateurs, de leurs habitudes et références implicites dans la lecture des textes et du type de corpus à traiter. Cela est valable autant pour les analyses manuelles, dans le cadre de l'enseignement en particulier, que pour les analyses informatisées.

En troisième lieu, les modèles linguistiques syntaxiques traitent des relations et des rôles, dans un cadre de référence. Tels que je les perçois, ils supportent de multiples adaptations, pour tenir compte de réalisations différentes, en fonction de la langue, du genre, du style ou du registre et enfin du grain d'analyse. Ce sont autant de paramètres caractérisant un corpus de manière différentielle, et non absolue, partant, autant de degrés de liberté à considérer dans une application.

Dans les chapitres suivants, je présente quelques logiciels d'étude dont j'ai assuré la conception, le plus souvent programmés par des stagiaires ou des collègues, pour donner une illustration de ma démarche, à travers des applications. Il s'agit d'une heuristique qui a évolué en plusieurs étapes, de la description vers la projection. Je reviendrai dans le dernier chapitre (p. 102) sur l'effort à faire pour mieux définir le métalangage, en particulier pour « projection » ou « induction ».

Dans la pratique, l'effort pédagogique que je mène consiste à passer d'une modélisation *ad hoc* sur des tâches très ciblées (corpus monolingue et mono-genre) à une modélisation des opérations plus générique. Il s'agit de conserver les opérations communes en travaillant sur un corpus bilingue ou multilingue, des genres différents, des styles très variables dans une collection. Les opérations, typiquement de comparaison, doivent alors être manipulées avec des variables au lieu d'opérandes dont la valeur est immédiate.

Chapitre 3

Multilinguisme et analyse automatique de textes

Le multilinguisme est un enjeu pour la recherche d'information. Alors que dans les années 80 il était supposé que l'anglais était la langue commune de la planète web, destinée à supplanter les autres, la place de l'anglais n'a cessé de diminuer dans les documents mis en ligne. Les standards de communication incluent avec la norme ISO-IEC 10646 — communément appelée Unicode — de nombreuses écritures que l'on croyait vouées à la disparition.

L'intérêt de la syntaxe est de permettre des comparaisons entre langues en fournissant un cadre d'analyse plus abstrait portant sur des relations. Un sujet en français correspond à un sujet en russe ou en allemand. Le rôle par rapport au verbe est le même, mais les façons de le marquer varient, positionnellement et morphologiquement. De la même façon, les rôles ou fonctions discursives sont identiques d'une langue à l'autre, mais les formes et leur position varient dans les textes, par exemple ce que l'on dit à la fin d'un article académique en japonais est dit au début d'un article académique en anglais. L'information se retrouve mais elle est disposée différemment.

Pour la majorité de nos applications à l'heure actuelle, un seul référentiel est utilisé ; pour faciliter le travail des étudiants, un *schema* est parfois établi pour une langue et un ordre de grandeur. Il suffit alors de classer les textes qui se prêtent à l'analyse visée et ceux qui se s'y prêtent pas. Ce classement ressemble à du profilage de textes comme le proposent Habert *et al.* (2000).

Les logiciels d'étude présentés ci-dessous ont une couverture bilingue, au minimum. En effet, partant d'une situation de traitements monolingues, dans laquelle seules les formes sont

représentées explicitement, il m'a fallu montrer la similitude du raisonnement déductif, c'est-à-dire du processus informatique de calcul sur un patron, puis sur un autre, pour avancer l'idée de constance de l'opération, sur des opérandes par nature variables. Des essais ont été réalisés sur un grand nombre de langues. Vergne s'appuie sur la notion de proposition pour analyser 5 langues avec des procédures endogènes (2002). Un plus grand nombre 7, 8, 14 langues sont traitées par un même algorithme à l'échelle du texte (Lucas & Giguët, 2005) comme on le verra p. 65 et p. 69 sq. A vrai dire il est vain de préciser combien de langues, puisque ce nombre ne tient qu'aux occasions, la coïncidence d'un corpus en un idiome et la disponibilité de lecteurs pouvant juger du résultat dans cet idiome. Les algorithmes sont construits davantage pour un genre ou une tâche que pour un idiome.

1. L'APPROCHE A BASE DE PATRONS ANCRES

Les travaux réalisés s'inscrivent dans la lignée initiée par Vergne et également illustrée par Bourigault en France, sous le nom de méthode robuste sans dictionnaire, méthode endogène ou encore distributionnelle (Bourigault, 1992 ; Vergne, 1998, 2002). Dans cette optique, l'analyse automatique est conçue explicitement comme une attribution de valeur syntaxique, dans le contexte de la phrase, à l'aide de patrons. Les relations traitées sont internes à la proposition — sujet-verbe, verbe-compléments directs, verbes-circonstants, nom-compléments — et entre propositions, principale-subordonnée (Vergne, 2002).

L'analyse distributionnelle a été formalisée par Bloomfield puis Hockett pour l'analyse syntaxique de langues amérindiennes encore inconnues (Hockett, 1958). La méthode consiste à séparer la position et la forme et à noter les formes remarquables ou fréquentes apparaissant aux positions remarquables, début et fin des unités prosodiques. La sélection de la marque est l'espace de recherche approprié pour une forme qui devient syntaxique ou grammaticale quand on a compris sa fonction dans le système de la langue étudiée. On l'appelle alors marque syntaxique.

Cette méthodologie est commode à transposer en analyse automatique et en apprentissage (Déjean, 1998a, b). Les principes distributionnels associés à des automates sont également utilisés en recherche d'information sur la toile (Kushmerick, 1997 ; Muslea *et al.*, 2001).

Les principes relationnels et l'indépendance par rapport à des attributs mémorisés dans des dictionnaires sont conservés dans mon approche du texte. J'ai conservé également un traitement déterministe. Par ailleurs, l'indépendance par rapport au lexique permet une couverture multilingue, moyennant quelques classes d'équivalence pour les patrons. Mon apport consiste à ancrer les patrons dans des espaces de recherche. Ensuite les formes sont traitées comme des variables puis comme des inconnues.

2. LA DETECTION DU DISCOURS RAPPORTE DANS LA PRESSE

Le discours rapporté a été étudié en linguistique du discours (Rosier, 1998a, 1998b, 2009). La détection automatique de discours rapporté dans la presse est une application jugée difficile et elle est très peu documentée. Dans l'optique dominante, celle de la consultation de ressources en mémoire, elle se heurte à l'impermanence des sujets d'actualité. Face à l'écueil du coût exorbitant de la maintenance de bases de données, concernant les personnes et leur rôle, les recherches ont marqué le pas. Cette problématique écartée ressurgit cependant sous l'appellation « détection des points de vue », dans des compétitions d'extraction d'information (TREC) ou de résumé (DUC). Les citations sont généralement détectées par l'analyse syntaxique de phrases, appliquée à toutes les phrases sur le modèle de corpus annotés, donc d'une description (Carlson *et al.*, 2003). On notera en Europe les travaux de Redeker qui critiquent l'approche mémoire à grain très fin de la RST revue par Carlson *et al.*, et proposent une vision moins rigide du texte (Redeker & Egg, 2006 ; Egg & Redeker, 2008). Dans l'approche française dite contextuelle, le grain reste celui de la phrase. Les indices stockés en mémoire sont les verbes de la classe {dire}, les noms propres et les noms de fonction comme *ministre, délégué, PDG* etc. (Mourad & Desclés, 2004).

La philosophie mise en œuvre dans mes travaux est qu'il vaut mieux peu de critères fiables qu'un grand nombre de critères hasardeux. D'autre part, dès le début, je pars de l'article, en appliquant une procédure de subdivision (approche projective descendante), et non d'une procédure de regroupement de segments identifiés un à un (procédure ascendante et cumulative).

Dans un premier temps, les travaux que j'ai réalisés ou encadrés ont fait appel à des procédures simples. Toutefois, comme les objets manipulés sont des *relations*, beaucoup d'étudiants ne parviennent pas à les manipuler. L'approche dite « ancrée » est une approche « positiviste » ou au premier degré, dans le sens où elle n'utilise que la présence de marques définies a priori, et échoue lorsque l'indice est absent. Une hypothèse est formulée humainement, et transcrite comme relation ordonnée (*pattern* ou *patron*). Elle sert de base de calcul dans les traitements automatiques d'un corpus pour répondre à une demande des utilisateurs.

La détection des patrons ou structures est faite par l'intermédiaire des expressions régulières¹⁷ et des techniques de reconnaissance de motifs dans les chaînes de caractères, procédures connues sous le nom d'algorithmique du texte (Crochemore, Hancart & Lecroq, 2001 ; Crochemore & Rytter, 2002 ; Navarro & Raffinot, 2007).

Dans un second temps, j'ai exploité les résultats de tests par des expressions régulières pour raffiner l'analyse en incluant les indices absents, plus exactement non trouvés. Cela permet de travailler au second degré. En effet, un avantage des méthodes projectives « robustes » tient à la

¹⁷ expression régulière ou expression rationnelle : notation pour représenter des ensembles de chaînes de caractères et les manipuler informatiquement, par exemple remplacer une chaîne par une autre.

sous spécification de certains éléments du patron, comme on le verra dans les exemples détaillés ci-dessous. On ne travaille plus sur la présence d'une chaîne *donnée* a priori, mais sur des distances d'édition, sur la récurrence ou non d'une chaîne *quelconque*. C'est aussi la raison majeure du rejet de ces méthodes, par les informaticiens formés au traitement mot à mot. Un patron qui n'est pas relevé point à point, exactement comme spécifié, est jugé mal défini dans l'approche positiviste et énumérative. Au contraire, dans l'approche différentielle et relative que j'ai adoptée, un certain degré de liberté est possible et on peut générer des ressources (Déjean, 1998, 2002). On n'a pas besoin de savoir si *said* est répété, mais seulement de savoir si une marque morphologique *m* est répétée quelque part. La répétition concerne n'importe quelle chaîne, connue ou non du programmeur, et le calcul renvoie « said. » ou « a-t-il ajouté. » ou « といった。 »

2.1. La détection de citations en anglais

Le postulat que j'ai formulé est que « les articles de presse contiennent du discours rapporté et le discours rapporté est marqué ». L'hypothèse linguistique est exprimée sous forme explicite de la manière suivante. Une citation est une structure actancielle définie par des patrons ordonnés en triplets (source —> relateur —> discours rapporté). La flèche se lit « suivi de ». Suivant l'ordre, on définit un motif normal, par exemple pour l'anglais (source —> relateur —> discours rapporté) et un motif inversé (discours rapporté —> source —> relateur). Ce critère est positionnel. La liste des motifs est établie en fonction du corpus traité, car les patrons ont des formes variantes en français et en anglais par exemple.

La relation ordonnée abstraite et typée sert de base de calcul dans les traitements automatiques, (dans un raisonnement déductif du programmeur), c'est donc une démarche différente de celle qui est habituellement appliquée par simple comparaison de formes sur les patrons. Chaque élément du triplet (un tagmème) a deux attributs, une position et une forme. En cas d'attribut manquant, soit la position soit la forme, le contexte permet de former un *pattern* ou patron de citation valide. Suivant la morphologie, les citations sont en outre caractérisées comme discours direct (avec des guillemets) ou discours indirect.

Dans le projet Linguix (1998-99), j'ai réalisé un module robuste de détection de citations dans des dépêches et articles de presse. Contrairement à l'approche courante, je n'ai retenu qu'un très petit nombre d'indices morphologiques sûrs (*said* et *told*) couplés à des indices typographiques pour former des patrons de discours rapporté. L'espace de recherche est défini à deux niveaux, global (le corps de dépêche) et local, le paragraphe (dans les dépêches, le paragraphe contient une ou deux phrases). Les catégories utilisées et coloriées dans la sortie sont les suivantes : l'en-tête en violet (lieu et date d'émission, source émettrice) ; le discours rapporté direct (rouge), le discours rapporté indirect (vert) et le refus de commentaire (magenta).

Dans les copies d'écran suivantes, le discours rapporté (DR) est colorié, j'ai ajouté en marge les catégories correspondantes : dans la figure 1 les types de DR (direct ou indirect).

Yemen get its first privately-owned airline
01:12 p.m Jun 11, 1998 Eastern

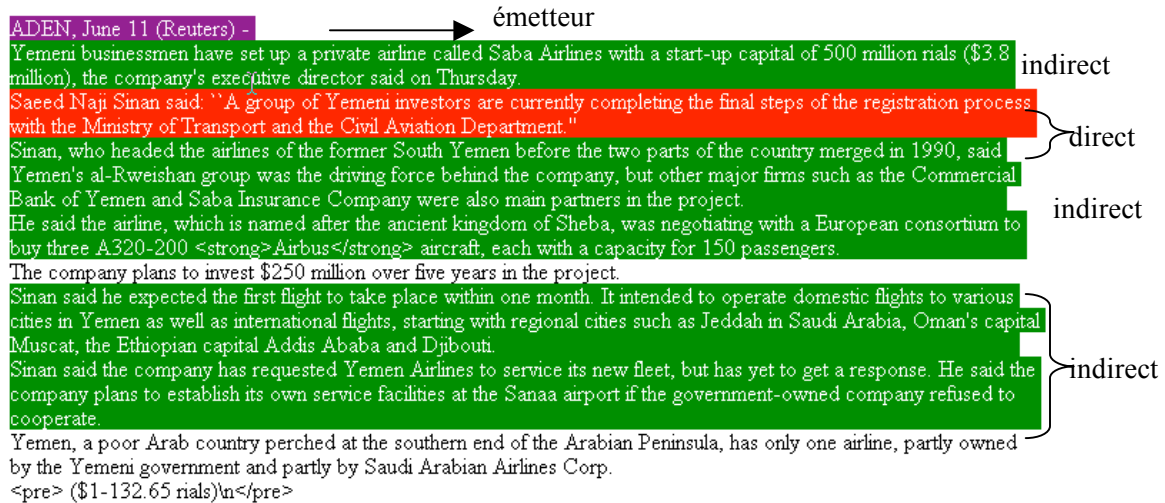


Figure 3.1. Coloriage d'une dépêche en fonction du type de discours rapporté

FOCUS-France seeks Aerospatiale flotation, allies
01:21 p.m May 27, 1998 Eastern

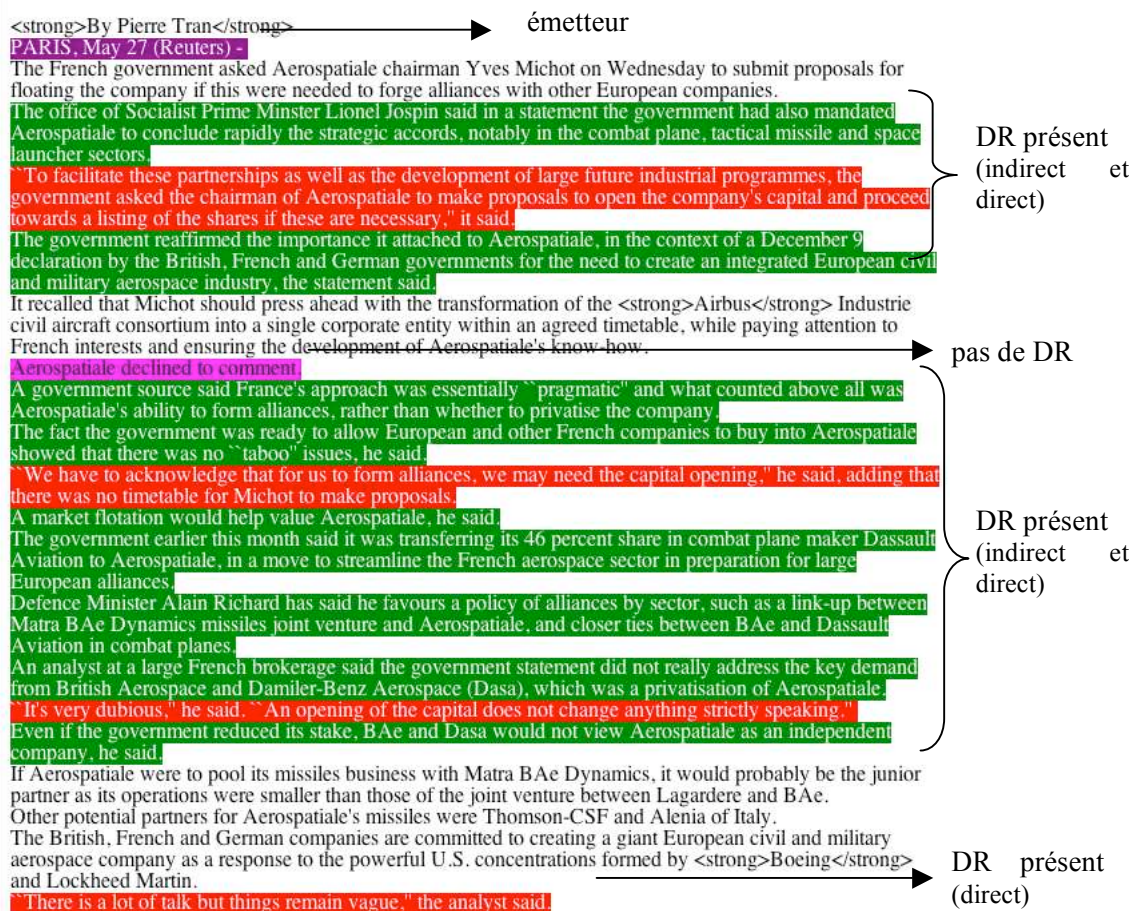


Figure 3.2. Coloriage d'une dépêche avec DR et absence de commentaire

Les indices morphologiques « ambigus » comme *reported*, *announced* ou *comment* ne sont pris en considération qu'à certaines conditions, leur position dans la dépêche et la co-location dans la dépêche de patrons sûrs.

Dans la figure 2, la distinction entre présence de DR et absence de commentaire (le fait qu'un discours attendu n'ait pas été prononcé) est ajoutée en marge.

Quoique les principes exploités soient fort simples, ce logiciel obtient de bien meilleurs résultats que les détecteurs lexicaux phrase à phrase. Mais c'est justement parce que les principes de calcul sont simples et donc contrôlés que ces résultats sont bons.

2.2. Détection de chaînes de citations

La détection de chaînes coréférentielles de citations est très demandée par les utilisateurs, car elle permet d'associer un locuteur avec ses déclarations. Mais comme les analyseurs classiques ne dépassent pas le cadre de la phrase, les anaphores, communes dans le genre journalistique, posent de gros problèmes (Siddharthan, 2003). Ainsi dans l'exemple 1, on ne peut trouver l'identité du locuteur, désigné par un pronom dans la fenêtre d'observation de la phrase.

Exemple 3.1. Reuters 31/01/2001 (FDJ 8) Phrase anaphorique.

Il a prévenu que la France avait l'intention d'utiliser la crise de la vache folle et ses conséquences pour convaincre ses partenaires européens et la Commission de Bruxelles de la nécessité d'un changement de cap dans les pratiques agricoles.

2.1. En français

La détection de chaînes de discours rapporté dans la presse française est réalisée par le logiciel CitaLoc (Giguet & Lucas, 2004). Un informant est une entité qui peut être désignée de plusieurs façons (reprises partielles, abréviations, pronoms, synonymes) mais réfère à la même personne physique ou morale. L'hypothèse est reformulée, « les articles contiennent du discours rapporté et il existe un informant marqué au moins ». Les patrons ordonnés concernent désormais des chaînes coréférentielles, en sus des citations. L'approche adoptée était en effet ascendante dans les premières implémentations. Dans un article (ou une dépêche), on attend une chaîne de citations (un informant au moins) la chaîne servant de modèle de référence. Méthodologiquement, une chaîne de citations est un couple ou un n -uplet ordonné de motifs de "citations" ayant un même informant, mais pas nécessairement le même locuteur individuel. On remarque que l'expression de l'informant par le journaliste revient à la forme initiale quand la chaîne est bouclée. La citation est donc le segment minimal ou maillon formant une chaîne.

L'espace de recherche est défini à trois grains, le corps d'article, le groupe de phrases et la phrase. Ces unités sont des tokens au sens statistique et au sens d'unités manipulées en machine. Une citation est observable soit dans un token phrase, soit dans un token dit contexte étendu (un groupe de k phrases ou un paragraphe). L'attribution d'une catégorie transitoire ou d'une fonction implique un certain nombre de tests concernant la présence ou l'absence d'indices typographiques

et morphologiques dans une fenêtre d'observation ou sélection. Peu de critères fiables, avons-nous dit. Le critère fiable exploité pour la détection des citations est syntaxique, c'est la collocation d'indices et l'ordre dans lequel un motif ou patron se présente.

Une citation est définie dans la sélection de la phrase par les triplets {source —> relateur —> discours rapporté}, appelé motif normal et {discours rapporté —> relateur —> source}, appelé motif inversé. Les indices sont typographiques ou morphologiques. En français, une grande variété de verbes est utilisée, il n'y a pas de stabilité sur une forme lexicale de relateur (alors qu'en anglais la forme canonique *said* est répétée). Le critère morphologique utilisé est donc un invariant grammatical (constance du temps). Les indices recherchés sont les morphèmes du temps le plus employé en français, c'est-à-dire le passé composé (caractères *a ...é, a ...it ou a ...u*).

Les indices ne sont jamais utilisés séparément, ils ne déterminent pas immédiatement l'affectation d'une variable. C'est l'identification *simultanée* (la co-occurrence) des trois variables ou de deux sur trois qui déclenche l'affectation. Autrement dit, il ne faut pas voir un patron comme une somme d'indices individuels, mais bien comme un tout, un 3/3. Si les positions des indices morphologiques relevés ne sont pas les positions attendues, alors ils sont oubliés. Si les positions sont conformes à un ordre attendu, mais que le nombre de critères morphologiques relevés n'est pas suffisant pour établir un calcul de déduction (l'attribution de la valeur "citation"), alors ils sont oubliés.

Une procédure de calcul est déclenchée en revanche si l'on dispose de deux variables sur trois dans un contexte borné (token phrase ou token groupe de phrases). Par exemple, si le contexte permet de construire le motif incomplet (?—> relateur —> discours rapporté), le recours au modèle permet de diagnostiquer que ? doit correspondre à la valeur de la "source". Dans le cas où une seule variable est calculée à partir d'un faisceau d'indices, aucune déduction fiable sur la position des deux autres n'est possible et elle est oubliée.

Voici d'abord une recherche locale aboutissant au résultat qu'il y a deux citations coréférentielles dans le passage extrait (un groupe de 2 phrases). Les indices recherchés sont en gras, les valeurs calculées d'après la position sont soulignées.

Exemple 3.2. Reuters 31/01/2001 (FDJ 8)

"Le contrat qu'elle avait passé avec les agriculteurs européens, il y a quarante ans, qui était de produire plus pour répondre à des problèmes d'autosuffisance alimentaire après la Seconde Guerre mondiale, est un contrat qui a été bien rempli par les agriculteurs, mais qui est totalement dépassé", **a relevé le ministre de l'Agriculture.**

Il a prévenu que la France avait l'intention d'utiliser la crise de la vache folle et ses conséquences pour convaincre ses partenaires européens et la Commission de Bruxelles de la nécessité d'un changement de cap dans les pratiques agricoles.

La recherche des indices dans le contexte de la phrase met en évidence 3 indices à exploiter, dans la première phrase, deux typo-dispositionnelles et une morphologique : " / *position début* marque la possibilité d'une ouverture de discours direct rapporté en tête et ", marque la fin

potentielle d'un discours rapporté. Les guillemets encadrent le discours rapporté direct. Les indices morphologiques (*a ...é* et *a ...u*) correspondent à un verbe au passé composé, potentiellement relateur. La séquence est incomplète (discours rapporté → relateur → ?), mais le recours au modèle permet de la faire correspondre à un motif inversé, et d'identifier que l'inconnue en position 3 correspond à la "source" (ou locuteur, à ce niveau de la citation individuelle).

Dans la seconde phrase, la recherche des indices met en évidence 3 indices à exploiter : *Il* / position début, marque la possibilité d'un locuteur anaphorique en tête ; *a ...u*, marque un relateur verbal potentiel ; *que*, marque potentiellement un discours rapporté indirect à suivre. L'exploitation des indices en co-présence et en ordre dans la phrase conduit à l'identification du motif en ordre normal (source → relateur → discours rapporté) dans la deuxième phrase.

On peut remarquer que les indices morphologiques (*a ...é* et *a ...u*) permettent effectivement de capter des verbes au passé composé. Ces verbes sont "citatifs" en contexte ou du moins introduisent effectivement une citation. Mais les indices morphologiques ne déterminent pas immédiatement l'affectation de la variable "relateur". C'est le contexte et notamment l'identification simultanée d'au moins une des deux autres variables qui déclenche l'affectation. En situation réelle, beaucoup d'indices sont relevés sans aboutir à une affectation effective, spécialement dans le cas de la variable "source" basée sur la reconnaissance de mots capitalisés consécutifs. Ils sont soulignés dans l'exemple suivant, où la co-présence dans le token phrase d'un passé composé et candidat d'un nom propre (capitalisé) n'aboutit pas à l'affectation de la valeur "citation".

Exemple 3.3. Détection d'indices abandonnés. *La Tribune*, 01/02/2001 (FQJ 41)

La Commission européenne a adopté, le 31 janvier, un projet de correctif du budget 2001 de l'Union européenne de 971 millions d'euros (un peu plus de 6,3 milliards de francs) destinés exclusivement au financement des mesures d'urgence décidées fin décembre par les ministres de l'Agriculture des Quinze pour faire face à la crise du secteur bovin.

Le fait que l'on puisse oublier un indice revient à tenir compte d'une opération effectuée (un test de présence/absence ou de co-présence) et à travailler sur les traces, non directement sur la forme ou sur l'étiquette de la forme. Cela suppose qu'on ait un algorithme projectif « dédié », qui dirige le traitement au niveau général.

Si tous les tests sont négatifs, le résultat du traitement doit être un *jugement négatif*, « il n'y a pas de discours rapporté dans cet article ». L'exploitation des tests négatifs n'est pas neuve, mais toujours rare et difficile à enseigner. L'exploitation des contre-exemples n'est déjà pas si fréquente (Ganascia, 1985 ; Yangarber, 2003). Cependant, la difficulté tient à la différenciation entre l'absence d'indice et l'absence de repérage d'indice ou une absence de marque et ce que les linguistes appellent une marque zéro. La réflexion sur le système est nécessaire (Ganascia, 1991).

La détection de chaînes de citation part du premier paragraphe. Le postulat est qu'il y a un informant au moins. L'informant plausible (un nom « propre » marqué par capitalisation) y est

recherché. Les citations suivantes sont rattachées à cet informant à condition qu'il n'y ait pas de rupture de temps, et pas non plus de nouveau nom propre.

La détection des chaînes de citation est fondée sur l'agrégation des citations en chaîne selon le principe du *nihil obstat* (tant qu'il n'y a pas de forte contre-indication) développé par Vergne pour l'analyse de phrase. Cette procédure suffit dans la plupart des cas (environ 80% des dépêches ont un seul informant et environ 70% des articles ont un informant ou quelques informants présentés successivement). Toutefois, il y a des erreurs dans les dépêches denses et dans les articles, présentant des points de vue contradictoires (5% des cas). Ce problème ne peut être résolu que dans le cadre d'un algorithme plus sophistiqué, tenant compte du point de vue.

Le logiciel CitaLoc pour le français est utilisé en entreprise. Il détecte les citations avec un taux de succès de 98%, donc très compétitif¹⁸. CitaLoc produit « en sus » par rapport aux systèmes à objectif comparable, la résolution des anaphores pronominales et nominales (en moyenne 84% de succès). Son « rapport qualité prix » est intéressant. Il ne nécessite pas de ressources externes, ni de dictionnaires de verbes, ni de dictionnaires de noms propres, très coûteux à mettre à jour, ni de dictionnaires de titres et fonctions. Ce logiciel a été réimplémenté par Emmanuel Giguët en 2004 sur la plate-forme Wims sous le nom de « WimsQuotes ». Je reviendrai plus loin p. 63 et 65 sur la mesure quantitative des résultats, en situant le plancher de comparaison au niveau des plus mauvais détecteurs de citations réalisés par les étudiants de première année du master de l'Université de Caen. Une critique des mesures est esquissée dans les points à développer, chapitre 6.

31Janv2001 FRANCE: **Glavany plaide pour une nouvelle PAC où l'on produirait "mieux"**.
PARIS, 31 janvier (Reuters) - Préconisant une rupture avec le modèle productiviste de l'après-guerre, Jean Glavany a **plaidé** pour une politique agricole européenne s'appuyant sur la qualité et le respect de l'environnement.

"Il faut passer un nouveau contrat avec les agriculteurs européens. Je pense que ce nouveau contrat doit être de produire mieux", a **estimé** [@Jean Glavany@] le ministre de l'Agriculture à l'Assemblée nationale.

Mettant l'accent sur le respect de l'environnement, la sécurité sanitaire et la qualité des produits, Jean Glavany a **estimé** que la PAC devait "être refondée en profondeur".

"Le contrat qu'elle avait passé avec les agriculteurs européens, il y a quarante ans, qui était de produire plus pour répondre à des problèmes d'autosuffisance alimentaire après la Seconde Guerre mondiale, est un contrat qui a été bien rempli par les agriculteurs, mais qui est totalement dépassé", a **relevé** [@Jean Glavany@] le ministre de l'Agriculture.

[#Jean Glavany#] Il a **prévenu** que la France avait l'intention d'utiliser la crise de la vache folle et ses conséquences pour convaincre ses partenaires européens et la Commission de Bruxelles de la nécessité d'un changement de cap dans les pratiques agricoles.

"Maintenant devant le débat public européen qui est posé, nous devons essayer d'aller plus loin au niveau de l'Europe, notamment pour tirer les leçons de la crise bovine", [#Jean Glavany#] a-t-il **lancé**.

"La détermination du gouvernement français est de tirer les leçons de cette crise et d'acter le pas dans cette reconversion de l'agriculture (...) vers ce modèle qualitatif que nous attendons", a **conclu** Jean Glavany.

(c) Colloque G-dit Reuters Limited 2001.

30

Lucas, Giguët et Vergne

Figure 3.3. Détection de citations et de chaîne de citation coréférentielle par CitaLoc

¹⁸ Nous n'avons pas pu les confronter à des systèmes industriels à base de ressources mémorisées, du fait de la compétition dans ce domaine, qui entraîne la rétention d'information de la part des concepteurs ou détenteurs de logiciels. De plus les bases de comparaison ne sont pas fournies. Voir chapitre 6 pour la discussion.

2.2. En anglais

Plusieurs logiciels d'étude exploitant un corpus stabilisé de dépêches en anglais ont été réalisés par des étudiants entre 2001 et 2002, puis en 2008 sur des dépêches en ligne, les fournisseurs de corpus étant laissés à la discrétion des programmeurs. Des patrons de chaînes sont définis dans l'espace de recherche de la dépêche.

Exemple 3. 4. dépêche ADJ174, une chaîne co-référentielle unique

FAA to insist on checks of Boeing plane rudders

05:44 p.m Jun 15, 1998 Eastern

By Tim Dobbyn

WASHINGTON, June 15 (Reuters) - Rudder pedals on nearly all Boeing aircraft will be checked after a pilot was forced to hand control to his first officer because of a malfunction, the Federal Aviation Administration said on Monday.

FAA spokeswoman Kathryn Creedy said an airworthiness directive for Boeing 737, 747, 757, 767 and 777 models would be issued shortly.

The directive would insist all U.S.-registered aircraft undergo checks recommended by Boeing Co. in a March 26 service bulletin to operators around the world.

The action was prompted by an incident last July when the captain of a Futura 737 had to hand control to his first officer after finding the rudder pedals wouldn't work just after touching down in Palma, Spain. Futura, a charter airline, is a unit of Aer Lingus.

A bolt meant to attach a push-rod to the rudder pedal was later found on the floor of the Futura plane's cockpit.

Another 737 was found to have a loose rudder pedal and inspections of 137 further 737s during regular maintenance revealed four more loose fasteners. "We decided to do something about this," Creedy said.

FAA said incorrect installation of the fasteners during manufacture was identified as the cause and Boeing had since introduced a new procedure and checks during assembly.

Creedy said models other than the 737 were included in the coming airworthiness directive because they all had similar rudder pedal assemblies.

The notice will affect 1,477 U.S.-registered planes. The 4,095 aircraft of these types around the rest of the world may be checked voluntarily or at the direction of local civil aviation authorities.

FAA will give airlines 90 days to make the checks that it estimates will take about an hour per plane.

The problem is thought to have no bearing on the crash of a USAir 737 just outside of Pittsburgh in 1994 although the continuing probe of that accident has focused heavily on a possible rudder malfunction.

"That was examined and found that it (the pedal bolt) was not related to the USAir crash," said Creedy.

The news of the impending rudder inspections comes as airlines are still examining 737s for possibly dangerous fuel tank wiring.

Copyright 1998 Reuters Limited. All rights reserved. Republication and redistribution of Reuters content is expressly prohibited without the prior written consent of Reuters. Reuters shall not be liable for any errors or delays in the content, or for any actions taken in reliance thereon.

Dans cet exemple en anglais, l'informant principal est la *FAA*, présent dans le titre et repris sous la forme étendue *Federal Aviation Administration* dans le premier paragraphe, associée au relateur *said* constant en anglais et à un circonstant temporel *on Monday*. L'indice du début de chaîne tient à cette conjonction dans la phrase et le paragraphe (la chaîne de caractères disjointe *said...day*). Concernant la forme de l'informant, une chaîne permettant d'amorcer la recherche est capitalisée (première lettre en majuscule) ou entièrement en majuscules. Elle doit être co-occurrence avec *said* dans la phrase. Concernant la position dans le corps de texte, le premier paragraphe est un lieu spécial, où les contraintes d'identité par rapport au titre et par rapport aux reprises ultérieures ne s'appliquent pas. Même la contrainte de présence de *said* est relâchée à cette position, car le DR n'est pas toujours présent : cela arrive souvent si le titre est citationnel. Autrement dit, c'est le principe de différence sur la forme de la citation qui est exploité, et non l'identité, entre l'antécédent

de la chaîne et les maillons qui s'y rattachent. La constance du point de vue en revanche est établie par l'ancrage circonstanciel.

Dans le second paragraphe, un locuteur est spécifié comme porte-parole de la FAA. Les reprises anaphoriques subséquentes contiennent une sous-chaîne de cette mention, en l'occurrence soit *Creedy* soit *FAA*. La chaîne co-référentielle est construite par agrégation, en adoptant le principe du *nihil obstat* (tant qu'il n'y a pas de forte contre-indication), ici pas de changement de verbe, ni de circonstant. La figure 4 schématise l'article cité sous forme de titre dominant un corps de texte représenté et annoté ici au grain paragraphe.

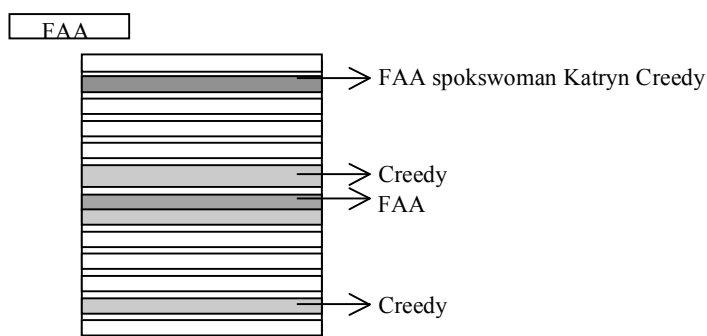


Figure 3.4. Position des maillons de chaînes de citation référant à un informant unique

Le résultat de l'analyse esquissée ci-dessus est illustré par une copie d'écran, où les maillons d'une chaîne ont la même couleur tant qu'elle a un même informant (Figure 5).

FAA to insist on checks of Boeing plane rudders
05:44 p.m Jun 15, 1998 Eastern

By Tim Dobbyn
WASHINGTON, June 15 (Reuters) -

Rudder pedals on nearly all Boeing aircraft will be checked after a pilot was forced to hand control to his first officer because of a malfunction, the Federal Aviation Administration said on Monday.

FAA spokeswoman Kathryn Creedy said an airworthiness directive for Boeing 737, 747, 757, 767 and 777 models would be issued shortly.

The directive would insist all U.S.-registered aircraft undergo checks recommended by Boeing Co. in a March 26 service bulletin to operators around the world.

The action was prompted by an incident last July when the captain of a Futura 737 had to hand control to his first officer after finding the rudder pedals wouldn't work just after touching down in Palma, Spain.

A bolt meant to push a push-rod to the rudder pedal was later found on the floor of the Futura plane's cockpit.

Another 737 was found to have a loose rudder pedal and inspections of 137 further 737s during regular maintenance revealed four more loose fasteners. "We decided to do something about this," Creedy said.

FAA said incorrect installation of the fasteners during manufacture was identified as the cause and Boeing had since introduced a new procedure and checks during assembly.

Creedy said models other than the 737 were included in the coming airworthiness directive because they all had similar rudder pedal assemblies.

The notice will affect 1,477 U.S.-registered planes. The 4,095 aircraft of these types around the rest of the world may be checked voluntarily or at the direction of local civil aviation authorities.

FAA will give airlines 90 days to make the checks that it estimates will take about an hour per plane.

The problem is thought to have no bearing on the crash of a USAir 737 just outside of Pittsburgh in 1994 although the continuing probe of that accident has focused heavily on a possible rudder malfunction.

"That was examined and found that it (the pedal bolt) was not related to the USAir crash," said Creedy.

The news of the impending rudder inspections comes as airlines are still examining 737s for possibly dangerous fuel tank wiring.

Figure 3.5. Détection des citations avec un informant et rattachement à une chaîne

Il peut se produire des erreurs. Dans un article avec plusieurs informants, 4 chaînes sont détectées, alors qu'il y en a 5 (Figure 6). La phrase entourée par une ellipse sur la copie d'écran ne contenait pas l'indice grammatical *said*, elle n'a pas été retenue comme candidate. On remarquera en conséquence une erreur de rattachement (*They said* en rose est rattaché à *Economic Class males* en rose). Les autres chaînes sont correctement détectées. De même, la cataphore, une difficulté insurmontable quand on travaille phrase à phrase, est surmontée quand on travaille sur le corps de texte. Deux cas de cataphore locale bien résolus sont signalés par une flèche blanche (Figure 6) ; une citation « en suspens » qui précède une citation avec mention de locuteur, cas qui cause des difficultés aux logiciels à petit grain (Siddarthan, 2003).

1999-04-20:
MALAYSIA: WHAT THEY SAY ABOUT PLANES HAVING DESIGNATED TOILETS. (BTMAL)

HOW do you feel about having designated toilets for the males and females in an aircraft? Airbus Industrie asks four groups of frequent flyers for their opinions and this is what they say.

Business Class females said they were not interested; Business Class males said they were not interested but felt sure the ladies would be.

Economy Class females liked the idea. They said the toilets would be cleaner. Economic Class males said they would not have to queue up behind the ladies for the use of the toilet, which may take some time.

The question is one of many from the company's worldwide market survey on passengers' requirements and perception of aircraft.

It was launched by Airbus Industrie's large aircraft division in cooperation with its commercial department.

The survey, which focused particularly on specific points of the A3XX cabin, is the largest ever conducted by the company. It involved interviewing some 1,200 frequent flyers and moving two specifically constructed cabin mock-ups between eight cities across three continents.

Mr Bob Lange, A3XX product marketing director, said by putting passengers in the mock-ups, "we were able to test ideas, check feasibility and avoid potential problems for all our future cabins.

"Our objective was to understand the relationship between passengers and aircraft throughout the travelling process," he said in Airbus Industrie's publication called Forum.

To ensure unbiased results, the survey was conducted anonymously with neither the interviewees nor the participants aware of Airbus' involvement.

The frequent flyers selected excluded anyone connected with the aviation industry or the media but included a representative cross-section of people from different countries.

The survey also yielded data on passengers' perceptions of Airbus' corporate image.

"People see Airbus as a dynamic modern company, a challenge to Boeing, better able to respond to passengers' needs.

"They think of our aircraft as new, equipped with the latest generation technology, spacious and quiet. However, they do not attribute these qualities to one Airbus aircraft in particular," Lange said.

The survey, according to Airbus Industrie, reinforces the conclusions of the recent International Air Transport Association survey on the importance to frequent travellers of non-stop flights, shorter journey time and cabin comfort.

The worldwide survey was followed by a smaller exercise based on 140 Airbus Industrie's frequent travellers.

"We repeated the experiment in Toulouse because we were curious to know how representative our own travellers were of the general public as this could save cost in the future.

"The results are still being analysed but the researchers commented that our own travellers showed a more imaginative view of the future and were better able to express their desires than the average traveller," he said.

BUSINESS TIMES (MALAYSIA) 20/04/1999 P1

Figure 3.6. Détection de chaînes de citation et de l'informant pour chaque chaîne dans un article à plusieurs informants. L'ellipse marque un antécédent manqué, la flèche noire marque une erreur de rattachement. La flèche blanche marque la cataphore correctement résolue.

D'après mes expériences et évaluations sur un corpus journalistique tout venant, ce type de logiciel macro syntaxique robuste endogène obtient un meilleur score (en moyenne 85% de succès dans la détection des citations) que les logiciels à base de ressources lexicales (en moyenne 76% de succès)¹⁹. La voie lexicale est évidemment dépendante de l'état du dictionnaire, en relation avec le

¹⁹ Sur l'ensemble des résultats de promotions successives.

corpus. Les résultats évalués l'ont été sur un corpus trié sur la thématique aéronautique, avec des ressources lexicales concernant les avionneurs. Les résultats s'effondrent si le corpus entrant n'est pas trié. Les analyseurs macro syntaxiques au contraire produisent des résultats stables.

Toutefois, les résultats sont variables en fonction du style des dépêches, lui-même corrélé à la source du corpus, pour les deux voies d'approche. Les dépêches financières par exemple sont complexes et plus denses que les faits divers, elles causent régulièrement des résultats dégradés. En outre les analystes réputés ont un style très personnel. Les résultats ci-dessous font état d'une analyse de dépêches financières (le corpus le plus difficile), avec le logiciel d'étude EDDAP2 (Marfouk & Gilain, 2001). On voit par là que les conditions d'évaluation peuvent faire changer considérablement les pourcentages annoncés, un constat établi pour d'autres tâches similaires (Labadié & Prince, 2008b).

Tableau 3.3. Détection de citations dans des dépêches financières par EDDAP2

Source	Nb de citations	Citations détectées	Silence	Bruit	Correct
Reuters business briefs	113	104	9 □ 8,6%	1 □ 0,9%	103 □ 91,1%
LexisNexis	122	101	21 □ 20,7%	4 □ 3,9%	97 □ 79,5%
BusinessWire	130	102	28 □ 27,4%	3 □ 3,8%	99 □ 76,1%
Total Dépêches financières	365	307	58 □ 18,8%	8 □ 2,6%	299 □ 81,9%

Le logiciel GRSP *Global-first reported speech parser* a été réalisé ultérieurement avec Emmanuel Giguet en approche descendante. Je ne parlerai pas ici de l'évaluation sur une seule implémentation, ce qui ne me semble pas très sérieux, d'autant que les mesures sur des chaînes (et non des citations individuelles) n'ont pas été validées de manière externe.

2.3. Multilingue

La détection de chaînes de citation a été réalisée sur 8 langues différentes dans le cadre d'un devoir de master en 2007-2008. Les étudiants groupés en trinômes avaient pour consigne d'établir la procédure sur des articles de presse en ligne, de l'évaluer sur au moins 2 langues connues et de l'appliquer à au moins une langue inconnue. Les langues traitées sont surtout des langues indo-européennes : français, anglais, allemand, néerlandais, italien, espagnol, russe, plus le malgache.

Les corpus de presse traités sont hétérogènes en idiomes, des tests d'évaluation ont été faits pour 7 langues indo-européennes. Les corpus sont parfois aussi hétérogènes en genre, ainsi des blogs ou des forums ont été captés avec des dépêches, avec pour conséquence une augmentation du taux d'erreur. A titre d'exemple, une des meilleures réalisations est proposée, pour montrer la difficulté du calcul mais aussi la difficulté de l'évaluation des résultats.

Dans cet exemple, les dépêches sont annotées à la volée sur un flux de dépêches en ligne (Lecavelier, Laurence et Valentin). Les deux vues illustrent respectivement un cas de réussite et un cas d'erreur de repérage du locuteur, en français.

Locuteurs : Jo-Wilfried Tsonga

Tsonga de justesse



Eurosport - J.G. avec AFP - 13/05/2008 11:30

Jo-Wilfried Tsonga s'est fait peur lundi pour son entrée en lice face à Nicolas Mahut mais sera bien présent au 2e tour. Michaël Llodra a battu Guillermo Canas. Paul-Henri Mathieu, lui, s'est incliné face au fantasque Nicolas Kiefer. Gilles Simon est passé dimanche.

MS HAMBOURG - 1ER TOUR:

Lundi:

Jo-Wilfried Tsonga (FRA/N.14) bat Nicolas Mahut (FRA) 0-6, 7-6 (7/5), 6-2.

Prochain adversaire: Andreev ou Soderling.

Auteur d'un départ calamiteux au cours duquel il n'a réussi à marquer que 12 en une heure et 55 minutes.

A 23 ans, Tsonga a ensuite nettement amélioré son jeu, réalisant cinq aces et inscrivant 33 points sur ses 39 premières balles de service dans les deuxième et troisième sets. "Nicolas était meilleur que moi dans le premier set", a déclaré Tsonga. "Il m'a mis sous pression au fond du court et j'ai dû me battre. Je me battais contre moi-même dans ma tête, j'avais dans l'idée que je pouvais gagner et cela m'a fait revenir".

"Je dois travailler très dur"

"Je me sens bien sur terre battue, a ajouté le Manceau. Je peux bien jouer sur cette surface mais cela me prend du temps pour faire quelques ajustements. Avec un peu plus d'expérience, je ferais mieux.", a-t-il ajouté. Tsonga a également reconnu que, depuis son accession en finale de l'Open d'Australie, son statut avait changé sur le circuit.

"Tous les joueurs me voient différemment et les matchs semblent plus difficiles. Ils savent tous que je peux bien jouer et ils jouent de leur mieux contre moi. Ils élèvent leur niveau de jeu et me mènent la vie dure", a-t-il expliqué. "Je dois travailler très dur". Tsonga rencontrera au deuxième tour le vainqueur du match entre le Russe Igor Andreev et le Suédois Robin Söderling.

Figure 3.7. Détection réussie sur un article mono locuteur avec une chaîne

On notera que le locuteur principal ou informant de la chaîne est correctement identifié par son nom complet, qui figure dans le chapeau d'article, traité comme antécédent de la chaîne anaphorique de citations co-référentielles. Dans l'exemple de la figure 7, la reconnaissance de l'informant est exacte, mais il manque un maillon à la chaîne, l'intertitre « *Je dois travailler très dur* » n'ayant pas été traité.

Locuteurs : Justine Henin - côté que la joueuse pourrait décider de faire simplement un break de quelques mois - Belge

Henin, clap de fin?



Eurosport - J.G. avec AFP - 14/05/2008 11:31

Justine Henin, numéro 1 mondiale, dont les rumeurs de retraite vont bon train en Belgique, convoque la presse cet après-midi. Vainqueur de sept titres du Grand Chelem, la Wallonne, 25 ans, connaît un début de saison décevant.

Coup de tonnerre ou rumeurs infondées? Justine Henin, la plus grande championne belge, pourrait se retirer du circuit. C'est en tout cas ce qu'annoncent plusieurs médias outre-Québécois. Le journal flamand "Het Nieuwsblad", qui cite "une source fiable", est le plus catégorique en affirmant que Henin va annoncer sa fin de carrière immédiate. Le site Internet du groupe "Vers l'Avenir" est plus évasif: pour lui, Justine va annoncer son départ soit immédiatement, soit après Roland-Garros (qui débute le 25 mai), soit après les Jeux Olympiques. "La Dernière Heure/Les sports" indique de son côté que la joueuse pourrait décider de faire simplement un break de quelques mois.

La semaine dernière, la Belge avait déclaré "manquer de flamme" après son élimination au 3e tour du tournoi de Berlin par la Russe Dinara Safina. Depuis le début de l'année, la leader du tennis féminin est apparue en difficulté. En avril, à Miami, elle avait concédé une lourde défaite (6-2, 6-0) contre l'Américaine Serena Williams et en avait été très affectée. Pourtant, il y a dix jours, elle affirmait qu'elle jouerait "encore deux, trois, quatre, voire cinq ans". Difficile de discerner le vrai du faux. Monstre de combativité et de courage, la "reine Justine", affectée par son divorce et sans cesse torturée sur un court, connaît-elle un ras-le-bol général? A Limelette (centre de la Belgique), en fin d'après-midi, on devrait en savoir plus. La parole est à la championne...

Figure. 3.8. Détection incertaine sur un article avec trois chaînes et plusieurs informants

Dans l'exemple de la figure 8, on peut voir que la notion de « locuteur » de maillon de chaîne (citation individuelle) et « informant » de chaîne n'a pas été implémentée séparément. Des interférences apparaissent du fait que l'enchâssement de citations n'a pas été modélisé : un journal

(non repéré) cite une source citant Hénin. Le DR et l'informant sont intervertis dans la seconde citation d'un autre journal (non repéré). Enfin, le lien de co-référence entre « la Belge » et « Justine Hénin » n'a pas été établi.

Les tests de qualité ont été faits manuellement sur un échantillon de 100 dépêches présentant des citations. Pour évaluer la qualité de la détection des chaînes, j'ai séparé la reconnaissance des chaînes et la qualité intrinsèque de chaque chaîne détectée (Popescu-Belis, 1999). Une chaîne est jugée correctement détectée si aucun maillon ne manque ni n'est en surplus. La reconnaissance de l'informant est jugée correcte si tous les maillons sont reliés sans erreur à un informant chapeautant éventuellement plusieurs locuteurs individuels.

Tableau 3.4. Mesure d'efficacité sur la détection et la qualification des citations en anglais (100 dépêches, 332 citations) $f = 2 * (\text{précision} * \text{rappel}) / (\text{précision} + \text{rappel})$

Manuel	Détection automatique			Groupe	Qualification des locuteurs pour les citations détectées		
332 citations	f-mesure	Précision	Rappel		incorrecte	correcte	Taux de succès
	0,76	1	0,62	A	24 /207	183 /207	88,4%
	0,97	0,96	0,98	B	77 /327	250 /327	76,4%
	0,99	1	0,99	C	28 /330	302 /330	91,5%

Tableau 3.5. Comparaison de la qualification des locuteurs des citations et de la qualification des informants des chaînes en anglais (140 chaînes, 292 citations)

Qualité de la qualification des locuteurs (332)			Groupe	Qualité de la qualification des informants (157)		
incorrecte	correcte	Taux de succès		incorrecte	correcte	Taux de succès
149/332	183/332	55,1%	A	48/157	109/157	69,4%
82/332	250/332	75,3%	B	33/157	124/157	78,9%
30/332	302/332	90,9%	C	94/157	63/157	40,1%

Le tableau 4 présente les résultats de trois groupes d'étudiants sur l'anglais avec à gauche les mesures de précision et rappel utilisées par d'autres auteurs en recherche d'information (Siddhartan, 2003). Dans la partie droite, le classement est binaire pour les groupes considérés, B le plus faible et C le meilleur. Sur les citations détectées, 76 à 91% sont correctement associées à leur locuteur, les erreurs sont surtout rencontrées dans le cas peu fréquent de citations enchâssées. Ces chiffres sont très flatteurs, le mode de calcul mettant en avant tout ce qui peut être jugé acceptable.

Dans le tableau 5, le calcul est fait de manière plus simple par rapport au jugement humain. Je montre que le nombre de citations correctement associées à un locuteur n'est pas directement lié au nombre de chaînes correctement associées à un informant (le référent de la chaîne). Les résultats du du groupe A et du groupe B sont meilleurs sur les chaînes que sur les citations individuelles et meilleurs que ceux du groupe C, qui devient alors le plus mauvais. Les stratégies adoptées sont très variables. Le groupe C avec une méthode cumulative, ascendante de la partie vers le tout, obtient 91% de citations individuelles correctement reconnues et reliées au locuteur individuel, mais

seulement 40% de chaînes correctement construites et reliées à l'informant. Les groupes A et B dissocient les deux tâches et obtiennent de meilleurs résultats sur les chaînes que sur les citations individuelles. Le groupe A ne repère pas les citations indirectes au niveau des citations individuelles et gagne beaucoup sur les chaînes ; le groupe B enregistre peu de différences entre la reconnaissance des informants de chaîne et le repérage des locuteurs des citations.

Les meilleurs résultats sur les citations individuelles sont qualitativement comparables à ceux du tableau 3 obtenus pour l'anglais. Dans les meilleurs cas, fondés sur une bonne analyse macro syntaxique, les résultats sont stables par famille de langues, ce qui est illustré essentiellement par la famille latine (la mieux représentée dans les échantillons).

3. DETECTION DE LA STRUCTURE THEMATIQUE DANS LES ARTICLES DE PRESSE

La détection de thèmes est entendue comme détection de passages concernant le même sujet (Labadié & Prince, 2008b). Les travaux dans ce domaine sont très nombreux, on pourra se reporter à Lamprier *et al.* pour une synthèse des acquis (2008). Les techniques courantes sont fondées sur la statistique lexicale, et tiennent compte de la distribution des mots qui caractérisent des passages de texte, à l'exclusion du reste de l'article (Hearst, 1994). Cette technique est dérivée du calcul de fréquence relative d'un terme dans un document appartenant à une collection dit *tf/idf* (*term frequency, inverse document frequency*) de Salton (1989). Elle donne de meilleurs résultats pour l'anglais que pour les langues latines, qui évitent la répétition des mêmes mots, ainsi que pour les langues germaniques, qui utilisent des mots composés.

L'algorithme de Choi nommé *c99* est un des plus utilisés (Choi, 2000). Rapide et de complexité linéaire en temps, il procède par division et regroupement sur une mesure de similarité des mots. Les résultats sont meilleurs pour séparer des textes réunis dans un fichier que pour séparer des passages dans un même texte (Labadié & Prince, 2008a). D'autres techniques à l'aide de représentations vectorielles sont fondées sur des représentations sémantiques lexicales, ou morpho-syntaxiques relativement coûteuses, car l'unité de compte est la phrase, ce qui nécessite de puissants moyens de calcul (Reitter, 2003 ; Prince & Labadié, 2007). Les travaux suédois montrent une intéressante comparaison de stratégies informatiques (vecteurs, graphes, arbres) pour exprimer les *propriétés d'un modèle* dirigé par la grammaire de dépendance, confiné à la phrase (Nivre *et al.*, 2006). Les fenêtres d'observation plus étendues ne dépassent guère le paragraphe (voir p. 40).

En 2001, j'ai entrepris un travail de modélisation (génie linguistique) sur le découpage thématique d'articles de presse où « thématique » ne signifie pas « lexical ». Trois analyseurs de discours sont fondés sur la théorie de l'exposition d'après Yamada. Ils exploitent le principe d'asymétrie de la construction expositive, selon lequel le thème précède le rhème et est significativement plus petit que le rhème. Je suis passée d'un processus à critères morphologiques

endogène à un processus purement proportionnel. THEMA et UniThem traitent d'articles de presse, ThemAgora de forums. Les grands traits de la modélisation informatique relèvent de la stylométrie, ou mesure du style, sur des principes différentiels.

3.1. En français

Le logiciel d'étude THEMA, destiné à segmenter des unités thématiques dans des textes informatifs courts ou de moyenne longueur a été conçu par moi, algorithme compris et implémenté par Pascalie Pinatel en projet de DESS (option Réseaux) 2002-2003. Ce programme est opérationnel, il prend en entrée des articles de journaux ou magazines et fournit la hiérarchie des différents thèmes abordés, sur trois niveaux d'enchâssement ; le coloriage thématique des différentes zones de texte, selon le niveau de détail choisi ; les mots-clés pour chaque zone coloriée. Il s'agit d'une conception très novatrice, en ce qu'elle ne nécessite ni dictionnaire, ni grammaire de phrase. Elle est descendante.

L'algorithme que j'ai proposé exploite le principe d'asymétrie du modèle de l'exposition d'après Yamada, et s'enracine dans le relevé systématique de la mise en forme matérielle ou MFM. On recherche des structures de liste du type A (BBB), caractérisées par contraste, successivement sans et avec répétitions d'items similaires. Les espaces de recherche traités comme *tokens* d'analyse sont le fichier, le titre et le corps de texte, le paragraphe, la phrase et le virgule (unité ponctuée par une virgule ou point virgule). Les espaces de recherche réduits appelés « sélection de la marque » sont les sous segments de début et fin de chaque *token* (Déjean, 1998).

Ce sont les proportions des sous segments de texte calculées d'après la mise en forme matérielle qui sont comparées pour le regroupement d'unités thématiques. L'algorithme est donc arithmétique mais guidé par le modèle, non par les données. Le diagramme UML de la Figure 9 montre que c'est le modèle et ses propriétés qui guide le moteur.

Le moteur calcule les effectifs des échelons de MFM et l'arité des relations compatible avec le modèle. Il n'y a donc plus de « déroulé » des opérations compréhensible humainement. Le calcul se fait sans étapes en fonction des contraintes spécifiées. Cela revient à détecter des séquences d'items similaires et des bornes contrastées « morphologiquement » c'est-à-dire par comparaison des chaînes de caractères entre elles. On extrait ensuite les chaînes de caractères répétés des espaces de sélection pour en extraire les mots-clés, par des procédures classiques (Boyer-Moore, Wu-Mamber).

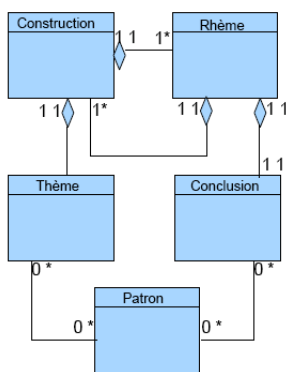


Figure 3.9. Diagramme UML structure des données du modèle

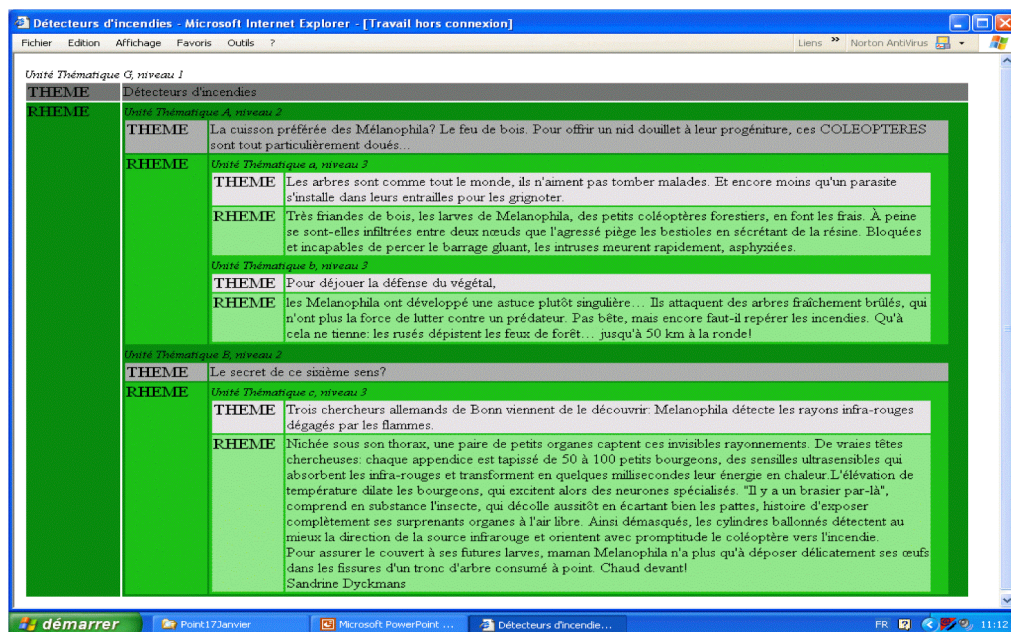


Figure 3.10. Segmentation et hiérarchisation d'un article de vulgarisation par THEMA

3.2. Traitements multilingues

UniThem, Analyseur thématique d'articles de presse sous Unicode, développé avec Emmanuel Giguët, traite de la hiérarchie expositive d'articles de presse en langues variées mais sans proposer de mots-clés (Lucas & Giguët, 2005). Dans ce logiciel d'étude, j'ai testé la possibilité d'abandonner les indices morphologiques, à l'exception des ponctuations, et de calculer les structures de listes par la stylométrie différentielle.

L'algorithme exprime les propriétés du modèle asymétrique et partiellement récursif sans vérification morphologique. L'analyse part de l'article entier et le subdivise. Le calcul tient compte uniquement de la constance ou de l'inconstance des effectifs de sous unités dans une unité (et non simplement de l'effectif avec des valeurs seuils comme dans les approches statistiques courantes).

Les bornes de constructions sont des unités typographiques ayant un cardinal d'éléments nettement plus petit ou plus grand que la moyenne. Les ponctuations rares ou hapax servent au niveau de granularité suivant (ou à l'échelon suivant).

UniThem exploite Unicode et a été testé dans les langues européennes à écritures latine, grecque et cyrillique et d'autres familles de langues non alphabétiques. Les tests ont été menés sur 14 idiomes : français, italien, espagnol, portugais, roumain, grec, suédois, norvégien, polonais, russe, chinois, japonais, coréen et arabe. D'autres idiomes peuvent être traités et l'ont été (bulgare, tchèque, finnois), nous ne mentionnons que ceux pour lesquels nous avons des locuteurs natifs disponibles pour évaluer la qualité des sorties.

Dans les copies d'écran, le bandeau *Structure thématique développée* indique que le texte entier est reproduit (quand les articles sont courts), tandis que *Structure thématique compacte* indique que le texte est évidé et non reproduit intégralement. Le texte de milieu est remplacé par [...].

La Figure 11 montre la structuration d'un exposé simple (voir exemple 3 p. 44-46).

STRUCTURE THEMATIQUE COMPACTE

Unité Thématique G , niveau 1

THEME Des traces de ricine dans des flacons gare de Lyon à Paris

RHEME Unité Thématique G1 , niveau 2

THEME PARIS (AFP),

RHEME le 21-03-2003 [...] indique le ministère.

Unité Thématique G2 , niveau 2

THEME "Les analyses effectuées ont permis de constater que les deux derniers flacons contenaient des traces de ricine dans un mélange qui s'est révélé être un poison très toxique", [...] la SNCF a fait appel à la police.

RHEME Unité Thématique G21 , niveau 3

THEME Celle-ci a dans un premier temps pensé à du charbon (anthrax),

RHEME avant que les produits ne soient confiés à un laboratoire spécialisé en Essonne, [...] tous les contenus des récipients n'ayant pas encore été identifiés.

Unité Thématique G22 , niveau 3

THEME C'est la première fois qu'est rendue publique en France la découverte d'une trace de ricine. Un tel produit avait déjà été retrouvé à Londres chez des islamistes soupçonnés de préparer des attentats.

RHEME L'enquête a été confiée à la section antiterroriste de la Brigade criminelle, [...] indique-t-on encore de même dans l'attente des résultats définitifs des analyses.

Unité Thématique G3 , niveau 2

THEME Un lien avec le déclenchement des hostilités en Irak "n'est pas non plus avéré en l'état des investigations", [...] mais cela est qualifié de "troublant".

RHEME La ricine est la toxine végétale la plus toxique. [...] Elle a été rendue célèbre par les services secrets bulgares et leur fameux parapluie.

Figure 3.11. « Ricine », exemple d'exposé en français hiérarchisé par Unithem

La Figure 12 montre la structuration d'une explication (voir exemple 5 pp. 46-47). Dans la représentation, un trait horizontal sépare l'entrée en matière de l'explication proprement dite.

STRUCTURE THEMATIQUE COMPACTE

Unité Thématique G, niveau 1

THEME Le point chaud de l'Afar sous surveillance**RHEME** Unité Thématique G1, niveau 2**THEME** Près de 90% des volcans naissent en bordure des plaques tectoniques, au niveau des dorsales et des plaques de subduction.**RHEME** Unité Thématique G11, niveau 3**THEME** Mais il existe un deuxième type de volcanisme,**RHEME** beaucoup moins répandu, [...] directeur du Département de sismologie de l'Institut de physique du globe de Paris (IPGP).

Unité Thématique G12, niveau 3

Parviennent-ils tous en surface? [...] régions où se trouve l'un des rares points chauds émergés.

Unité Thématique G2, niveau 2

THEME Organisée dans le cadre du programme " Corne de l'Afrique " de l'Insu, [...] explique Jean-Paul Montagner.**RHEME** Unité Thématique G21, niveau 3**THEME** Ces ondes se propagent plus lentement dans les milieux chauds.**RHEME** En repérant les anomalies de vitesse, [...] les chercheurs parisiens ont sillonné le Yémen à la recherche de zones épargnées par le " bruit culturel " (les vibrations produites par l'activité humaine).

Unité Thématique G22, niveau 3

THEME C'est finalement au nord d'Aden qu'une nouvelle station a été mise en place,**RHEME** venant enrichir le dispositif de surveillance déjà installé dans l'année écoulée — [...] nous devrions être en mesure de fournir une image détaillée du sous-sol de la corne africaine."

Figure 3.12. « Afar » exemple d'explication en français hiérarchisé par Unithem

Les Figures 13 et 14 montrent des résultats sur des textes en italien et russe, à écriture alphabétique. Les Figures 15 et 16 montrent des résultats sur des textes en arabe et japonais, écritures non alphabétiques.

STRUCTURE THEMATIQUE COMPACTE

Unité Thématique G, niveau 1

THEME I Savoia fra tre mesi torneranno in Italia**RHEME** Unité Thématique G1, niveau 2**THEME** Di Franco Chirico [...] la biografia del Re di Maggio:**RHEME** Unité Thématique G11, niveau 3**THEME** Umberto, re senza corone**RHEME** ROMA - Il Parlamento abolisce l'esilio perpetuo dei discendenti maschi di Casa Savoia. [...] Come si vede quindi movimenti in bilico fra nostalgia e archeologia politica che difficilmente hanno da soli i mezzi per mobilitare le masse.

Unité Thématique G2, niveau 2

THEME Fra gli interventi che hanno preceduto il definitivo voto parlamentare di oggi (che non è avvenuto a scrutinio segreto nonostante la richiesta in tal senso di Maura Cossutta) ne vanno comunque segnalati alcuni.**RHEME** Quelli di chi da sinistra (ma lo ha fatto anche il repubblicano Giorgio La Malfa) è tornato a chiedere che i Savoia facciano comunque una pubblica dichiarazione di fedeltà alla Repubblica. Ma anche quelli di chi più prosaicamente ha chiesto loro di rinunciare ai beni nazionali che a suo tempo conservarono anche dopo l'esilio.

Figure 3.13. Exemple d'exposé en italien hiérarchisé par Unithem

<i>Unité Thématique G , niveau 1</i>	
THEME	Колокольный звон над колонией
RHEME	<i>Unité Thématique G1 , niveau 2</i>
THEME	раздастся скоро - здесь будет построена церковь [...] не опускают руки поборники доброты и нравственности.
RHEME	<i>Unité Thématique G11 , niveau 3</i>
THEME	Беспрецедентное событие не только для Зеленограда,
RHEME	Москвы, [...] которая станет Подворьем Данилова мужского монастыря Москвы.
	<i>Unité Thématique G12 , niveau 3</i>
THEME	Церемонию закладки в фундамент церкви капсулы с грамотой на освящение проводил архимандрит Алексей, [...] взявшие на себя обязательства по финансированию строительства храма.
RHEME	Перед Богом все равны, [...] истинное.
	<i>Unité Thématique G2 , niveau 2</i>
THEME	- Прихожане здесь отличаются от обычных искренностью, [...] стал священнослужителем и ведет приход в Подмоскowie.
RHEME	<i>Unité Thématique G21 , niveau 3</i>
THEME	...Когда наступил момент торжественной службы, [...] что права гражданина соблюдаются.
RHEME	Церкви есть практически в каждой колонии, [...] Л.ПОМАНОВА

Figure 3.14. Exemple d'exposé en russe hiérarchisé par Unithem

STRUCTURE THEMATIQUE DEVELOPEE	
<i>Unité Thématique G , niveau 1</i>	
THEME	22.5 مليون دولار لمواجهة تشويه صورة العرب
RHEME	<i>Unité Thématique G1 , niveau 2</i>
THEME	وافق وزراء الإعلام العرب في ختام اجتماعاتهم أمس بالقاهرة علي تنفيذ الخطة الإعلامية الموحدة التي اقترحتها مصر
RHEME	لواجهة الحملات الإعلامية الصهيونية لتشويه صورة العرب في الخارج وخصص الوزراء 22.5 مليون دولار لتنفيذ الخطة. وصرح السيد عدنان عمران وزير الإعلام السوري ورئيس الدورة الحالية لمجلس وزراء الإعلام العرب بأنه تم اعتماد الاستراتيجية الإعلامية العربية وميثاق العمل الإعلامي العربي وقال إن المجلس وافق أيضا علي إنشاء مرصد عربي في أوروبا والولايات المتحدة لتجميع المواد الإعلامية التي تتناول صورة العرب في الخارج وإعداد الردود اللازمة عليها.
	<i>Unité Thématique G2 , niveau 2</i>
THEME	وقد بعث الوزراء برقية شكر للرئيس حسني مبارك لدعمه العمل العربي المشترك
RHEME	وهناؤا الإعلام المصري بافتتاح مدينة الإنتاج الإعلامي.

Figure 3.15. Exemple d'exposé en arabe hiérarchisé par Unithem

Les résultats évalués par des locuteurs des langues considérées ont montré que les présupposés sous-tendant l'analyse automatique étaient en effet valides pour le genre journalistique dans les langues à écriture alphabétique. Ils sont satisfaisants aussi pour le japonais, à écriture mixte

syllabique et idéographique. Cependant, les limites du calcul en relation avec la grille d'interprétation sont observées pour les langues à écriture compacte, comme le chinois, purement idéographique, ou l'arabe. Le recours au jugement des lecteurs a montré des défauts d'interprétation, défauts accentués dans les articles longs. La distribution de l'information est différente dans ces langues. Le résultat produit ne tient pas compte des reformulations en fin d'article par exemple. Le modèle devrait pour cela accepter une construction inversée (rhème suivi d'un thème). Cet ajustement est possible si l'on conserve les contraintes de distance d'édition entre chaînes de caractères, dites contraintes morphologiques. Les articles souffrent aussi d'une mauvaise résolution en arabe, dans le sens où les unités manipulées mot et paragraphe ne sont pas les plus informatives. Le calcul sur les caractères et les phrases serait plus approprié.

STRUCTURE THEMATIQUE COMPACTE

<i>Unité Thématique G, niveau 1</i>	
THEME	必要資金額の調達は困難 仏口は拠出表明せず イラク復興支援会議開幕
RHEME	<i>Unité Thématique G1, niveau 2</i>
THEME	【マドリード 23 日共同】イラク復興支援会議が二十三日、[...] 欧州諸国など約七十カ国・機関が参加してスペインのマドリードで開幕した。
RHEME	<i>Unité Thématique G11, niveau 3</i>
THEME	世界銀行が試算した五百五十億ドル（約六兆円）の必要資金に対し、
RHEME	米国の要請を受けた各国がどの程度の資金拠出を表明するかが焦点。[...] 必要資金の調達は困難な状況だ。
	<i>Unité Thématique G12, niveau 3</i>
THEME	米国主導のイラク戦争と戦後統治への各国の政治姿勢は、
RHEME	資金協力への対応に大きく反映している。
	<i>Unité Thématique G2, niveau 2</i>

Figure 3.16. Exemple d'exposé en japonais hiérarchisé par Unithem

Puisque le logiciel a été construit en fonction du genre journalistique, ses limites ont été éprouvées sur des articles académiques, des forums et des textes littéraires. Les résultats ne sont pas aussi mauvais qu'escompté, le modèle est donc encore sous-spécifié en genre. On pourrait probablement gagner en qualité en resserrant les contraintes du style collectif journalistique. J'ai choisi de spécifier davantage les contraintes pour les forums, comme on le verra dans le chapitre suivant qui aborde les questions d'échelle et de traitement de la toile.

3.3. Des critères différents par niveau

Pour conclure provisoirement cette question, je voudrais souligner que dans l'approche que je préconise, les conditions pour le calcul de patrons à un niveau en grammaire de texte ne sont pas dépendantes des niveaux adjacents, ou du moins jamais entièrement. Il n'y a pas d'homothétie récursive dans les constructions. Ce qui est visible à un grain ne l'est plus au niveau suivant. Ainsi, une citation à verbe variant est manquée lors de la détection de citations individuelles en anglais dans le paragraphe (exemple 5). Cette citation au discours indirect n'a pas de locuteur immédiatement identifiable. Mais elle est trouvée et comprise dans une chaîne co-référentielle de citations, et reliée à l'informant comme on pourra le voir dans l'exemple 6 proposant l'article entier.

Exemple 3.5. ADJ 38 Citation hors contexte sans indice morphologique *said*

A municipal analyst also agreed that low absolute rate levels have dimmed retail interest, but added that some monies have shifted out of emerging market funds and stock funds and into governments.

Exemple 3.6. ADJ 38 Chaînes de citation dans le contexte de la dépêche

11/12/97 USA: RETAIL INTEREST IN MUNIS HO-HUM DESPITE STOCK WOES.

By Kathie O'Donnell

NEW YORK, Nov 12 (Reuters) - While long-term municipals are cheap versus U.S. Treasuries, retail investors remain largely faithful to stocks and unimpressed by bonds' skimpy absolute yields, traders said.

"Retail has to get burned before they get their feet out of equities, 700 to 1,000 points might do it, 554 didn't," a muni trader said, [...]

The trader added that if the muni market does see more retail participation in coming days, the increased activity may have more to do with Nov 1, Dec 1 and Jan 1 redemptions. Those dates are "just huge coupon payment dates as well as call dates," he said.

"I think you'll see money from that, which is just a rollover, but I'm not so sure new money comes in," the trader said.

A second trader said she has not seen any pickup in retail activity to speak of.

"Stocks maybe off, but people are still glued to the screen," the trader said. "I can't say that we've seen much of a follow through on it."

While municipals are cheap to governments, buyers are looking for more yield.

"The cheapest thing you have are New York Cities, and where are they trading?...5.60," she said. "People are still looking for that magic six."

A municipal analyst also agreed that low absolute rate levels have dimmed retail interest, but added that some monies have shifted out of emerging market funds and stock funds and into governments.

"At least it's a positive step that people have re-allocated some cash back into fixed income I guess," the analyst said.

But while most traders reported either no discernible pickup or a modest increase from low levels, Michael Appelbaum, a senior vice president in investments at PaineWebber Inc cited "a great deal" of retail interest in municipals since the stock market volatility began.

"With all the volatility going on in the stock market, I think a lot of people are looking to bonds," Appelbaum said. "And, it depends on your tax bracket, but municipals are real cheap."

If an investor has a triple-A tax-free municipal bond yielding 5-1/4 percent, the taxable equivalent is roughly 8.60 percent, he said. That makes for a nice return, considering that 30-year government bonds were yielding 6.15 percent at mid-afternoon Wednesday, Appelbaum said.

Ce résultat est la conséquence du choix de critère fiables et fortement contraints par l'espace de recherche. On voit aussi par là que la granularité du traitement est dépendante de la densité d'information propre au style de l'auteur, ce que j'appelle aussi la résolution offerte par le texte passant dans un outil de traitement. Le passage cité en 5 n'a pas vraiment de sens hors du contexte de l'article, en dehors de l'espace de lecture dans lequel il est présenté *de facto* à l'utilisateur.

Les logiciels d'étude sont destinés à mettre à l'épreuve des hypothèses et n'ont pas de prétention à l'excellence ni d'objectif applicatif immédiat. En réalité, bon nombre de ces logiciels sont ou ont été exploités. Certains ont été mis gratuitement à la disposition des utilisateurs, d'autres ont eu un débouché commercial.

Pour clore ce chapitre, je voudrais revenir sur la notion d'« algorithme maître » qui dirige le traitement au niveau général. Un « algorithme maître » diffère d'une « architecture » en ce que l'architecture est ordinairement conçue comme le cadre permettant de relier des modules, elle évoque l'aspect contrôlé d'une approche compositionnelle. L'algorithme maître est un processus de résolution, permettant l'émergence éphémère ou non de certains opérandes qui ne sont pas tous prévus au départ. J'entends par opérandes des tagmèmes (formes à une position), par exemple *l'Agriculture des Quinze* dans le 1^{er} paragraphe de l'exemple 3.3. p. 58. Seules les formes sont

considérées ordinairement et retenues comme « candidates » à un étiquetage dans le métalangage commun. Dans mon approche, les formes sont simplement examinées en contexte, l'étiquetage ne concerne que des relations, à l'intérieur d'une construction (dans un référentiel) et non les termes de la relation.

On voit bien aussi que les calculs ne sont pas calqués sur une reconnaissance humaine à tout moment des chaînes manipulées, par exemple des noms propres et des « candidats locuteurs ». Je ne procède pas par reconnaissance cumulative du locuteur puis du relateur puis du discours rapporté, mais je décide si oui ou non je peux vérifier sur *l'ensemble de la dépêche* mon hypothèse qu'il y a une différence entre le discours du journaliste et le discours d'autrui. De la même manière, je vérifie sur *l'ensemble de la dépêche* mon hypothèse qu'il y a un thème principal avec des reprises.

Il y a donc une certaine liberté par rapport à la description d'items, liberté qui doit être tempérée par la connaissance des limites du système, la parfaite maîtrise des espaces de recherche qui permettent de compléter des séries ou comme je le dis habituellement d'affecter des valeurs à des instances.

Pendant de longues années, le domaine de l'apprentissage a été disjoint de celui de l'analyse, mais on constate que des progrès importants sont advenus grâce à la convergence de réflexions sur les conditions de propagation de contraintes, la notion d'induction, la réflexion sur les propriétés des systèmes et des inférences permises (Ganascia, 1993 ; 1998 ; 2001). C'est ce que nous allons maintenant évoquer à travers les travaux menés sur la toile ou dans les bibliothèques numériques, à une autre échelle que le texte pris individuellement. Il devient évident alors que le programmeur doit renoncer à son désir de contrôle à tout moment sur les objets manipulés, et admettre une certaine forme d'« émergence » de propriétés générées par le système informatique.

Chapitre 4

La gestion de l'échelle dans les applications

La gestion de l'échelle est très liée à la notion de bonne résolution pour mener l'analyse automatique d'un texte. Mais elle intervient aussi pour le traitement d'un sous-ensemble de textes dans une collection, ce qui est le cas en fouille d'Internet et en fouille de textes. Je présente ces travaux en commençant par une application avec amorce (*bootstrap*) pour montrer comment on se détache progressivement du support graphique et de sa valeur immédiate. On s'appuie sur du connu pour appréhender de l'inconnu repéré automatiquement.

1. L'APPROCHE INDUCTIVE EN EXTRACTION D'INFORMATION SUR LA TOILE

La problématique de l'adaptation est traitée en recherche d'information tout spécialement en fouille de textes ou *web mining*. Le terme d'induction²⁰ est utilisé par certains auteurs pour capter ce processus (Kushmerick, 1997, 2000 ; Muslea, 2001, 2002, 2003). Il est aussi utilisé en analyse syntaxique (Nilsson *et al.*, 2006). Dans des contextes différents, d'autres auteurs parlent de correction ou encore de reconfiguration adaptative en insistant davantage sur l'idée de conservation de l'hypothèse servant au raisonnement (Yangarber, 2005).

²⁰ Ce terme est employé pour indiquer qu'un grand nombre d'exemples sert à dériver une interprétation. En algorithmique on parle d'abduction. Les définitions ne sont pas établies de manière très rigoureuse, et correspondent souvent davantage au résultat perçu qu'aux procédures informatiques. Souvent l'induction recouvre un raisonnement abductif de la part du programmeur.

Comme indiqué, les logiciels réalisés par des étudiants sous ma direction sont des automates déterministes. Cette appellation évoque la déduction. L'hypothèse, formulée humainement et transcrite comme relation ordonnée, sert de base de calcul dans les traitements automatiques. À l'intérieur de ce modèle, il faut cependant admettre certaines variantes ou valeurs manquantes, pour pouvoir compléter un patron ou revenir sur un point de décision.

1.1. Etape d'acquisition de patrons

Dans le contexte du filtrage d'information multilingue, nous avons réalisé un système basé sur ces principes (Stienne & Lucas, 2003). Le système d'interrogation en ligne Cinéphile offre des informations sur les sorties de films, et recherche les titres, les dates, les noms de réalisateurs et les noms d'acteurs sur des sites de cinéma. Il a travaillé sur des langues européennes. Le jeu de déductions (prévu par le programmeur) est fondé sur des postulats propres à la tâche. L'hypothèse est l'uniformité du « thème » ou centre d'intérêt, le cinéma, et sur une périodicité régulière du renouvellement de l'information recherchée (qui deviendra une constante dans le calcul).

Sachant que les sites de cinéma offrent des informations qui fournissent la matière des interrogations possibles, nous produisons automatiquement un lexique de noms propres associé à un lexique de noms de métiers du cinéma, au lieu de l'injecter à partir d'un travail de recensement manuel. Dans cette expérience, il y a alternance de processus de déduction et d'induction ou abduction, vus du côté du programmeur (Morand, 2004). Ainsi le système semble « acquérir des connaissances » sur un certain nombre de sites de cinéma, et les ordonne pour traiter de nouveaux sites, dans une langue différente. Bien sûr l'entité « film » change de nom, les titres étant ordinairement traduits. Le principe de calcul est de n'attribuer de valeur à une nouvelle forme que si la structure d'ensemble est suffisamment renseignée pour que l'extrapolation soit permise.



Site cinéphile européen

titres du film dans différents pays

- **Swimming Pool** (titre en France)
- 8 donne e un mistero (titre en Italie)
- Swimming Pool (titre en Belgique)

différents articles

- [Supereva](#) (italien)
- [Yahoo](#) (français)
- [Cinebel](#) (français)
- [Chronic'Art](#) (français)
- [Chronic'Art](#) (français)
- [Chronic'Art](#) (français)
- [Chronic'Art](#) (français)
- [Allociné](#) (français)
- [L'Express](#) (français)
- [L'Express](#) (français)
- [Monsieur Cinéma](#) (français)
- [Chronic'Art](#) (français)
- [Chronic'Art](#) (français)
- [Chronic'Art](#) (français)
- [Chronic'Art](#) (français)
- [L'Express](#) (français)
- [L'Express](#) (français)
- [Chronic'Art](#) (français)

recherche [film](#)

recherche [acteur](#)

recherche [libre](#)

[films](#) de la semaine

[sites de référence](#)

[données disponibles](#)

[à savoir](#)

réalisateur

- [François Ozon](#)

quelques acteurs

- [Charlotte Rampling](#)
- [Ludvine Sagnier](#)
- [Charles Dance](#)
- [Marc Fayolle](#)
- [Jean-Marie Lamour](#)
- [M. Mossé](#)

Figure 4.1. Interface de *Cinéphile* montrant la synthèse de l'entité film

1.2. Acquisition de nouvelles données

Rappelons que le postulat est que dans les sites visités, on parle de films, et que le titre, les noms de réalisateurs et d'acteurs sont donnés ensemble, ce qui forme une « entité film », une structure à géométrie variable. La présentation du film est localisée en tête de page. D'abord, l'induction (humaine) porte sur le repérage des segments informatifs, la présentation des films dans des sites de presse. Le dépouillement des sites est mené par la détection de frontières ou *wrappers*, encadrant l'information recherchée, sur le modèle de Kushmerick (1997, 2000). Des patrons définissant l'ordre des éléments recherchés sont fournis comme base de calcul en machine. « L'induction » ou « acquisition » (automatique) joue quand le patron prévu est remplacé par des variantes trouvées en grand nombre dans le même contexte. On appelle souvent ce processus acquisition de « connaissances ». Le système traite de patrons avec des valeurs manquantes (comme pour la détection de citation), ou bien propose des patrons formés à partir d'un seul indice avec des voisins jugés équivalents à ceux d'origine.

Quand on change de site, l'adaptation est semi-supervisée (active) sur le modèle de Muslea (2001, 2002, 2003), ce qui consiste à extrapoler ou propager certaines valeurs. D'un site à l'autre on retrouve les mêmes noms propres de personne. On stocke en mémoire ce qui est trouvé à la position prévue par le patron, et un nouveau patron est appliqué avec les formes les plus fréquentes dans le site traité (la meilleure configuration produit un patron reflétant l'ordre de présentation adopté dans le site). Tant que les résultats sont jugés bons, on n'arrête pas le processus d'acquisition. Lorsque le système extrapole à tort, on interrompt le processus. On peut mettre des interdictions sur les configurations fautives à la manière de Yangarber (2003) ou bien contraindre la propagation en ajoutant des conditions, ce que nous avons fait. Par exemple, lorsque le système extrapole qu'un ou deux noms capitalisés suivant un titre sont des noms de réalisateurs, Victor Hugo est traité comme réalisateur dans « *Les Misérables* » d'après l'œuvre de Victor Hugo. On peut annoter négativement le segment fautif ou ajouter une condition (absence de *selon/d'après* dans le segment ponctué).

Ihre freie Frage: **Darsteller Berry**

- [Richard Berry](#)
- [Halle Berry](#)
- [Joséphine Berry](#)
- [The Wild Thornberys](#)
- [Den ville familien Thornberry](#)

Cette méthode d'interprétation travaille sur la recherche de film

[Le Petit prince a dit](#) Darsteller Richard Berry
[Ah ! Si j'étais riche](#) Darsteller Richard Berry
[Ah, si j'étais riche!](#) Darsteller Richard Berry

Figure 4.2. Interface allemande, interrogation sur un acteur « Berry »

L'interface permet d'interroger le système sur des films, des réalisateurs, des acteurs, dans l'une des langues acquises. La première partie de la copie d'écran sollicite le choix d'un nom, si l'on choisit Richard Berry, la seconde s'affiche et propose trois titres de film. Nous avons soigneusement établi le diagnostic de tâche, à savoir le ciblage de l'information recherchée, et l'examen des critères repérables informatiquement pour initier la détection. J'ai procédé par transposition des indices pour établir les patrons initiaux. Cela veut dire que le système traite des *relations* ordonnées dans un espace de recherche et les exploite pour traiter des formes également reliées, éventuellement dans un ordre différent, dans le même espace de recherche. Ceux-ci sont le site, la page, le paragraphe, la phrase et le virgule (espace ponctué par une virgule). Lorsque l'on change de site, le patron évolue, car les formes sont différentes. Mais elles sont jugées équivalentes si les relations sont constantes dans l'espace de recherche considéré. Le système a ainsi une fonction de rétroaction non supervisée, il effectue la partie calcul d'une déduction, ce qui s'est révélé être un pari intéressant.

L'extrapolation est évidemment très risquée si elle n'est pas fortement contrainte, ce que nous avons réalisé à travers un double contrôle de la sélection (l'espace dans lequel une relation peut être établie). Le double contrôle s'exerce ainsi à deux niveaux : sur la sélection du site entier, à travers les liens, spécialement le lien allant de l'index à la page du film et, dans cette page, sur une sélection restreinte en tête de page, la zone descriptive du film.

Les résultats ne sont bien sûr pas parfaits, comme on le voit à la Figure 11, deux titres ont été conservés pour un seul film, à cause de leur graphie différente et du fait que les acteurs cités ne sont pas les mêmes. Le logiciel ayant tourné quelques mois, les résultats sont évalués par rapport à une vérification manuelle sur un échantillon de 207 films dans le tableau 5.

Tableau 4.1. Evaluation de la pertinence de la base *Cinéphile*

	Titres de film	Réalisateur(s)	Acteurs
Ensemble de référence P	207	130	403
Total enregistré B	189	154	306
Correctement détecté C	165	110	289
Sur-généré	24	35	17
Non détecté	66	20	140
Autres erreurs (inversions...)		9	9
Rappel C/P	0,8	0,85	0,72
Précision C/B	0,87	0,71	0,94
Silence (1- rappel)	0,20	0,15	0,28
Bruit (1- précision)	0,12	0,29	0,06

En termes linguistiques, les constantes du genre (journalistique) sont exploitées, pour permettre une mémorisation éphémère (en mémoire de travail) des critères de forme et d'ordre local propres à un idiome, c'est-à-dire en pratique une collection de sites, dans un contexte borné. Du point de vue du génie logiciel, l'externalisation du contrôle d'induction est assurée par l'opérateur de

raffinement. C'est ce qui permet de vérifier la cohérence des critères retenus pour satisfaire la requête qui porte sur des relations.

Les sources d'erreur tiennent à une hypothèse abusive, qui est qu'une chronique de cinéma sur une page renvoie à un film et un seul. Les rétrospectives échappent à ce postulat et expliquent en grande partie le silence. Par ailleurs, des cas particuliers comme les films animaliers (sans noms d'acteurs) mettent le processus en difficulté. Mais ces résultats sont plus qu'honorables pour un travail réalisé par un étudiant de DESS.

Les idées force qui résultent de ces travaux sont les suivantes :

- la caractérisation des informations recherchées ensemble est décisive ;
- la description exhaustive de toutes les formes est inutile ;
- la représentation des espaces de recherche est capitale ;
- l'induction au niveau de l'interprétation doit être sévèrement contrainte : elle ne peut porter que sur un critère à la fois.

Par la suite, plusieurs applications ont mis en œuvre tout ou partie de ces enseignements.

2. LA FOUILLE DE COLLECTIONS DE TEXTES

La fouille de données textuelles s'est développée récemment, depuis le tournant du siècle, avec une relative stabilisation des pratiques dans un cadre applicatif qui est celui de la recherche et de l'extraction d'information dans les collections d'articles de biologie et médecine (Hunter & Bretonnel Cohen, 2006). Devant la masse considérable des archives spécialisées en génomique, la recherche de liens entre les données stockées dans des bases de données sur le génome et les articles interprétant ces données fait même l'objet d'un sous-domaine appelé « la biologie *in silico* » (Ananiadou *et al.*, 2006). A partir de ce foyer initial, la fouille de textes s'est étendue à d'autres domaines : la recherche d'information dans les brevets ou dans les rapports financiers, entres autres.

Les méthodes purement statistiques, utilisant l'ensemble des mots d'un texte pour le décrire, se sont complexifiées. Dans la problématique de la fouille de textes, comme dans la fouille de données chiffrées, des systèmes de caractérisation sont utilisés en amont des classifieurs, dans le but de mieux caractériser des collections d'articles.

2.1. Les descripteurs multi-niveaux

Le terme « multi-échelle » traduit le terme anglais *multi-scale*, employé pour la description des textes ou des collections à différents grains. Nous lui préférons le terme de multi-niveaux, tant que la notion d'ordre de grandeur n'a pas été éclaircie. Le passage d'un ordre de grandeur à l'autre n'est en effet pas si simple.

Le premier projet associant la description des textes et leur caractérisation par des techniques de fouille de données a porté sur la correction grammaticale et stylistique sur un corpus d'articles académiques en anglais, toutes disciplines confondues (Lucas & Crémilleux, 2004). Une collaboration industrielle a été sollicitée par Jouve en 2000. Il s'agissait d'améliorer la lisibilité des articles après la révision éditoriale et avant l'impression définitive. Une part du corpus est écrite par des anglophones, une autre par des chercheurs de diverses origines. Le travail comprend un volet apprentissage et un volet signalement et correction d'erreurs.

Pour le volet apprentissage, j'ai travaillé en collaboration avec l'équipe DoDoLa du GREYC (co-encadrement de Leny Turmel avec Bruno Crémilleux). Il s'agissait de caractériser les textes en fonction de leur degré de lisibilité, en s'appuyant sur des paires d'articles avant et après correction humaine : en quelque sorte d'apprendre quels articles et quelles portions d'articles sont corrigés. Les textes non corrigés sont considérés comme corrects.

Pour le second volet, signalement et correction d'erreurs, l'équipe Island a engagé un ingénieur en 2001-2002, que j'ai dirigé sur une période de trois mois. La réalisation informatique a porté sur l'alignement des corpus (avant et après correction humaine et avant et après correction automatique). Il est nécessaire en effet de pouvoir montrer à un correcteur humain l'original et les modifications proposées pour qu'il puisse juger de la qualité. Il faut aussi comparer des états (original et corrigé) pour comptabiliser la quantité des corrections apportées du point de vue apprentissage mécanique. Un automate de détection et de correction de fautes a été testé, pour l'emploi des démonstratifs, un autre pour l'accord sujet-verbe. J'ai défini l'algorithme de ces modules, programmés par Nicolas Stienne.

Les techniques de fouille nécessitent la création de matrices (bases de données) représentant des collections de textes et leurs descripteurs (attributs) pour extraire des « connaissances » à partir des collections. J'ai proposé une nouvelle stratégie pour hiérarchiser les descripteurs du texte et tenir compte de l'organisation typo-dispositionnelle dans les paires d'articles. Les descripteurs de texte sont indépendants et annotés pour chaque niveau dans une fenêtre d'observation différente appelée sélection de la marque. Ceci renvoie à l'idée que les espaces de recherche situés au début ou à la fin des segments typo-dispositionnels sont souvent plus faciles à caractériser que le segment in extenso.

Les textes ont été « éclatés » en cinq niveaux distincts, autrement dit, la fouille a été menée indépendamment sur des espaces de recherche ou contextes différents, du plus petit au plus grand : le virgule (espace ponctué par une virgule), la phrase, le paragraphe, la section, la partie. Chaque segment de texte est annoté par le jeu de descripteurs relatif à son niveau. Il y a ainsi 5 matrices à exploiter pour caractériser la collection.

It is interesting to note that the calcium imaging technique used in these experiments allows one to indirectly follow the propagation of a single action potential through a branching nerve terminal, confirming that intermittence of neurotransmitter release, at the level of the single

varicosity, is **not** due to failure of the action potential to invade the secretory terminals (Brock and Cunnane, 1988, 1992).

Figure 4.3. Exemple d'analyse d'un paragraphe au niveau du paragraphe

La figure 3 montre un paragraphe annoté en tant que paragraphe, caractérisé par début et fin et la figure 4 ce même paragraphe, constitué d'une seule phrase, annoté en tant que début au niveau de la partie, enfin la figure 5 ce même passage annoté finement au niveau des virgules.

It is interesting to note **that the** calcium imaging technique used in **these** experiments allows one to indirectly follow the propagation of a single action potential through a branching nerve terminal, confirming that intermittence of neurotransmitter release, at the level of the single varicosity, **is not** due to failure of the action potential to invade the secretory terminals (Brock and Cunnane, 1988, 1992).

Figure 4.4. Exemple d'analyse d'un paragraphe au niveau de la partie

It is interesting to note that the calcium imaging technique used in these experiments allows one to indirectly follow the propagation of a single action potential through a branching nerve terminal, confirming **ing** that intermittence of neurotransmitter release, **at** the level of the single varicosity, **is** not due to failure of the action potential to invade the secretory terminals (Brock and Cunnane, 1988, 1992).

Figure 4.5. Exemple d'analyse d'un paragraphe au niveau du virgule

Tableau 4. 2. Echantillon de descripteurs multi-niveaux

Sélection	Formes remarquables		
Virgules	-ly -ed Its / Their -ing ...	and despite such indeed because ...	: ; "" () ()
Phrases	Idem + majuscule ou point		
Paragraphes	There is In fact As well ...	Déterminants indéfinis Adverbes Pronom Conjonction subordination	Date Référence biblio
Sections		Conjonctions majeures Aspect Voix ...	Chiffres Ordinal
Parties	in spite of for this reason as well as	Connecteurs adverbiaux Personnel / Impersonnel Futur /passé Anaphore Conjonctions <i>wh</i> ...	

Les 18 paires d'articles du corpus initial étaient trop peu nombreuses pour que la caractérisation des articles ou des passages avec et sans faute soit possible. Toutefois, les résultats de cette étude préliminaire ont permis de tester des hypothèses, en particulier de montrer que le principe

d'héritage, reliant les divers niveaux, était indispensable à l'extraction d'associations, ou à l'extraction de motifs émergents, deux techniques complémentaires pour caractériser des textes. Les textes sans faute étaient mieux caractérisés par la méthode des associations fréquentes, et au niveau des virgules. Les textes avec faute étaient mieux caractérisés par la méthode des motifs émergents, au niveau des sections et des paragraphes.

Ces travaux ont rencontré un très bon accueil et ont donné lieu à plusieurs communications (Lucas *et al.*, 2003 ; Turmel *et al.*, 2003) ainsi qu'à un article de revue (Lucas & Crémilleux, 2004).

1.2. Filtrage des collections d'articles

Dans le cadre d'une Action Concertée Incitative sur les masses de données, le projet "Bases de données inductives et Génomique" (Bingo), coordonné par Bruno Crémilleux du GREYC, rassemble le Centre de Génétique moléculaire et cellulaire (CGMC) de Lyon-Villeurbanne ainsi que trois laboratoires d'informatique : outre le GREYC, l'équipe universitaire de recherche informatique de Saint-Étienne (EURISE) et le laboratoire d'Informatique en Image et Systèmes d'information (LIRIS, Lyon). L'application visée est l'exploitation des données sur le génome. Elle comprend un volet sur l'exploitation des données textuelles, à partir de textes médicaux, qui est mené par T. Charnois (équipe DoDoLa du GREYC) pour les aspects lexicaux et par moi pour la pondération de l'importance des articles et des zones de texte sélectionnées. Ce dernier aspect est le plus prospectif et entraîne une recherche des critères de pertinence exposés ci-dessous.

Dans le cadre de Bingo, je me suis intéressée à une demande des utilisateurs, en termes généraux : retrouver des articles qui valident ou invalident une hypothèse h au temps t . Le problème est étudié à partir d'un cas de recherche documentaire faite par les partenaires de Lyon. Les textes traités sont des textes de médecine de la bibliothèque numérique Pub Med Central (17 millions de documents à ce jour). Partant d'un ensemble de textes filtrés par le moteur de recherche EntrezPubMed (environ 10 000 documents), l'on cherche à isoler une dizaine de documents « pertinents » pour la requête.

La catégorisation par sous-genre à l'intérieur du genre académique est une des voies utilisée pour réduire l'espace de recherche. Elle consiste à définir des types d'articles (lettre à l'éditeur, article de synthèse, de recherche, etc). Cette voie a été partiellement mise en œuvre par Nadia Zerida, doctorante que j'ai co-dirigée avec Bruno Crémilleux (octobre 2005 novembre 2008). Les descripteurs de textes que nous avons utilisés sont contextuels et très différents des mots-clefs du domaine généralement exploités (Prince & Roche, 2009). Ils permettent d'appréhender différents niveaux de granularité à l'aide de jeux de matrices.

Les descripteurs de textes structuraux implémentés par N. Zerida ont permis de catégoriser des collections d'articles de médecine en trois classes répondant à des critères de choix des utilisateurs : articles de synthèse, articles de recherche et cliniques. Ces classes ne représentent pas l'ensemble

des sous-genres, mais seulement des types d'articles recherchés. Les descripteurs sont rhétoriques, stylistiques, métriques ; ils sont établis et testés séparément sur des unités repérables typographiquement, depuis les fragments de phrases jusqu'au corps de texte (six niveaux pour représenter la hiérarchie des textes). La méthode de fouille utilisée est celle des motifs émergents. Ces travaux ont donné lieu à des communications (Zerida *et al.*, 2006a, 2006b, 2007).

L'approche multi-niveaux est guidée par la nécessité de filtrer les articles en plusieurs étapes sans avoir à les annoter exhaustivement dès le début. Les articles sont catégorisés en articles de synthèse, articles de recherche et cliniques en annotant les textes jusqu'au niveau des sections seulement. Autrement dit, une sous-collection pertinente pour l'étape ultérieure de fouille (recherche d'hypothèses) est extraite grâce à un prétraitement réduit.

J'ai également travaillé avec Franck Thollard (Université de Saint-Etienne) et Arnaud Soulet (Université de Tours) sur la classification d'articles de médecine. J'ai établi 5 catégories rhétoriques (Argumentation, Explication, Exposé descriptif, Compte-rendu, plus une classe indéfinie sans étiquette). J'ai réalisé trois programmes d'annotation utilisant chacun un jeu de descripteurs différents, lexical, grammatical et discursif. Ces deux derniers jeux de descripteurs nécessitent une analyse automatique sommaire. L'annotation a été réalisée sur un millier d'articles de médecine. Deux expériences de catégorisation ont été tentées, l'une à Saint-Etienne exploitant les séquences, à l'aide de fouille d'automates et de calculs de probabilités sur les séquences, l'autre au GREYC puis à Tours exploitant des contraintes à critères paramétrables sur des motifs émergents (Soulet, 2006). L'exploitation des séquences s'est avérée peu probante et coûteuse. La catégorisation par le solveur Music DFS de Soulet à partir des critères positifs et négatifs que j'ai spécifiés a donné d'excellents résultats (Soulet, 2007). L'utilisation du solveur MicMac a permis de gagner encore en fiabilité (Soulet & Crémilleux, 2008). Elle a permis de définir le meilleur jeu de descripteurs permettant d'assigner une classe avec certitude. Les résultats présentent une forte convergence, très inattendue. Il y a très peu de descripteurs retenus, mais ils sont très discriminants.

Les retombées pratiques de cette expérience sont que l'annotation des textes peut se faire par des moyens peu coûteux, sans analyse (macro)syntaxique approfondie du texte entier. Les descripteurs obtenus à partir des critères lexicaux et grammaticaux observés sur les titres internes des documents sont suffisants pour caractériser une collection. Ces travaux ont été soumis pour communications. L'approche par des descripteurs différents suivant le grain fait l'objet d'un chapitre d'ouvrage (Lucas, 2009).

Les expériences se poursuivent dans le cadre du projet ANR « Bingo2 » : Découverte de connaissances par et pour des requêtes inductives dans des applications en post-génomique (2008-2010). Elles portent sur l'annotation des corpus, en particulier le choix des descripteurs en fonction d'une méthode de caractérisation, dans le lot de travail sur la recherche d'information dans les textes bio-médicaux.

Dans le cadre d'un accord bilatéral avec la république tchèque (PHC Barrande du CNRS) "Heterogeneous Data Fusion for Genomic and Proteomic Knowledge", je poursuivrai également la coopération initiée dans le cadre de Bingo avec Jiří Klema sur la fouille de textes.

3. L'ANALYSE DES FORUMS

Le genre et le registre des textes sont des catégories bien connues en linguistique, mais rarement mentionnées en informatique linguistique, si ce n'est implicitement à travers le filtre des applications et des programmes dédiés. Le registre désigne des catégories stylistiques entre les pôles informel et formel (par exemple, *familier, soutenu, guindé*).

Les forums et autres nouvelles formes de communication écrite (NFCE) qui se développent avec les moyens électroniques posent le problème du registre et du style collectif. Les forums sont souvent des documents pour l'action, des discussions qui sont menées à distance et en accompagnement d'une autre activité, étudier, faire des devoirs, pour les forums d'enseignement, s'informer pour des démarches ou des achats, pour les forums de consommateurs, et jouer pour les forums de jeu (Zacklad, 2004). Certains forums sont des lieux de discussion sans autre but que la détente, l'établissement ou le maintien de liens sociaux. On y observe un type d'écrit informel. Autre particularité, par rapport aux analyses de discours, ils sont polyphoniques.

Les forums ne sont pas traités en tant que tels par les méthodes de TAL. Les analyses classiques sont faites au niveau phrase, et présupposent un style soutenu, au moins que les mots soient correctement orthographiés. Or dans le registre des forums, les limites de phrase ne sont pas fiables, et l'orthographe est peu respectée (spécialement en français qui a une orthographe difficile).

Les forums que j'ai étudiés sont des forums d'enseignement et des forums de jeu. Le type de forum, autrement dit la situation d'échange est plus importante que la langue dans la définition du comportement du discours collectif, donc il sert de référence pour définir les constantes de l'algorithme.

3.1 Les forums d'enseignement

Les forums que j'ai étudiés sont des forums d'enseignement à distance ou en présence, en français puis en anglais et diverses autres langues (espagnol, japonais). Les logiciels ont une couverture plus vaste, puisque je modélise un comportement du discours collectif et que je n'utilise pas de lexique idiomatique défini a priori. Les forums traités automatiquement ont été évalués par des formateurs sur le français, l'anglais, le grec et le vietnamien. Encore une fois, les retours sur expérience sont intéressants quand nous disposons d'un corpus d'utilisateurs et de leur jugement sur le résultat.

Le cadre institutionnel de cette étude est celui d'une équipe de recherche en technologie de l'éducation (ERTé) CALICo (Communautés d'apprentissage en ligne, instrumentation,

collaboration) coordonnée par Eric Bruillard. Calico regroupe des formateurs et chercheurs de 8 laboratoires et IUFM, pour le premier volet (2006-2007). Le renouvellement a été accepté pour 2008-2009 avec 11 partenaires. J'assure l'animation du site de Caen (IUFM, FT R&D et GREYC). Plusieurs membres de l'équipe Island participent à ce projet : Emmanuel Giguët, Luigi Lancieri, Anne Lavallard (doctorante à FT R&D soutenance en juillet 2008), Thibault Roy (doctorant sous la direction de Jacques Vergne et Pierre Beust, thèse soutenue en 2007) et Sylvie Normand (chercheur associé).

Les utilisateurs d'outils d'analyse de forums sont les équipes pédagogiques (enseignants et tuteurs, formateurs). Ils n'ont pas le temps de lire l'intégralité des forums d'une plate-forme et ont besoin de moyens de suivi. La réponse apportée à ce problème consiste en indicateurs graphiques, utiles pour comparer des forums. Ils sont issus de données statistiques sur le nombre de messages, de réponses etc. Une alternative consiste à étudier les besoins des utilisateurs (Kanellos *et al.*, 2007). Pour une synthèse récente, voir (Bratitsis & Dimitracopoulou, 2008).

Pour permettre le suivi des forums d'étudiants, tout en conservant l'accès direct au contenu, j'ai mis au point un algorithme traitant du discours collectif, dont le grain est le forum, le segment fonctionnel un ensemble de messages appelé « mouvement » et l'atome est le message. J'ai réalisé ThemAgora avec Emmanuel Giguët en 2006. C'est un logiciel d'étude innovant, adapté au registre informel des forums, il représente une première au niveau international.

En effet, selon les pédagogues, l'unité phrase n'est pas utile, car elle ne correspond pas aux unités interprétables d'après les diverses théories pédagogiques, qui mettent l'accent sur la compréhension, l'esprit critique ou les facultés de communication des étudiants. Par ailleurs, les analyses classiques au niveau phrase supposent que les mots soient correctement orthographiés, ce qui n'est pas souvent le cas.

L'objectif est de proposer une vue synthétique des forums, pour permettre un diagnostic au temps t, notamment pour juger si la discussion est active ou au contraire languissante, si le tuteur doit intervenir ou non. Au départ conçu pour mettre en valeur l'exposition dans le discours collectif, le logiciel ThemAgora était destiné aux forums orientés par une tâche, spécialement les études de cas ou les devoirs. Il a été ensuite adapté aux forums libres. Les forums analysés sont des forums de petite taille (de moins de 100 messages à 500).

Les exemples ci-contre sont des copies d'écran où le symbole [...] indique soit une élision dans un message, soit une élision de plusieurs messages dans un ensemble. Sur l'écran, le texte entier s'affiche si l'on clique sur ce symbole.

Le choix de l'analyse à gros grain permet de surmonter les difficultés de segmentation à grain fin : la phrase n'est guère isolable typographiquement — mais elle n'est pas non plus une unité sémantique. La procédure utilisée est distributionnelle, elle consiste à détecter des messages

différents servant de séparateurs entre des « moments » du discours. Ceux-ci correspondent à des unités sémantiques regroupant des suites d'échanges. Les principes différentiels permettent de relier la stylométrie à la structure discursive, en fonction de récurrences de marques (la sélection de la marque pour un mouvement correspond à un message).

Unité Thématique G, niveau 1	
Thème	Compte tenu du changement de chef d'établissement, la documentaliste a-t-elle laissé un message ou une description de cette IRD avant l'élaboration des emplois du temps ? [...] Heure d'IRD L'heure d'IRD concerne toutes les 6èmes donc elle se répartit tout le long de la semaine en fonction des emplois du temps des élèves et des collègues (ELSA) ... et du mien : je ne fait que (sic) 30 heures...
Rhème	L'action a été voté, pas le créneau ! [...] En fait c'est la suppression de la lecture-info sur ELSA qui a fait disparaître " par inadvertance " la formation IRD... et c'est là que l'Haddock se vexe !
Unité Thématique G2, niveau 2	
Thème	A la lecture de vos tableaux et analyses, plusieurs axes de réflexion se dégagent IMPORTANCE D'UNE FORMATION A L'IRD: comment faire comprendre, reconnaître sa nécessité? [...] C'est pourquoi les IDD sont une chance de faire des piqûres de rappel et de faire bouger les représentations des collègues.
Rhème	Tout à fait d'accord, je vis à peu près des choses similaires dans mon établissement, en 6ème, les élèves suivent une initiation de septembre aux vacances de la Toussaint, soit seulement 3 séances (nous devons gérer 7 classes, soit 14 groupes à venir au CDI) : au fur et à mesure des années, les heures d'IRD se rétrécissent comme une peau de chagrin ! Les réinvestissements dans les disciplines sont assez limités.
Unité Thématique G21, niveau 3	
Thème	Dans mon collège, les élèves de sixième et de cinquième disposent d'une heure au CDI tous les 15 jours, inscrite à leur emploi du temps. [...] Ces heures au CDI nous permettent d'initier les élèves à la recherche documentaire notamment et demandent un travail de préparation relativement important.
Rhème	Dans mon établissement, l'initiation 6ème est prévue et inscrite au projet d'établissement. [...] Ne faudrait-il pas y ajouter parallèlement une démarche active (au niveau de chaque personne, mais aussi au niveau des institutions, dont les IUFM) pour une reconnaissance de l'importance de cette IRD !
Unité Thématique G, niveau 1	
Thème	Compte tenu du changement de chef d'établissement, la documentaliste a-t-elle laissé un message ou une description de cette IRD avant l'élaboration des emplois du temps ? [...] Heure d'IRD L'heure d'IRD concerne toutes les 6èmes donc elle se répartit tout le long de la semaine en fonction des emplois du temps des élèves et des collègues (ELSA) ... et du mien : je ne fait que (sic) 30 heures...
Rhème	L'action a été voté, pas le créneau ! [...] En fait c'est la suppression de la lecture-info sur ELSA qui a fait disparaître " par inadvertance " la formation IRD... et c'est là que l'Haddock se vexe !
Unité Thématique G2, niveau 2	
Thème	A la lecture de vos tableaux et analyses, plusieurs axes de réflexion se dégagent IMPORTANCE D'UNE FORMATION A L'IRD: comment faire comprendre, reconnaître sa nécessité? [...] C'est pourquoi les IDD sont une chance de faire des piqûres de rappel et de faire bouger les représentations des collègues.
Rhème	Tout à fait d'accord, je vis à peu près des choses similaires dans mon établissement, en 6ème, les élèves suivent une initiation de septembre aux vacances de la Toussaint, soit seulement 3 séances (nous devons gérer 7 classes, soit 14 groupes à venir au CDI) : au fur et à mesure des années, les heures d'IRD se rétrécissent comme une peau de chagrin ! Les réinvestissements dans les disciplines sont assez limités.
Unité Thématique G21, niveau 3	
Thème	Dans mon collège, les élèves de sixième et de cinquième disposent d'une heure au CDI tous les 15 jours, inscrite à leur emploi du temps. [...] Ces heures au CDI nous permettent d'initier les élèves à la recherche documentaire notamment et demandent un travail de préparation relativement important.
Rhème	Dans mon établissement, l'initiation 6ème est prévue et inscrite au projet d'établissement. [...] Ne faudrait-il pas y ajouter parallèlement une démarche active (au niveau de chaque personne, mais aussi au niveau des institutions, dont les IUFM) pour une reconnaissance de l'importance de cette IRD !

Figure 4.6. Petit forum d'éducation, analysé par ThemAgora, zoom sur la hiérarchie des thèmes

Les résultats permettent de comparer des forums présentés sous une forme compacte tenant sur un écran, et de juger de la forme prise par le discours collectif : peu de moments hiérarchisés en débats successifs indiquent une forte concentration des étudiants, beaucoup de moments distincts avec peu de profondeur indiquent une discussion moins approfondie. Les résultats d'analyse ont été évalués positivement par les formateurs et confrontés à d'autres approches (volumétrie, réseaux sociaux) pour suivre les forums dirigés d'enseignement à distance de l'université de Picardie et les forums d'appui à l'enseignement en présence de l'IUFM de Caen. En effet, les représentations sont significatives : elles coïncident bien avec le type des forums, ceux attachés aux groupes de travail, ou ceux destinés à la discussion libre (voir aussi Figure 4.7).

Les premiers retours des utilisateurs ont permis d'améliorer l'interface, de présenter des vues très compactes réduites aux départs de discussion, ce qui permet de fournir une sorte de « résumé visuel » du forum. Dans un deuxième temps, nous avons traité des forums à la volée (au cours de la

discussion). Ce logiciel est actuellement mis à disposition des membres de l'ERTé Calico à travers une plate-forme d'échange réalisée par Emmanuel Giguet.

ThemAgora a été remanié en 2008 pour traiter de forums non orientés par une tâche, ainsi que des listes de discussion d'enseignants (Agora), et aussi de forums plus importants (de 500 à 5 000 messages), en tenant compte de séparateurs constitués de groupes de messages. De manière assez révélatrice, le logiciel Agora n'a pas été techniquement et nommément séparé de ThemAgora. Le terme de version a été utilisé puis abandonné, parce qu'un outil destiné à analyser des forums reste compris par rapport à son *objet* et non par rapport à sa *méthode* ou sa *visée*.

Des forums d'éducation en anglais ont été analysés pour montrer l'intérêt des méthodes différentielles permettant une couverture plus large des forums d'enseignement en termes de langue. La Figure 4 montre un forum « dense », attaché à la réalisation d'un projet, comparé à un forum de discussion libre en anglais. La Figure 5 détaille les zooms successifs à partir de la représentation compacte du forum libre pour la section G6.

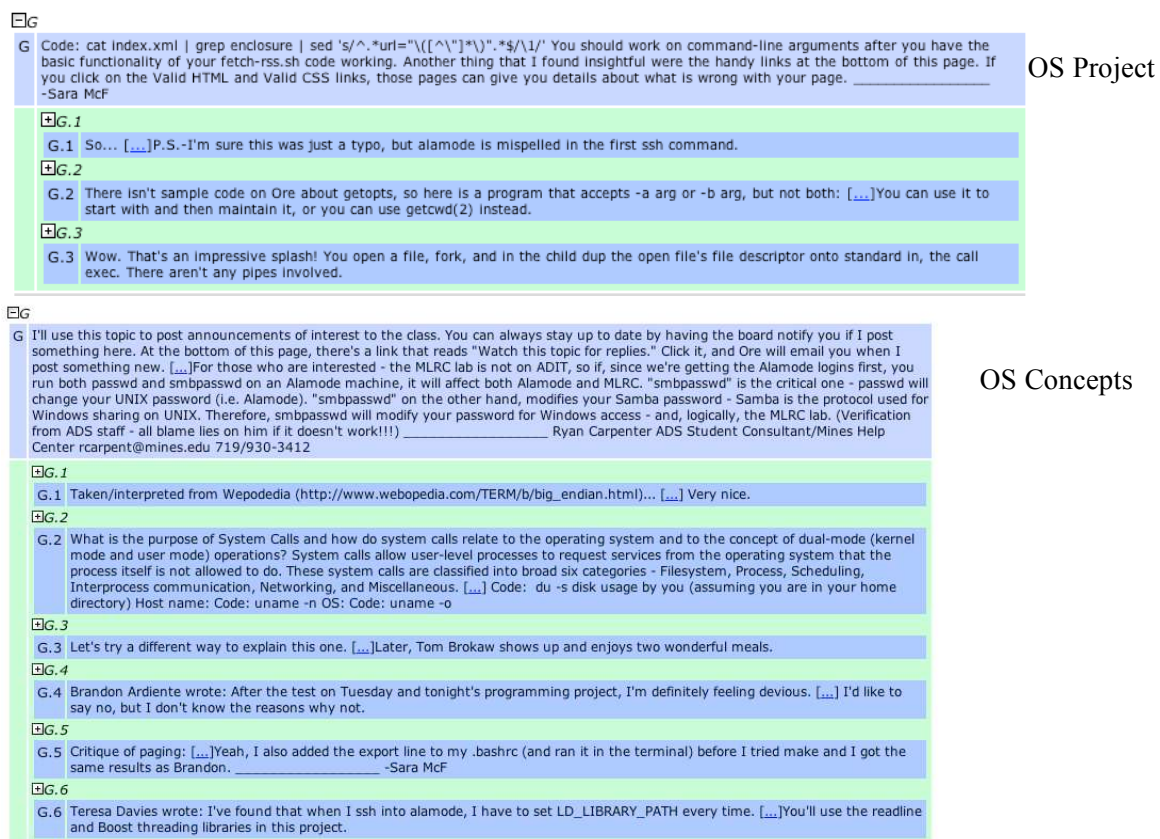


Figure 4.7. Comparaison d'un forum associé à une tâche (OS Project) et d'un forum associé à un cours (OS Concepts)

Après le français et l'anglais, le grec et le vietnamien ont été traités, pour montrer l'intérêt de l'approche discursive stylométrique, une expérience est en cours sur le turc.

Plusieurs communications ont été publiées ou acceptées, certaines au niveau international (Lucas & Giguet, 2008 ; Giguet & Lucas, 2009), ainsi qu'un article de revue (Sidir *et al.*, 2006) et un chapitre d'ouvrage (Sidir & Lucas, 2007). Des forums d'éducation en grec et en vietnamien ont également été traités pour valoriser le choix d'Unicode sur le plan technique et la méthode différentielle sur le plan méthodologique.. Les résultats ont été évalués par des formateurs, très satisfaits d'avoir accès au contenu sans avoir à constituer de ressources lexicographiques ou à annoter des forums manuellement.

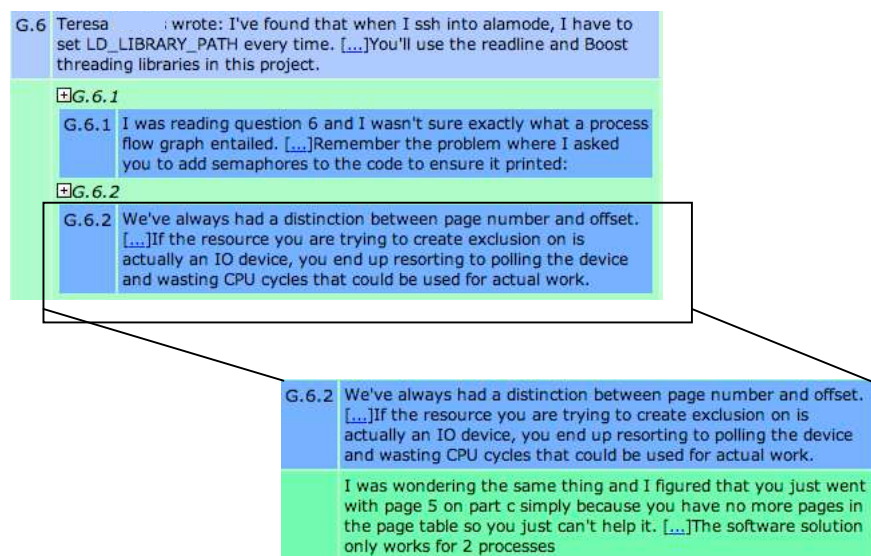


Figure 4.8. OS Concepts, zoom sur le dernier « moment » de discussion et sur un débat

2.2. Les forums de jeu

Des forums de jeu en ligne ont également été analysés en collaboration avec Anne Lavallard. Ce sont de très gros forums (des centaines de milliers de messages) à rythme de croissance rapide et à durée longue. Ces forums posent le problème du passage à l'échelle à tous les niveaux du traitement, depuis le rapatriement des archives jusqu'au choix de la visualisation, en passant par le choix des unités de traitement. Les fils de discussion sont groupés en canaux et ceux-ci en sous-forums. Lavallard propose des outils de visualisation et d'analyse dans un système nommé ForumExplor (Lavallard, 2008).

Agora ne traite encore que des fichiers contenant moins de 10 000 messages, ce qui correspond approximativement dans les gros forums de jeu, à des canaux ou des fils de discussion. Il serait donc nécessaire de traiter de la structure globale des très gros forums, en exploitant les spécificités de ces forums, leur hiérarchie, l'usage des « stickies » ou messages restant en tête des canaux, ainsi que les rythmes d'archivage.

2.3. Conclusion sur les forums

Cette expérience a montré l'intérêt d'intégrer des données relevant de la stylométrie à celles de la structure discursive, en rapport avec le contenu même du forum, pour mettre en valeur la dynamique collective. Cet aspect était d'ailleurs à l'origine du travail dans le cadre de Calico et a été très apprécié. Jusqu'alors la perception et le suivi de la classe virtuelle ou de la promotion étaient quasiment impossibles, le souci de suivre des individus et des relations interpersonnelles dominant les approches et les visualisations.

L'expérience a montré également les limites de la grille d'analyse utilisée, conçue pour des forums dirigés. Quoiqu'elle ait été modifiée dans la version Agora, l'implémentation n'a pas été suffisamment reprise pour rendre la visualisation paramétrable. Cet aspect devrait faire l'objet de recherches ultérieures, pour permettre une adaptation du format de présentation de la sortie. De plus, un processus de diagnostic du type du forum serait nécessaire pour permettre un traitement satisfaisant des forums tout venant.

Le travail sur les forums est cependant innovant et prometteur, il attire des collaborations tant industrielles qu'académiques. Un atelier international est organisé par l'équipe Calico en 2009 dans le cadre de la conférence *Computer Supported Collaborative Learning*.

Sur le plan méthodologique, on comprend que le lexique des forums et autres nouvelles formes de communication écrite étant mouvant et peu normé, si l'on peut les traiter dans un idiome, sans se briser sur l'écueil du sociolecte, alors on peut aussi les traiter dans un autre sociolecte/idiome.

Enfin, le travail sur les forums montre aussi de façon patente que le choix de l'unité de compte et du référentiel n'est pas immédiat et n'est pas non plus indifférent. Ce travail est clairement démarqué d'une part des approches dites sémantiques s'appuyant sur un lexique défini *a priori* ou sur des annotations manuelles, d'autre part des approches purement statistiques. L'approche stylométrique *différentielle* permet d'attribuer une valeur à des formes dont on ne possède pas l'inventaire, mais que l'on peut relever et exploiter avec une grande souplesse, du moment qu'elles entrent dans une construction interprétable du discours collectif. Les notions utiles sont non seulement la récurrence mais aussi la distribution des patrons dans un espace organisé par le modèle (Ganascia, 2001).

Au confluent des cultures de sous-domaines de l'informatique, on se dirige à grands pas vers le paramétrage des critères par l'utilisateur, une idée sous-jacente à certains outils de fouille désormais appelés solveurs, car il ne s'agit plus seulement de classifieurs (Ganascia, 2007 ; Soulet & Crémilleux, 2008 ; Crémilleux *et al.*, 2009). On note ainsi un certain brassage du vocabulaire traversant les frontières de spécialité (Besson *et al.*, 2008).

Chapitre 5

Perspectives et questions ouvertes

Dans les chapitres précédents, j'ai présenté des applications variées : la détection des citations et des locuteurs dans la presse, la hiérarchisation thématique des articles, la caractérisation des collections d'articles de médecine, la segmentation et le compactage des forums. Ces quelques exemples d'analyse textuelle automatisée montrent à la fois des perspectives prometteuses pour l'analyse des textes et des insuffisances, dont je suis peut-être trop consciente, ce qui a pour effet fâcheux une insuffisante publicité des résultats. Pourtant, si les applications paraissent diverses voire dispersées, cela montre aussi que la méthodologie sous-jacente est suffisamment établie dans la pratique et qu'elle est transmissible, puisque la plupart des exemples sont des travaux réalisés par des étudiants.

Les travaux réalisés sont reliés par des « axes de variation » selon Biber : la langue, le genre, le registre (Biber, 1988). J'y ajouterai le style personnel. Ces axes de variation représentent des jalons, pour explorer en largeur le traitement robuste des textes. La couverture bilingue ou multilingue est une conséquence de l'indépendance par rapport au lexique. La variété des genres et des styles est une autre dimension abordée. La relativisation des indices morphologiques est un apport bien intégré dans l'équipe. La prise en compte des sélections (espaces de recherche) et des critères positionnels l'est également, dans la continuité des analyses distributionnelles de Déjean (1998).

L'analyse différentielle se caractérise par le fait qu'on utilise dans l'interprétation le fait que deux segments à comparer sont différents, sans qu'il soit nécessaire de décrire ces segments

intrinsèquement et exhaustivement. Il suffit de préciser en quoi (sur quels critères mesurables) on établit la différence. Un différentiel sert à caractériser des objets linguistiques qui sont en relation de co-variance ou de complémentarité, ce qui va au-delà de la simple caractérisation de la similarité ou non sur une paire. Les enjeux sont bien sûr une plus grande souplesse dans l'affectation des valeurs. La problématique de l'échelle et du choix de la meilleure résolution est mon apport principal, sur lequel je m'étendrai davantage, car c'est aussi l'aspect le plus difficile à modéliser informatiquement.

1. LES ACQUIS SCIENTIFIQUES

La faisabilité des traitements robustes à l'échelle du texte a été démontrée ainsi que leur qualité par rapport aux approches principalement basées sur le lexique et la phrase. La résolution des problèmes de co-référence en particulier est un gain considérable quand on travaille au bon grain, celui où le problème est posé globalement et peut être résolu globalement.

Une démarche hypothético-déductive est à mettre en œuvre en pratique pour identifier les erreurs avec toute la rigueur nécessaire, ce qui n'est pas toujours facile. Le choix de travailler sans ressources externes élimine une incertitude sur les incomplétudes du dictionnaire. Cela permet de vérifier l'adéquation du modèle d'interprétation retenu en relation avec le problème posé. Cela demande aussi une reformulation plus géométrique ou arithmétique des propriétés du modèle relationnel choisi. Il reste à trouver le métalangage mathématique décrivant le comportement des algorithmes qui est le seul garant de crédibilité dans le monde académique, ou à inventer une notation algorithmique consensuelle.

1.1. Des critères morphologiques et positionnels

La différence importante entre les travaux de syntaxe du TAL fondés sur la consultation de dictionnaires et nos réalisations est que les indices exploités sont caractérisés à la fois par leur position relative et leur forme, non pas par leur forme exclusivement. Le problème de l'ambiguïté est évité, car aucune forme n'a de valeur en soi mais seulement dans un patron. Les indices recherchés sont des indices fiables, ils sont peu nombreux.

Ces méthodes faisant place aux positions des indices nécessitent de gérer les espaces de recherche et par conséquent de bien reconnaître la structure typo-dispositionnelle du document. Les techniques de « maillage » des documents, c'est-à-dire la représentation de toutes les positions des segments à tous les grains, mériteraient de plus amples recherches.

J'ai participé en ce sens au projet Résurgence (pôle ITIC Basse-Normandie sur fonds européens) associant le GREYC et la PME Memodata. L'objectif est de recouvrer la structure logique de documents électroniques, notamment des articles académiques. Le format pdf fidèle au rendu visuel est utilisé comme pivot, ce qui permet d'accorder à la disposition toute son importance. Les

méthodes de calcul sont fortement contextualisées, au niveau de l'article et de la page. Elles s'appuient sur la détection des éléments de mise en forme matérielle fréquents et répétés et tiennent compte de leur position relative. Cette approche permet d'améliorer les résultats par rapport à l'approche purement descriptive et énumérative, associant une source éditoriale et une DTD, qui se heurte à l'incomplétude des catalogues et aux fréquents changements de format des revues.

1.2. Des critères positifs et négatifs

Le choix de travailler sur des jeux de relations, en m'appuyant sur des constantes du corpus traité, a permis de résoudre des problèmes jugés extrêmement difficiles, comme celui de la co-référence. Les informations sur l'absence d'une forme recherchée sont exploitées autant que la présence de la forme. Une marque zéro, dans le métalangage des linguistes, est une marque, à condition de pouvoir gérer l'extrapolation de la valeur à partir d'un contexte.

Cette méthode exploite un niveau métalinguistique, la notion de structure interprétable, avec *prééminence des relations sur les éléments* (cf p. 20). Elle nécessite de séparer « l'attendu », le patron ou la structure recherchée d'une part, et la chaîne de caractères traitée d'autre part. L'entrée et la sortie ne sont pas identiques, ni placées sur le même plan. L'entrée est captée par des détecteurs informatiques. L'attendu correspond à une possibilité d'interprétation, il est donc relié à une grille de lecture, un modèle relationnel. Or, dans les pratiques habituelles, le segment d'entrée est aussi le segment de sortie, plus une étiquette. Les options de segmentation, lorsqu'elles sont appliquées en routine, et en début de traitement, bloquent irrémédiablement le processus de pondération d'hypothèses. Si la projection du modèle n'est pas explicite et réservée à la production de la sortie, on observe une confusion entre le relevé des critères utiles et la projection des critères d'interprétation. Une procédure hâtive produit des résultats faux.

La prise de recul sur l'objet à traiter n'est pas possible dans les traitements classiques effectués en déplaçant une petite fenêtre d'observation de mot en mot, étiquetés au fur et à mesure du passage sur la chaîne d'entrée. Il n'y a alors qu'une « dimension », la dimension chronologique, et qu'une opération, la concaténation, comparable à l'addition sur des entiers.

2. EVALUATION

2.1. Les quantifications consensuelles

La question de l'évaluation est épineuse. Dans les concours organisés, la confrontation des résultats suppose une approche commune qui sert de base aux métriques d'évaluation. L'évaluation est intrinsèque, les informaticiens évaluent les systèmes en dehors des conditions d'usage, la plupart du temps en posant qu'il existe une référence, un étalon de qualité fourni par une analyse manuelle sur un petit échantillon. Les points évalués reflètent des préoccupations consensuelles.

Par exemple, les évaluations quantitatives que j'ai faites sur les chaînes de citation n'ont pas été reçues comme convaincantes, car le décompte des citations individuelles donne des résultats excellents. Or, c'est la chaîne qui est traitée et l'exactitude de la détection de chaîne qui devrait être mesurée, ce que je propose. Pourtant, seul le décompte des citations individuelles est ordinairement pratiqué pour comparer les systèmes, donc reconnu. Pour les chaînes, l'exemple 3.8 (p. 64) montre la difficulté à séparer les rôles, un journal et son informateur, un locuteur et l'organisme qui l'emploie. Ce genre d'erreur cristallise l'attention, car le problème est identifié dans la recherche des « points de vue », mais dans la pratique courante, à un grain où il est impossible de déduire rigoureusement.

Dans le cadre de DEFT, le défi européen de fouille de textes, qui est a priori une compétition intéressante, le corpus est proposé sous une forme standardisée, mais impropre à l'analyse du document comme un tout, car la MFM originale n'est pas conservée. Les présupposés sont que l'information extraite correspond à des groupes de mots, et que cette information sera extraite de phrases. Sur ce type de matériau, les méthodes sont le plus souvent probabilistes (Prince & Kodratoff, 2007). L'accent mis sur le découpage en phrases est bien sûr le reflet des pratiques dominantes, mais comme il se fait au détriment de l'unité corps de texte, le prétraitement nous empêche d'exploiter une méthode descendante d'analyse, du corps de texte vers des unités plus petites. De la même façon, une participation à BioCreative a été envisagée, mais le corpus XML était un « sac de phrases » « nettoyé » de toute information sur la typo-disposition, donc, inexploitable par nous car l'annotation multi hiérarchique nécessite le repérage de la hiérarchie typodispositionnelle dans chaque document entrant.

Lorsque les logiciels sont orientés par la tâche et les besoins de l'utilisateur, l'évaluation est faite par les utilisateurs, elle est dite extrinsèque et ne peut servir de base de comparaison. C'est davantage un retour d'expérience. De plus, dans la majorité des applications, les critères retenus par les utilisateurs ne sont pas communiqués. Enfin, dans nombre de cas, la rançon de l'innovation est que les résultats ne sont pas comparables avec des réalisations existantes, encore moins évaluables dans un cadre établi, comme on l'a vu avec les chaînes de citation co-référentielle. En produisant des résultats « scandaleusement bons » mesurés sur les citations individuelles, on se heurte à l'incrédulité.

Mais comme je l'ai signalé, un score de 0.99 en f-mesure sur les citations individuelles ne veut pas dire que le taux de réussite réel sur les chaînes soit bon (p. 65). Va-t-on défendre cette nouvelle unité chaîne ? Quelles sont les mesures utiles pour l'évaluer ? Doit-on continuer à extrapoler des valeurs par rapport aux mesures de précision et rappel utilisées en recherche d'information, qui n'ont plus de réelle signification dans les nouveaux contextes où il n'existe pas naturellement de référence extrinsèque ?

Les utilisateurs sont surtout intéressés par les points de vue — un calcul qui dans mon approche est au-dessus des chaînes — et non par les locuteurs individuels. La proposition d'unités créées pour l'interprétation consensuelle irait plutôt à la détection des points de vue et de la polarité, un besoin exprimé.

Il est souhaitable de publier les règles servant à l'évaluation des logiciels d'étude par les concepteurs, et de susciter des groupes de réflexion sur la meilleure manière de mener l'évaluation sur des traitements à gros grain. Cette tâche à vrai dire ne peut être menée de front avec l'expérimentation. Il est préférable que la conception des logiciels puisse librement progresser et que les questions d'évaluation reviennent à ceux qui ont l'usage de ces logiciels.

2.2. L'éthique

Dernier point à évoquer, l'éthique, curieusement absente de la réflexion collective. Sans aller jusqu'au catastrophisme de certains²¹, il est légitime de se poser des questions sur les applications contraires aux recommandations de la Commission Informatique et Liberté et plus simplement au respect de la démocratie et de la liberté d'opinion.

Y a-t-il nécessité à répondre à des demandes nettement discriminatives comme « trouver les journalistes exprimant une opinion contraire à l'intérêt de [la firme / l'organisme étatique] pour les recadrer » ? L'intitulé de la tâche 3 de DEFT 2009 n'ouvre-t-il pas des perspectives hasardeuses ?

La détermination du parti politique auquel appartient l'orateur de chaque intervention dans le même ensemble de débats au parlement européen que précédemment, fera l'objet de la troisième tâche. Le parti sera à déterminer dans un ensemble fermé de partis européens. (<http://deft09.limsi.fr/>).

3. LES QUESTIONS A APPROFONDIR

Les questions à approfondir sont nombreuses. Le métalangage de l'informatique, discipline récente et encore exhubérante, entraîne d'une part une créativité lexicale forte, une grande polysémie des termes, comme on l'a vu avec « induction » ou « projection » et une grande synonymie, parfois relevée par les informaticiens eux-mêmes (la problématique *self*-* qui rebaptise certains processus des systèmes naguère appelés « intelligents »).

3.1. Questions d'épistémologie

Il semble que la représentation et l'enseignement des processus informatiques soit insuffisante, qu'en quelque sorte, on ne dispose pas encore d'une notation stable et effective dans la transmission des connaissances, notamment en génie logiciel. Le problème des opérations complexes en arithmétique comme en physique n'a pas été résolu en cinquante ans. Il serait curieux que les

21 La littérature de science-fiction est le genre qui porte le débat, par exemple Claude Van Ecken *Le Monde tous droits réservés*, Editions du Belial, 2005 ou Cory Doctorow «Scroogled» *Radar* septembre 2007 traduction française *Engooglés* <http://cfeditions.com/scroogled/>

opérations complexes en informatique le soient en si peu de temps. La réflexion amorcée par Anne Nicolle devrait être reprise et développée.

On sent bien actuellement une explosion du métalangage, autour de la notion d'induction, d'adaptation, d'autonomisation, de contrôle ou de réflexivité et de projection. Quelles sont les opérations communes ou les processus communs qui se cachent sous ces différents termes ?

3.2. Quelques propositions à développer

3.2.1. L'approche descendante

L'approche descendante traduit le souci de *détermination du global sur le local* (cf p. 20-21). Dans la mesure où il y a plusieurs grains d'analyse — et plusieurs sélections dans un grain — cette idée évoque pour certains informaticiens l'analyse multi-niveaux proposée naguère par Vauquois comme moyen de différer les choix en analyse de phrase, quand la décision est incertaine. Elle a été implémentée pour le système de traduction Ariane 78 (Vauquois & Boitet, 1985).

Le principe consistant à maximiser la fiabilité est le même en analyse de texte. Sans doute faut-il différencier ici plus soigneusement le modèle projeté et les étapes d'implémentation. Le modèle projeté stipule que le tout a valeur sémantique est le texte, qui est subdivisé en suivant les indices de structuration typo-dispositionnels et morphologiques et ensuite placé dans une grille de lecture matérialisée par un coloriage.

3.2.2. Le paramétrage du grain

Dans les travaux présentés, le grain d'analyse est pris en compte, ce qui n'est pas courant. Le grain est davantage qu'un niveau, ou même qu'une fenêtre d'observation. Le paragraphe est de plus en plus souvent représenté comme unité d'analyse du texte (Bilhaut, 2006, 2007). Mais souvent cette unité est aussi contraignante que celle de la phrase, aussi mono-dimensionnelle. La démarche est « positiviste » dans le sens où seule la présence de formes remarquables est détectée ou, dit autrement, tout paragraphe doit correspondre à une description répertoriée. Un paragraphe « non marqué » cause de l'embarras. Deuxième problème, l'unité paragraphe n'est pas toujours pertinente pour l'interprétation. L'unité de sens d'un texte dépend à la fois du style collectif (le genre ou le sous-genre) et du style personnel.

Nous avons procédé par étapes pour gagner en souplesse d'analyse. Du point de vue de l'implémentation, la pratique de Jacques Vergne, comme celle d'Emmanuel Giguët, sont enracinées dans une procédure ascendante, à partir des atomes de traitement (les plus petits *tokens*). En règle générale, les difficultés surgissent lorsqu'il s'agit de passer d'un niveau d'analyse au suivant. Le moteur de Linguix 99, à la conception duquel j'ai participé, est encore hybride, dans le sens où il reste dédié principalement à l'analyse de phrase et où il pondère deux niveaux adjacents, chunk et phrase. Certes, il est étendu à l'analyse de textes courts au grain paragraphe, mais sans que l'unité

d'analyse soit complètement déclarative, c'est-à-dire passée comme paramètre dans une fonction informatique. Par exemple, il est impossible avec ce logiciel d'analyser un ouvrage en spécifiant qu'on veut le faire au grain chapitre, avec le paragraphe comme atome.

Linguix ressemble fortement à une procédure multi-niveaux embryonnaire, car deux niveaux adjacents de structuration interagissent, chunk et phrase (dans l'analyse grammaticale), ou phrase et paragraphe (dans l'analyse des textes courts). Il ne s'agit pas d'une cascade d'automates, l'analyse d'un grain g alimentant le suivant, mais bien d'une interaction des critères issus de g et $g-1$ dans le processus d'affectation de valeur.

Cependant, en théorie, une analyse multi-niveaux devrait, comme dans l'idée de Vauquois, être à n niveaux, autant que nécessaire. L'enjeu est d'importance, tant pour l'analyse de textes dans l'objectif du résumé que dans la perspective de la traduction. Une structure relationnelle n'est pas toujours marquée au même grain dans deux langues, comme on s'en est déjà aperçu en opposant les langues « positionnelles » à morphologie pauvre et les langues à morphologie riche. La qualification est faite à grain constant, par consensus dans le cadre d'interprétation de la proposition. Mais une langue à morphologie pauvre au grain g est habituellement riche à $g+1$ ou à $g-1$.

3.3. L'adaptation au style individuel

Pour obtenir le degré de liberté voulu, dès la mise en œuvre du moteur de Linguix, nous avons relevé les cas faciles dans les contextes bien marqués, et appelé procédures « réflexes » celles qui déclenchent une affectation de valeur lorsqu'un patron complet est trouvé. Sinon, une procédure dite « de raffinement » ou procédure différée est appliquée. En tenant compte de l'ordre et des positions « vides », on déclenche un calcul de la valeur manquante lorsque la sélection contient des indices fiables. Lorsque la sélection ne contient pas d'indices en nombre suffisant, on ne statue pas, mais on diffère l'affectation de valeur à un autre grain, en cherchant un patron de plus grande portée dans une nouvelle sélection.

La procédure ascendante de Linguix n'est pas la seule implémentée. Le nombre de grains adjacents à explorer ensemble n'est pas non plus limité à deux. Jusqu'à présent, la procédure descendante, du plus gros grain vers les plus petits, n'est implémentée que dans Thema et ses dérivés. Le gain en abstraction concerne la définition du grain ou plutôt des grains d'analyse d'un corpus. Cependant, l'expérience montre que le degré de liberté n'est pas suffisant. Même à l'intérieur d'un genre, il existe des différences de culture et de style collectif.

Dans le genre journalistique, pour détecter des citations par exemple, une unité d'analyse « groupe de phrases » est appropriée pour le français, dont les paragraphes sont plus longs que les paragraphes anglais. On sait aussi qu'un article de sciences humaines est plus long qu'un article de sciences exactes. Pour gérer ces paramètres de variation, il est souhaitable de faire un diagnostic du corpus texte à texte, ce qui n'est pas réalisé à l'heure actuelle.

Le profilage de textes consiste à attribuer une classe aux textes constituant une collection réputée homogène (Habert *et al.*, 2000). On cherche à établir une relation entre un sous corpus et un traitement. C'est à peu près ce qui est fait actuellement, par exemple en évaluant humainement la taille moyenne des segments constituant sur une collection pour fixer les sélections. Pour objectiver ce savoir-faire, on calcule les chaînes ponctuées et on les ordonne automatiquement par fréquence et longueur, suivant la loi de Zipf (Déjean, 1998 ; Vergne, à paraître).

Le paramétrage de la sélection permet de gérer les différences de taille sur la base du style collectif, mais ne permet pas de tenir compte du style personnel, ce qui est une source d'erreur, statistiquement faible mais irritante. Tant que l'analyse reste dépendante des hypothèses distributionnelles, seules les positions remarquables sont testées, et le moteur de règles ne parvient pas à traiter correctement des textes denses, dans lesquels des patrons reconnaissables existent jusqu'à l'intérieur d'une unité typographique.

Un diagnostic de texte entrant a pour ambition de renseigner l'analyseur sur l'adéquation entre le texte entrant et les opérations visées, ce qui est plus ambitieux qu'un profilage de textes et relève de l'adaptation dynamique, la configuration proactive ou l'apprentissage immédiat, ou encore de la vérification de cohérence d'un système. Pour que les méthodes différentielles soient réellement « endogènes » et robustes, il est nécessaire de prévoir un relevé automatique de la taille et de la densité du texte entrant. Le diagnostic dynamique de texte doit permettre l'adaptation des mesures. Ce diagnostic introduit une relation à trois termes entre l'entrée de l'analyseur, les règles de l'analyseur et les possibilités d'interprétation. On cherche à vérifier si, en vertu des propriétés du modèle et des propriétés du corpus, on a bien une chance raisonnable d'obtenir pour tout texte du corpus une interprétation fiable en sortie.

C'est pourquoi j'envisage un maillage du document, partiellement implémenté. Le maillage consiste à relever systématiquement toutes les positions remarquables dans un texte. Ceci devrait permettre de trouver les délinéaments des structures marquées pour chaque texte, en relâchant les contraintes positionnelles d'amorçage le cas échéant. Le degré de liberté sur le grain d'analyse, par calcul automatique sur la collection et sur chaque texte, est plus nécessaire encore pour permettre l'analyse de textes polyphoniques, d'ouvrages collectifs ou de forums. Dans ces documents, en effet, il est fréquent que les ruptures de style entraînent des erreurs de segmentation et génèrent de l'imprécision dans le résultat.

L'objectif général pour l'interprétation fiable des textes est de calculer automatiquement la résolution optimale pour l'analyse automatique, en fonction de la tâche, du type de document, de la grille d'interprétation projetée et des variantes de style individuel.

Chapitre 6

Perspectives en encadrement de la recherche

L'objectif général pour l'interprétation fiable des textes est de calculer automatiquement la résolution optimale pour l'analyse automatique, en fonction de la tâche et des caractéristiques du document entrant. Pour atteindre cet objectif ambitieux, il est possible de prendre appui sur les réalisations existantes. Les exemples présentés montrent des perspectives suffisamment prometteuses pour que l'effort de transmission soit développé. Le GREYC et l'équipe ISLanD ont participé à la reconnaissance d'une démarche originale, fondée sur une meilleure prise en compte de notions théoriques de la linguistique, le traitement *différentiel* des indices relevés dans les textes. Cette approche s'est avérée utile pour résoudre des questions de génie logiciel, notamment en différenciant la prise de décision dans les analyses. Les décisions consistent à attribuer des valeurs, elles ne sont prises qu'au moment où les indices reliés sont suffisants pour étiqueter un segment fonctionnel. De nouvelles techniques de paramétrage des variables manipulées en sont issues et intégrées dans la plate-forme Wims. Celle-ci est d'une conception novatrice et peu reconnue : les corpus traités sont associés aux programmes utilisés et aux résultats qui en résultent²².

Toutefois, les implémentations avaient pour but de démontrer la faisabilité de l'approche et ne sont pas pérennes : les questions non résolues sont nombreuses, et de surcroît les questions résolues empiriquement n'ont pas fait l'objet d'une évaluation et d'une réflexion suffisamment approfondie pour que les techniques et les algorithmes soient formalisés et reconnus. Il est donc nécessaire de

²² Le travail d'explicitation sur le plan informatique revient à son concepteur, Emmanuel Giguet.

reproduire les expériences pour disposer d'un ensemble d'implémentations comparables et d'en tirer de nouvelles connaissances utiles aux informaticiens.

J'envisage donc de poursuivre mes efforts et de d'encadrer des recherches en informatique linguistique ainsi que sur l'évaluation. Les principaux objectifs importants du point de vue méthodologique, en relation avec l'informatique théorique, sont présentés dans la première section. Ils permettent de dégager des angles d'attaque innovants récapitulés en section 2. De nouvelles applications pourraient ainsi être proposées, comme on le verra en section 3 en particulier dans la recherche et l'extraction d'information, le résumé automatique et la traduction automatique.

1. METHODOLOGIE

Plusieurs expériences d'analyse de textes en diverses langues ont été citées, notamment la méthode à indices pour le repérage très spécifique des citations et des points de vue, l'analyse stylométrique pour les constructions expositives et la méthode inductive pour la recherche d'information sur un sujet défini. Ces méthodes ne sont pas concurrentes mais complémentaires.

Toutefois, de multiples difficultés ont été rencontrées dans la transmission des connaissances. Dans plusieurs cas, j'ai constaté que des logiciels nouveaux étaient confondus avec de « nouvelles versions » écrasant le logiciel précédent comme une amélioration, alors qu'elles étaient conçues comme cumulables, ou « au choix » de l'utilisateur. Ensuite, le savoir-faire est souvent produit, sans que les programmeurs puissent expliquer « comment et pourquoi ça marche ».

D'une manière plus générale, il semble que le manque de recul et de consensus sur des notions essentielles soit pénalisant : en particulier sur la notion d'information sémantique (Floridi, 2003).

1.1. Le défi de l'algorithmique

Il semble nécessaire de poursuivre la recherche en alliant les méthodes à règles linguistiques et la définition d'algorithmes moins coûteux que ceux de l'algorithmique du texte actuelle, la « chaînologie », qui ne tient pas compte des positions dans la hiérarchie du texte. Les techniques de ce que j'ai baptisé maillage mériteraient une approche plus poussée en coopération avec des algorithmiciens.

Il semble important de formaliser davantage ces modes de résolution de problèmes, ce que j'appelle algorithme — improprement, du moins dans le contexte français, tant que les méthodes ne sont pas exprimées sous une forme démontrable. Il est souhaitable de les traduire en termes mathématiques ne nécessitant aucune compréhension des données ou du problème linguistique pour être appliquées. Les formules de calcul seraient ainsi utilisables comme celles de Salton et deviendraient des algorithmes au sens de l'algorithmique théorique. L'exploitation des contraintes des figures et nombres figurés de Pythagore, que j'utilise comme métaphore imagée des processus a déjà fait l'objet de tentatives de généralisation.

La collaboration entreprise avec la fouille de données est importante et mérite d'être approfondie. La notion « d'apprentissage instantané » ou « apprentissage immédiat » semble être assez proche de la notion d'adaptation dynamique et d'auto-contrôle. Les méthodes à règles linguistiques définies a priori sont aussi à utiliser en conjonction avec les méthodes d'acquisition ou de découverte d'indices en corpus pour permettre l'adaptation et établir en retour des procédures de diagnostic de texte de manière plus systématique. Les points plus spécifiques abordés par la suite supposent la mise en œuvre de ces moyens complémentaires.

1.2. Le défi de l'opérateur

Le terme de projection est employé de plusieurs façons différentes et associé à celui d'induction (Hall *et al.*, 2007 ; Nivre, 2008). Le terme d'induction est employé dans une palette d'acceptions très vaste. Les travaux fondés sur une culture philosophique comme ceux de Ganascia (1998b) ou Morand (2004) ne couvrent pas l'usage courant de ce terme dans différents domaines, l'intelligence artificielle, la syntaxe, l'apprentissage ou la fouille de données. La polysémie est la règle.

En pratique, la plupart du temps, les étudiants produisent des diagrammes UML sous forme de schémas de données statiques. Ils ne voient pas l'intérêt de représenter schématiquement des méthodes, des activités, ou de parler des processus, des opérations. Quoique cela existe en informatique, il leur paraît très étrange d'associer des opérations ou des opérateurs informatiques à des données texte, je n'ose plus dire linguistiques ni textuelles.

Cela laisse à penser qu'il reste encore beaucoup à faire pour garder des traces utiles des opérations effectuées en machine, qui sont par nature impermanentes. Le métalangage, les notations de l'informaticien semblent encore assez frustes (Nicolle, 2006 ; Morand, 2008).

2. ANGLES D'ATTAQUE

2.1. Le défi du multilinguisme

Dans la plupart des cas, le multilinguisme est conçu comme une complication, nécessitant une somme de traitements à base de ressources monolingues, et non comme une situation favorable permettant la factorisation du programme de traitement. Dans la mesure où les traitements ne s'appuient pas sur les mots, mais sur un ensemble de contraintes propres à la grille d'interprétation (les *relations* syntaxiques par exemple), il est possible de factoriser les traitements (Hall *et al.*, 2007). La factorisation peut également concerner les relations macrosyntaxiques. Dans le cas des langues latines, par exemple, les macro-indices morphologiques se trouvent à la même position, et l'analyse de textes peut s'appuyer sur les mêmes présupposés d'ordre des arguments.

De façon plus générale, les méthodes que j'envisage devraient permettre d'explorer un nouveau champ d'étude, entre projection et découverte de règles, production et découverte d'indices nouveaux, ce qui est parfois appelé « apprentissage immédiat ».

2.2. La gestion de l'échelle

Les données que je manipule sont traitées à différents grains, et toujours avec différentes sélections ou fenêtres d'observation simultanément, donc différentes mesures. Ceci échappe généralement à l'attention des collègues, qui n'ont pas du tout l'habitude de penser le traitement linguistique autrement que de façon linéaire.

J'ai cité les forums pour évoquer les traitements à très gros grain. Je n'ai pas cité les travaux que j'ai menés sur les ouvrages, car ils n'ont pas abouti à des programmes. Pourtant, il est envisageable de généraliser les principes de granularité variable pour les traitements de type macro syntaxique ou rhétorique. Ils seraient applicables à des textes très structurés comme les manuels scolaires, les ouvrages académiques, mais aussi les ouvrages de vulgarisation, les rapports techniques.

Cependant, il est nécessaire dans les applications ambitieuses comme le résumé ou la traduction, de choisir au cas par cas le référentiel qui semble le mieux adapté en fonction du document entrant. Il y a alors réglage ou paramétrage du moteur ou choix de la grille de lecture dirigé par le texte (au singulier), pour tenir compte de variantes stylistiques à partir de la géométrie des marques relevées. Ceci relève du réglage auto-contrôlé du système ou du *diagnostic proactif dynamique*.

Pour aboutir à un réel diagnostic multidimensionnel de corpus ou de texte entrant, il est souhaitable de définir un maillage de corpus, en examinant systématiquement les positions remarquables sur tous les segments typo-dispositionnels pour repérer automatiquement à quels niveaux le marquage morphologique est le plus exploitable. L'objectif est d'adapter le traitement en situation de constance du genre avec un corpus hétérogène en langue, ou monolingue avec un corpus hétérogène en genre. Pour poursuivre la métaphore médicale, cela reviendrait à adapter le traitement à la physiologie, au poids et aux antécédents du patient.

2.3. Rapport texte et illustration

Les trois axes d'adaptation concernent la tâche elle-même (interprétation de documents, aide à la lecture ou extraction d'information), les règles d'adaptation au corpus (genre ou famille de langues), enfin les règles d'adaptation aux préférences de l'utilisateur (garder ou non les illustrations, certaines d'entre elles).

Il paraît en effet raisonnable de conserver l'illustration dans une sortie d'analyse pour une explication de type fiche pédagogique. De même, dans un débat, il est utile de conserver les données chiffrées appuyant l'argumentation. A titre d'illustration je propose ici une copie d'écran provenant des essais que j'ai menés à partir des programmes Thema et Unithem pour conserver les illustrations et tableaux comme partie intégrante dans la structuration du document électronique. Dans un article de médecine clinique, du sous-genre lettre à l'éditeur, deux tableaux de chiffres sont traités comme thèmes (Figure 1). Le tableau 1 mentionné en G2 constitue l'information principale à expliquer, le tableau 2 est considéré comme subordonné. Le tableau 3 en revanche, sur lequel

s'appuie l'argumentation ultérieure, est placé en G3 au même niveau de profondeur que le tableau 1 en G2.

Unité Thématique G , niveau 1	
THEME	J Med Genet 2000; 37 : 947-949 (December)
RHEME	Unité Thématique G1 , niveau 2
THEME	Letters to the editor
RHEME	Unité Thématique G11 , niveau 3
THEME	Suggestive evidence for a site specific prostate cancer gene on chromosome 1p36
RHEME	E DITOR A report was recently published on the localisation of a chromosome segment at 1p36 which appeared to be linked (two point lod score=4.74) to a large number of families with multiple cases of early onset (mean age at diagnosis of 66 years) prostate cancer (PC) in which a brain tumour had been reported in a first or second degree relative of a PC case. 1 This result is consistent with epidemiological evidence suggesting a familial relationship between brain cancer and PC as well as numerous studies of LOH at 1p36 in brain tumours. 1 As part of the ACTANE (Anglo/Canadian/Texan/Australian/Norwegian/EU Biomed) familial PC Consortium, [...] the family histories of all cancers were abstracted from the databases of several Consortium members and included in the analysis.
	Unité Thématique G12 , niveau 3
THEME	Table 1 presents the characteristics of the families; [...] NPL, and hlod scores maximised at D1S1160.
RHEME	Overall, [...] View this table: [in this window] [in a new window]
	Unité Thématique G2 , niveau 2
THEME	Table 1 Number of ACTANE families by source and number affected
RHEME	Unité Thématique G21 , niveau 3 View this table: [in this window] [in a new window]
	Unité Thématique G22 , niveau 3
THEME	Table 2 Linkage results from nine prostate-brain cancer families [...] Table 3 presents the linkage analysis results for the entire ACTANE pedigree set subdivided according to mean age at diagnosis of affected men in the family.
RHEME	The four age groups presented in table 3 were chosen to give approximately equal representation and are listed from the youngest (59 years) to the oldest (80 years). [...] View this table: [in this window] [in a new window]
	Unité Thématique G3 , niveau 2
THEME	Table 3 Linkage results from all 207 ACTANE families [...] Acknowledgments
RHEME	Unité Thématique G31 , niveau 3
THEME	This study was supported by The Cancer Research Campaign, [...] UK
RHEME	Correspondence to: [...] WA 98195, USA
	Unité Thématique G32 , niveau 3
THEME	References 1. Gibbs M, [...] Linkage map integration.
RHEME	Genomics 1996; 36 :157-162 [Medline] . [...] Am J Hum Genet 1995; 56 :265-271 [Medline] .

Figure 6.1. Structuration thématique d'un article de clinique, avec ses tableaux

Ces recherches innovantes peuvent être menées sous l'égide de la sémiotique, qui propose une cohérence méthodologique fédérant l'analyse rhétorique, l'apprentissage et la fouille de données textes et images (Brier, 2004 ; Kanellos & Mauceri 2008 ; Datta *et al.*, 2008 ; Popescu *et al.*, 2008).

3. RETOMBÉES APPLICATIVES ATTENDUES

J'évoque ici les prolongements de ma recherche expérimentale à très court terme, car il est bien entendu impossible de prédire les retombées de la recherche de fond, même à moyen terme. Du reste, je pensais que les traitements multilingues seraient bien accueillis au niveau européen, il n'en a rien été. C'est à travers la résolution de l'anaphore, un problème identifié, par le traiteemnt multigrain, que j'ai pu arguer pour un traitement multilingue, qui heurte les conceptions politiques

communément admises : le slogan « une langue, un peuple, une nation » cristallise toujours le sentiment d'identité.

3.1. Recherche/extraction d'information configurable MMM

3.1.1. Recherche/extraction d'information multigrain

Une application particulièrement utile serait d'assurer l'extraction d'information pour la veille sanitaire sur un corpus multilingue dans le cadre du système européen MedISys. Actuellement, seules les informations en anglais sont traitées, reflétant en cela les limitations de l'état de l'art. Le système PULS existant a été réalisé à l'université d'Helsinki. Une collaboration entre l'équipe Doremi, qui l'a réalisé, et l'équipe ISLanD du GREYC est en cours pour rendre ce système bilingue dans un premier temps (version MultiPULS alpha, français multigrains). La co-direction de stage a commencé en automne 2008 (Gaël Lejeune sous la direction d'Antoine Doucet et moi-même au GREYC et Roman Yangarber à Helsinki). Comme les résultats sur le français sont d'emblée bien supérieurs aux résultats sur l'anglais, une version anglaise multigrain est envisagée (l'anglais est la vitrine de choix pour les décisions politiques). Un projet d'échanges bilatéraux PICS du Cnrs a été accepté pour 2009-2011. La co-direction de thèses est prévue.

3.1.2. Recherche/extraction d'information multilingue

Il est bien sûr souhaitable de mener des recherches sur les langues européennes, ce qui est entrepris dans l'équipe Island en testant les limites des méthodes par familles de langue (Vergne, 2005, Vega & Vergne, 2005). Je souhaite également inclure les langues asiatiques et les langues « faiblement dotées » c'est-à-dire n'ayant pas ou peu fait l'objet de descriptions accessibles par des moyens informatiques. Les expériences encore limitées menées sur les articles de presse et sur les forums gagneraient à être étendues à d'autres langues, de façon à définir explicitement les limites de chaque méthode sur l'axe de variation des langues.

Un projet de collaboration au niveau européen se dessine avec l'université d'Helsinki et le Joint Research Center JRC pour rendre le système MedISys réellement multilingue et autonome (version MultiPULS bêta).

3.1.3. Recherche/extraction d'information multimodale

Les travaux que j'ai menés sur la structure dynamique du texte concernent les manières de lire et comprendre les documents illustrés, notamment les documents géographiques, articles et atlas. Ils n'ont pas donné lieu à des programmes complets de traitement des documents, mais les conditions sont réunies pour tester des hypothèses. Les cartes de géographie sont selon les géographes, un langage à part entière, elles font souvent partie de l'argumentation (Lucas, 2009b). Il en va de même en médecine, où les illustrations sont complexes et nombreuses, et dans d'autres disciplines. Les mathématiques s'illustrent par un langage graphique spécialisé, les formules, souvent traité

informatiquement comme une image (Herreman, 2001). Dans ces domaines, une approche résolument fondée sur la sémiotique du discours, plutôt que sur la juxtaposition de données traitées séparément, s'avère payante (Kanellos & Mauceri 2008 ; Popescu *et al.*, 2008). Les questions de disposition, de grain d'analyse, de style mériteraient de plus amples investigations.

De même, dans les manuels scolaires, les articles académiques, les articles de presse, il paraît indispensable de tenir compte de l'illustration pour faciliter la recherche d'information, notamment dans un contexte multilingue. C'est une position défendue au sein du consortium européen *Cross-Language Evaluation Forum* CLEF.

3.2. Documents volumineux

Le nombre croissant d'ouvrages électroniques disponibles suscite de nouveaux besoins d'accès au contenu. Les concours qui se mettent en place portent sur diverses problématiques, l'extraction de passages, la comparaison de collections de livres ou le partage de points de vue des lecteurs. Le GREYC et l'équipe Island, avec la participation d'Antoine Doucet au comité d'organisation pour la recherche d'information dans les ouvrages électroniques (*Book search track*) dans le cadre international *INitiative for the Evaluation of XML Retrieval* (INEX), pourraient être moteur dans ce domaine émergent.

Comme nombre d'ouvrages libres de droit sont des documents anciens, filmés ou scannés, le travail réalisé sur l'image des pages dans le projet Résurgence est un point de départ pour tenter le recouvrement de la structure logique des livres. Il s'agit d'optimiser les traitements pour gérer des collections importantes et très lourdes en taille de fichiers. Le développement de techniques de « maillage » déjà évoquées est une clé pour la gestion des unités typo-dispositionnelles à multiples échelons.

3.3. MMM

3.3.1. Extraction de texte et illustration

Une problématique en rapport avec la recherche d'information multimedia concerne les résumés d'articles avec extraction d'image, ou la production de passages de texte (microdocuments extraits d'ouvrages) avec illustration. Cette demande est forte en médecine, et dans les domaines techniques, mais aussi dans la gestion du patrimoine artistique et culturel.

3.3.2. Méthodologie transversale

Le domaine émergent MMM pour multi document (les collections), multilingue et multi media correspond à un besoin des utilisateurs face à la diversité des documents signalés. Les questions récemment soulevées dans le domaine des techniques multimodales tenant compte de l'utilisateur font écho à nos préoccupations (Jaimes & Sebe, 2007 ; Datta *et al.*, 2008). L'interprétation est

nettement distinguée de l'analyse automatique des données brutes. Les principaux points soulevés pour l'analyse d'image sont applicables aussi au texte :

- l'étendue du contexte nécessaire à l'interprétation des données ;
- les capacités d'adaptation d'un système ;
- les avantages et limites des techniques à couverture large versus les techniques à objet spécifique ;
- la fusion des sources d'information.

3.4. Résumé

Dans le prolongement du traitement d'ouvrages, le résumé est incontournable. L'intérêt des méthodes d'analyse textuelle est d'autant plus grand qu'elles s'appliquent aux textes longs, ceux qu'il est intéressant de résumer en pratique. La gestion de l'échelle et le choix d'une bonne résolution semblent pré requis dans cette optique. Il est tentant d'approfondir la recherche de points de vue, ou de thème, pour guider des algorithmes de résumé de documents complexes.

Dans le contexte académique actuel, le résumé concerne des textes courts, généralement des textes journalistiques, et se limite à l'extraction de passages de texte en un nombre fixe de mots (100, 250). Le résumé ou plutôt synopsis de collections d'articles, en contexte multilingue, semble un défi abordable, proposé dans le cadre d'une tâche CLEF. L'analyse globale d'articles par la procédure descendante et différentielle permet de pondérer contextuellement les passages importants. Cette technique pourrait concurrencer les systèmes existants, qui se fondent sur la pondération de phrases à l'aide d'indices lexicaux.

Une problématique déjà mentionnée en rapport avec la recherche d'information mono ou multi document concerne le résumé d'articles avec extraction d'image.

3.5. Traduction

Le traitement multilingue fait partie des priorités affichées au niveau européen. Il comprend la gestion de documents alignés — traduction l'un de l'autre — ou de documents très similaires (sur le même sujet et traités de façon comparable). L'équipe ISLanD s'est engagée dans cette voie (Giguet & Luquet, 2006), ainsi que dans celle de la traduction par analogie (Lepage, 2003).

Les paramètres de style collectif (propres au genre ou au sous-genre), de style personnel, de longueur et densité du texte sont importants dans la perspective de la traduction. L'analyse globale et à de multiples échelons de textes en différentes langues devrait permettre d'acquérir des procédures de comparaison utiles pour la génération de traduction dans une langue cible. C'est le travail de thèse en cours de Romain Brixtel, co-encadré par Jacques Vergne et Emmanuel Giguet.

Il semble également important de relier les recherches sur la résolution optimale pour l'analyse automatique des textes à la problématique de la traduction. En effet, la traduction humaine ne se fait ni mot à mot ni même phrase à phrase, de nombreux indicateurs étant répartis dans le contexte à différents niveaux suivant les langues. Ceci est d'autant plus sensible que les familles de langues sont distantes, par exemple, pour traduire du japonais en français.

Même si l'aide à la traduction semble un objectif plus lointain que les autres orientations citées, l'hybridation des méthodes que j'envisage entre projection et découverte de règles, production et découverte d'indices nouveaux pourrait avoir des retombées positives.

REFERENCES BIBLIOGRAPHIQUES

- ABDALLA, Rashid & Simone TEUFEL (2006). A bootstrapping approach to unsupervised detection of cue phrase variants. *21st International Conference on Computational Linguistics Coling 06*, Sydney, Australia. pp. 921-928.
- ADAM, Jean-Michel (1977). "Ordre du texte, ordre du discours" *Pratiques* (13) pp. 103-111.
- ADAM, J-M. (1999). *Linguistique textuelle. Des genres de discours aux textes*. Paris, Nathan. 208 p.
- ADAM, Jean-Michel, Jean-Blaise GRIZE & Magid ALI BOUCHA (Eds.) (2004). *Textes et discours: Catégories pour l'analyse*. Dijon: Editions universitaires de Dijon. 272 p.
- AHO, Alfred V. & Margaret J. CORASICK (1975). Efficient string matching: An aid to bibliographic search, *Communications of the ACM*, vol. 18 (6) pp. 333-340. DOI 10.1145/360825.360855
- AÏT-EL-HADJ, Smaïl (2006). Intelligence économique et conception : les apports de la systémique technologique. In B. Yannou & P. Deshayes (Eds.), *Intelligence et innovation en conception de produits et services*. Paris: L'Harmattan. pp. 127-140.
- ANANIADOU, Sophia, Douglas B. KELL & Jun-Ichi TSUJII (2006). "Text mining and its potential applications in systems biology" *Trends in Biotechnology* 24 (12) pp. 571-579.
- ARGAMON, Shlomo, Casey WHITELOW, Paul J. CHASE, Soban RAJ HOTA, Navendu GARG & Shlomo LEVITAN (2007). "Stylistic text classification using functional lexical features" *Journal of American Society for Information Science & Technology (JASIST)* 58 (6) pp. 802-822.
- ASHER, Nicholas & Alex LASCARIDES (2003). *Logics of Conversation*. Cambridge University Press, Cambridge, England.
- AUROUX, Sylvain, Ed. (1989). *Histoire des idées linguistiques: la naissance des métalangages en Orient et en Occident*. Liège, Bruxelles : Pierre Mardaga. 510 p.
- AUROUX, Sylvain, Ed. (1992). *Histoire des idées linguistiques: le développement de la grammaire occidentale*. Liège, Bruxelles : Pierre Mardaga. 683 p.
- AUROUX, Sylvain, Ed. (2000). *Histoire des idées linguistiques : l'hégémonie du comparatisme*. Liège, Bruxelles : Pierre Mardaga. 608 p.
- BARTHES, Roland (1966). "Introduction à l'analyse structurale des récits" *Communications* 8. pp. 1-27.
- BAZIN, Patrick (1996). "Vers une métalecture" *Bulletin des bibliothèques de France* (1). pp. 8-15.
- BESSON Jeremy, Christophe RIGOTTI, Ieva MITASIUNAITE, Jean-François BOULICAUT (2008). Parameter Tuning for Differential Mining of String Patterns, *2nd Int. Workshop on Domain Driven Data Mining DDDM'08 with IEEE ICDM'08*, Pisa, Italy. IEEE Computer Society, pp. 77-86.
- BILHAUT, Frederik (2006) Analyse automatique de structures thématiques discursives : application à la recherche d'information. Thèse de l'université de Caen.
- BILHAUT, F. (2007). Analyse thématique automatique fondée sur la notion d'univers de discours *Discours*, 1/2007. <http://discours.revues.org/>
- BLOOMFIELD, Leonard (1958 [1933]). *Language*. London, George Allen & Unwin, réed 1958. 566 p.
- BOURIGAULT, Didier (1992). Surface grammatical analysis for the extraction of terminological noun phrases. *CoLing 92*, Nantes, ACL. pp. 977-981.
- BOYER, R. S. & J. S. MOORE (1977). A fast string-searching algorithm *Communications of the ACM*, vol. 20 (10). pp. 762-772.
- BRATITSIS, Tharrenos & DIMITRACOPOULOU, Angélique (2008). Interpretation issues in monitoring and analyzing group interactions in asynchronous discussions. *International Journal of e-Collaboration*, 4 (1), pp. 20-40.

- BRIER, Søren (2004). Cybersemiotics and the problems of the information-processing paradigm as a candidate for a unified science of information behind library information science *Library Trends* Winter (27p).
- BÜHLER, Karl (2009 [1934]). *Théorie du langage* trad. Didier Samain [titre original *Sprachtheorie. Die Darstellungsfunktion der Sprache*. Stuttgart, New York : Fischer]. Marseille : Agone. 687 p.
- CHARAUDEAU, Patrick & Dominique MAINGUENEAU, Eds. (2002). *Dictionnaire d'analyse du discours*. Paris : Seuil. 666 p.
- COURSIL, Jacques (1992). *Grammaire analytique du français contemporain. Essai d'intelligence artificielle et de linguistique générale*. Thèse d'Etat Université de Caen, Caen.
- COURSIL, Jacques (2000). *La fonction muette du langage*. Matoury : Ibis rouge. 112 p.
- CREMILLEUX, Bruno, Arnaud SOULET, Jiri KLEMA, Céline HEBERT & Olivier GANDRILLON (2009). "Discovering Knowledge from Local Patterns in SAGE data" in .P. Berka *et al.* (Ed.) *Data Mining and Medical Knowledge Management: Cases and Applications*. Hershey, Pennsylvania, USA : IGI Global. pp. 251-267.
- CRISTEA, Dan, Nancy IDE & Laurent ROMARY (1998). Veins theory: A model of global discourse cohesion and coherence. *17th international conference on Computational Linguistics*, August 10-14, Montreal, Quebec, Canada. pp. 281-285.
- CROCHEMORE, Maxime, Christophe HANCART & Thierry LECROQ (2001). *Algorithmique du texte*. Paris: Vuibert. 347 p.
- CROCHEMORE, Maxime & Wojciech RYTTER (2002). *Jewels of Stringology*. Singapore: World Scientific Publishing. 392 p.
- DANEŠ, František (1966). The Relation of Centre and Periphery as a Language Universal, *Travaux linguistiques de Prague 2, Les problèmes du centre et de la périphérie du système de la langue*. 1^{ère} éd. Prague : Academia [Editions de l'Académie tchécoslovaque des Sciences], réed. distrib. Amsterdam : John Benjamins, 1996. réed [en ligne] Jihočeská Univerzita [Université de Bohême du Sud Faculté de philosophie] 2009 http://www.ff.jcu.cz/research/ee/tcclp-cp_textes.php (consulté février 09)
- DANEŠ, František (1994). "Prague school functionalism as a precursor of text linguistics" *Cahiers de l'I.L.S.L.* 5 (L'école de Prague: l'apport épistémologique). Institut de Linguistique et des Sciences du Langage de l'Université de Lausanne.
- DATTA, Ritendra, Dhiraj JOSHI, Jia LI & James Z. WANG (2008). "Image Retrieval: Ideas, Influences, and Trends of the New Age" *ACM Transactions on Computing Surveys*. vol. 40 (2), 60 p.
- DEJEAN, Hervé (1998a). Inférence automatique de contextes distributionnels. *Cinquième conférence annuelle: Le Traitement Automatique des Langues Naturelles, TALN'98*, Paris. pp. 229 – 235.
- DEJEAN, Hervé (1998b). Concepts et algorithmes pour la découverte des structures formelles des langues. Caen : thèse de l'Université de Caen. Spécialité : informatique.
- DEJEAN, H. (2002) Learning Rules and Their Exceptions. *Journal of Machine Learning Research* 2. pp. 669-693.
- DEJEAN, Hervé & Jean-Luc MEUNIER (2007). Logical document conversion: combining functional and formal knowledge. *ACM Symposium on Document Engineering*, pp. 135-143.
- DELCROIX, Maurice & Walter GEERTS (1981). *"Les chats" de Baudelaire: Une confrontation de méthodes*. Namur: Presses universitaires de Namur avec les Presses universitaires de France. 374 p.
- DESSALLES, Jean-Louis (2008). *La pertinence et ses origines cognitives : nouvelles théories*. Paris : Lavoisier. 204 p.
- DOURY, Marianne & Sophie MOIRAND, Eds (2005). *L'argumentation aujourd'hui : positions théoriques en confrontation*. Paris : Presses Sorbonne nouvelle. 200 p.

- DUMAIS, S. T., G. W. FURNAS, T. K. LANDAUER, *et al.* (1998). Using latent semantic analysis to improve access to textual information, *Conference on Human Factors in Computing (CHI'98)*, Washington, DC : ACM Press, pp. 281-285.
- DUTOIT, Dominique (2000). *Quelques opérations sens --> texte et texte --> sens utilisant une sémantique linguistique universaliste a priori*. Thèse de Doctorat, Université de Caen.
- EGG, Markus & Gisela REDEKER (2008). Underspecified discourse representation. In: Anton Benz & Peter Kühnlein (eds), *Constraints in Discourse*, Amsterdam: Benjamins. pp. 117-138.
- FAHNESTOCK, Jeanne (1999). *Rhetorical figures in science*. Oxford, New York: Oxford University Press. 234 p.
- FIRBAS, Jan (1964). "On defining the theme in functional sentence analysis" in *Travaux linguistiques de Prague nouvelle série I* (Hajicova *et al.*, Ed.). Prague, distrib. Amsterdam : John Benjamins, pp. 267-280.
- FIRTH, John Rupert (1957). *Papers in Linguistics, 1934-1951*. Oxford : Oxford University Press. 233 p.
- FLORIDI, Floriano (2003). From data to semantic information *Entropy* (5) pp. 125-145. en ligne, consulté février 2009 www.mdpi.org/entropy
- FLØTTUM, Kjersti (2001). "Personal English, indefinite French and plural Norwegian scientific authors?" *Norsk Lingvistisk Tidsskrift* (21) pp. 21-55.
- FLØTTUM, Kjersti & François RASTIER, Eds. (2003). *Academic discourse, multidisciplinary approaches*. Oslo, Novus. 222 p.
- FOLTZ, Peter W., Walter KINTSCH & Thomas K. LANDAUER (1998). "The Measurement of Textual Coherence with Latent Semantic Analysis" *Discourse Processes* 25 (2-3) pp. 285-307.
- GANASCIA, Jean-Gabriel (1985). Comment oublier à l'aide de contre-exemples. *RFIA*, Grenoble.
- GANASCIA, Jean-Gabriel (1991). "Deriving the learning bias from rule properties" in Hayes, J.E., D. Mitchie & E. Tyugu *Machine intelligence 12: towards an automated logic of human thought* New York : Clarendon Press. pp. 151-167.
- GANASCIA, J.-G. (1993). "Algebraic structure of some learning systems" in *Algorithmic Learning Theory*. Berlin Heidelberg : Springer. Lecture Notes in Computer Science 744 pp. 398-409. doi : 10.1007/3-540-57370-4_63
- GANASCIA, J.-G. (1998). *Logic and induction: an old debate*, CiteSeer Scientific Commons. 2009: 43 p.
- GANASCIA, J.-G. (2001). Extraction of Recurrent Patterns from Stratified Ordered Trees. *12th European Conference on Machine Learning*, L. de Raedt & P. A. Flach, (ed) Berlin Heidelberg : Springer. Lecture Notes in Computer Science 2167 pp. 167-178. doi : 10.1007/3-540-44795-4_15
- GANASCIA, J.-G. (2007). Ethical System Formalization using Non-Monotonic Logics. *Cognitive Science conference (CogSci2007)*, Nashville, USA. pp. 1013-1018. en ligne consulté mars 2009 <http://www.cogsci.rpi.edu/CSJarchive/proceedings/2007/>
- GARNIER, Catherine (1982). *La phrase japonaise: structures complexes en japonais moderne*. Paris : Presses Orientalistes de France. 181 p.
- GARNIER, C. (2001). "La phrase japonaise, composants et structure" *Faits de langues* (17) pp. 149-156.
- GIGUET, Emmanuel & Nadine LUCAS (2002). Intégration d'Unicode : Conception d'un agent de recherche d'information sur internet. *Document numérique*, 6 (3-4), pp. 225-236.
- GIGUET, Emmanuel & Nadine LUCAS (2004). "La détection automatique des citations et des locuteurs dans les textes informatifs" in *Le discours rapporté dans tous ses états: Question de frontières*. J. M. López-Muñoz *et al.* (Ed.). Paris, L'Harmattan. pp. 410-418.
- GIGUET, Emmanuel & Pierre-Sylvain LUQUET (2006). Multilingual lexical database generation from parallel texts in 20 European languages with endogenous resources. *Computational Linguistic Conference COLING/ACL 2006* Sydney, Australia, pp. 271-278.
- GIGUET, E., N. LUCAS & C. CHIRCU (2008). Le projet Résurgence: Recouvrement de la structure logique des documents électroniques, *JEP-TALN-RECITAL'08 Session "Show & Tell"*. Avignon.

- GIGUET, Emmanuel & Nadine LUCAS (2009). Creating discussion threads graphs with Anagora. *Computer-supported Collaborative Learning Conference CSCL 2009*, Rhodes. pp. 616-620.
- GRIZE, Jean-Blaise (1990a). *Logique et langage*. Paris : Ophrys. 153 p.
- GRIZE, Jean-Blaise (1990b). "La construction du discours: un point de vue sémiotique" in *Le discours: représentations et interprétations*. M. CHAROLLES, S. FISCHER & J. JAYEZ, (Ed.). Nancy, Presses universitaires de Nancy. pp. 11-18.
- GROSZ, Barbara & Candace SIDNER (1986). "Attention, intention and the structure of discourse" *Computational Linguistics* 12 (3) pp. 175-204.
- GROSZ, Barbara & Candace SIDNER (1990). "Plans for discourse" in *Intentions in Communication*. P. Cohen *et al.* (Ed.). Cambridge, Mass. : MIT Press pp. 417-444.
- GROSZ, Barbara, Aravind JOSHI & Scott WEINSTEIN (1995). "Centering: a framework for modelling the local coherence of discourse" *Computational Linguistics* 21 (2) pp. 203-225.
- HABERT Benoît, Adeline NAZARENKO & André SALEM (1997). *Les linguistiques de corpus*. Paris : Armand Colin. 192 p.
- HABERT, Benoît, Gabriel ILLOUZ, Pierre LAFON *et al.* (2000). Profilage de textes: Cadre de travail et expérience. *5èmes Journées d'Analyse des Données Textuelles (JADT)*, Lausanne.
- HABERT, Benoît, Ed. (2004). Linguistique et informatique: nouveaux défis. *Revue Française de Linguistique Appliquée*. vol. IX n°1, juin 2004, Paris. 140 p.
- HABERT, Benoît (2004). Outiller la linguistique : de l'emprunt de techniques aux rencontres de savoirs *Revue Française de Linguistique Appliquée*. vol. IX n°1, juin 2004. pp. 5-24.
- HABERT, Benoît, Gabriel ILLOUZ & Helka FOLCH (2005). Des décalages de distribution aux différences d'acception. In A. Condamines (Ed.), *Sémantique et corpus*. Paris: Hermès Lavoisier. pp. 278-318.
- HAGEGE, Claude (1982). *La structure des langues*. Paris : PUF. 127 p.
- HAJICOVA, Eva (1983). [Hajičová] "Topic and Focus" *Theoretical Linguistics* 10 (2/3) pp. 268-273.
- HAJICOVA, Eva, Petr KUBON & Václav KUBON (1990). Hierarchy of Salience and Discourse Analysis and Production. *CoLing 90. 13th International conference on Computational Linguistics*, Helsinki, H. Karlgen (Ed.). pp. 144-148.
- HAJICOVA, Eva, Tomas HOSKOVEC & Petr SGALL (1992). "Discourse Modelling Based on Hierarchy of Salience" *Prague studies in Mathematical Linguistics* 64 (11) pp. 5-24.
- HAJICOVA, Eva (2007). "Information Structure from the Point of View of the Relation of Function and Form " *The Prague Bulletin of Mathematical Linguistics* (88) pp. 53-72.
- HALL J., J. NILSSON, J. NIVRE, G. ERYIGIT, B. MEGYESI, M. NILSSON & M. SAERS (2007) Single Malt or Blended? A Study in Multilingual Parser Optimization *CoNLL Shared Task Session of EMNLP-CoNLL 2007*, pp. 933-939.
- HALLIDAY, Michael & Ruqaiya HASAN (1976). *Cohesion in English*. London : Longman. 392 p.
- HALLIDAY, Michael A. K. (1985). *An introduction to functional grammar*. London : Edward Arnold.
- HALLIDAY, M.A.K. & J. R. MARTIN (1993). *Writing Science: Literacy and Discursive Power*. University of Pittsburgh Press. 283 p.
- HARRIS, Zellig (1951). *Methods in structural linguistics*. Chicago : University of Chicago Press. 384 p.
- HARRIS, Zellig (1952). "Discourse analysis" *Language* (28) pp. 1-30.
- HARRIS, Zellig (1991). *A theory of language and information: A mathematical approach*. Oxford: Clarendon Press.
- HARRIS, Zellig (2002). "The structure of science information" *Journal of Biomedical informatics* (35) pp. 215-221.
- HAUSSER, Roland (2001). *Foundations of Computational Linguistics: Human-Computer Communication in Natural Language*. Berlin : Springer. 578 p.

- HEARST, M. (1994). Multi-Paragraph Segmentation of Expository Text. *32nd Annual Meeting of the Association for Computational Linguistics*, Las Cruces, NM, June 1994. pp. 9-16.
- HEBERT, Louis (2006). « Les fonctions du langage », dans Louis Hébert (dir.), *Signo*, Rimouski (Québec) [en ligne] <http://www.signosemio.com>. (consulté le 9 février 2009)
- HEINONEN, Oskari (1998). Optimal Multi-Paragraph Text Segmentation by Dynamic Programming. *COLING-ACL'98*, Association for Computational Linguistics. pp. 1484-1486.
- HERREMAN Alain (2001) La mise en texte mathématique. Une analyse de l'« algorithme de Frankenthal ». *Methodos* (1), pp. 61-100. mis en ligne 2004 <http://methodos.revues.org/document45.html> (consulté le 9 février 2009)
- HINDS, John (1983). "Contrastive Rhetoric: Japanese and English" *Text* 3. pp. 183-196.
- HOCKETT, Charles F. (1958). *A course in modern linguistics*. New York: MacMillan. xi, 621 p.
- HOEY, Michael (2001). *Textual Interaction*. London, Routledge. 203 p.
- HOUDE, Olivier, Daniel KAYSER, Olivier KËNIG, Joëlle PROUST & François RASTIER (1998). *Vocabulaire de sciences cognitives*. Paris : Presses universitaires de France. 409 p.
- HUNTER, Lawrence & Kevin BRETONNEL COHEN (2006). "Biomedical Language Processing: What's Beyond PubMed?" *Molecular Cell* (21): 589-594.
- IIDA, Masayo (1997). "Discourse coherence and shifting centers in Japanese texts" in *Centering in Discourse*. M. Walker et al. (Ed.). Oxford : Oxford University Press. pp. 161-180.
- JAIMES, Alexandro & Nicu SEBE (2007). Multimodal Human Computer Interaction: A Survey. *Computer Vision in Human-Computer Interaction*, Springer. LNCS 3766 / 2005 pp. 1-15.
- JAKOBSON, Roman. (1960-1988). *Selected writings* (8 vol.). The Hague, Paris: Mouton.
- JAKOBSON, Roman (1963). *Essais de linguistique générale*. Paris : Editions de Minuit. 260 p.
- JAKOBSON, Roman (1973). *Essais de linguistique générale: rapports internes et externes du langage*. Paris : Editions de Minuit. 317 p.
- JONES, Linda Kay (1977). *Theme in English expository discourse*. Lake Bluff, Ill. : Jupiter Press. 308p.
- KAESER, Pascal (1997). *Nouveaux exercices de style: Jeux mathématiques et poésie*. Paris: Diderot.
- KANDO, Noriko (1997). Text-Level Structure of Research Papers: Implications for Text-Based Information Processing Systems. *19th Annual BCS-IRSG Colloquium on IR Research*, Aberdeen, Scotland 8-9 April 1997, J. Furner & D. Harper (Eds), Berlin : Springer. pp. 68-81.
- KANELLOS Ioannis (2008). Lecture(s) et genre(s) du document numérique (Conférence invitée). *11ème Colloque International sur le Document Electronique, CIDE 11* 28-31 octobre, Rouen.
- KANELLOS, Ioannis & Christian MAUCERI (2008). Une conscience interprétative face à un univers de textes. Arguments en faveur d'une Analyse de Données Interprétative. *Syntaxe & Sémantique*, novembre 2008, n° 9. pp. 37-52.
- KANELLOS Ioannis, Thomas LE BRAS, Ioana SUCIU & S. DANIILIA (2007). Interpretative e-Learning Personalization: Methodology, Formal Aspects and generic Scenarios of Individual/Group Dynamics. A case of a course in art history. *11th International Conference on User Modeling : Workshop " Personalisation in E-Learning Environments at Individual and Group Level"*, June 25-29, Corfu, Greece. pp. 75-76.
- KARCZMARCZUK, Jurek (2007). Complexité des images: Attribut scientifique ou mythe? In *Art et Complexité A la mémoire de Bernard Caillaud* sous la dir. de Jean Vivier. Caen : Université de Caen, Modesco. en ligne <http://www.unicaen.fr/mrsh/publications/online.php>
- KARP, Richard M. & RABIN, Michael O. (1987). Efficient randomized pattern-matching algorithms. *IBM Journal of Research and Development* 31, 2 (Mar. 1987) pp. 249-260.
- KNOBLOCK, Craig, Daniel LOPRESTI, Shourya ROY & L. Ventaka SUBRAMANIAN (2007). "Special issue on noisy text analytics" *International Journal of Document Analysis and Recognition* (10) pp. 127-128.

- KUSHMERICK, Nicholas (1997). *Wrapper Induction for Information Extraction* Washington, DC : University of Washington, PhD thesis.
- KUSHMERICK, Nicholas (2000). "Wrapper Induction: Efficiency and Expressiveness" *Artificial Intelligence* (118) pp. 15-68.
- KUBON, Vladislav, Marketa LOPATKOVA, Martin PLATEK & Patrice POGNAN (2007). A linguistically-based segmentation of complex sentences. *20th International Florida Artificial Intelligence Research Society Conference, FLAIRS*, Key West, Florida. pp. 368-373.
- LABADIE, Alexandre & Violaine PRINCE (2008a). Finding text boundaries and ending topic boundaries: Two different tasks?. B. Nordstrom, A. Ranta, (eds) *6th International Conference on Natural Language Processing (GoTAL'08)* Berlin : Springer-Verlag (Lecture Notes in Artificial Intelligence vol. 5221) pp. 260-271.
- LABADIE, A. & V. PRINCE (2008b). The impact of corpus quality and type on topic based text segmentation evaluation. *Computer Linguistics Applications CLA'08*. Wisla, Pologne. IEEE. pp. 313-319.
- LACHERET, Anne & Jacques FRANÇOIS (2004). "De la notion de détachement topical à celle de constituant thématique extrapositionnel" *Cahiers de praxématique* 40 Linguistique du détachement (dir. F. Neveu).
- LAFOURCADE, Mathieu & Violaine PRINCE (2001). Relative synonymy and conceptual vectors. *Sixth Natural Language Processing Pacific Rim Symposium NLP RS2001*, Tokyo, Japan. pp. 127-134.
- LAMPRIER, S., A. TASSADIT, B. LEVRAT & F. SAUBION (2008). Thematic segment retrieval revisited. In *Artificial intelligence: Methodology, systems, and applications*. Berlin, Heidelberg: Springer. (Lecture Notes in Computer Science Vol. 5253) pp. 157-166.
- LARDILLEUX, Adrien & Yves LEPAGE (2008). A truly multilingual, high coverage, accurate, yet simple, sub-sentential alignment method. *Proceedings of the 8th Conference of the Association for Machine Translation in the Americas (AMTA 2008)*, Waikiki, Hawaï, USA, oct 2008. pp. 125-132.
- LAVALLARD, Anne (2008) *Exploration interactive d'archives de forum : le cas des jeux de rôle en ligne*. thèse d'université, Caen : Université de Caen et de Basse Normandie.
- LEPAGE, Yves (2003). De l'analogie rendant compte de la commutation en linguistique. Mémoire d'HDR, Université Joseph-Fourier - Grenoble I, Grenoble. <http://tel.archives-ouvertes.fr/tel-00004372/en/>
- LIDDY, Elizabeth D., Kenneth McVEARRY, Woojin PAIK, Edmund YU & Mary McKENNA (1993). Development, implementation and testing of a discourse model for newspaper texts. *Seminar on Summarizing Text for Intelligent Communication*, Dagstuhl. pp. 159-164.
- LONGACRE, Robert E. (1976). "The paragraph as a grammatical unit" in *Discourse and Syntax*. Givon & Talmy (Ed.). New York : Academic Press. pp. 115-134.
- LONGACRE, Robert E. (1983). *The grammar of discourse*. Topics in Language and Linguistics. New York, London: Plenum Press. xxi, 423 p.
- LOPEZ-MUÑOZ, Juan Manuel, Sophie MARNETTE & Laurence ROSIER, Eds. (2004). *Le discours rapporté dans tous ses états: Question de frontières*. Paris : L'Harmattan. (Sémantiques) 664 p.
- LOPEZ-MUÑOZ, J. M., S. MARNETTE & L. ROSIER, Eds. (2006). *Dans la jungle du discours rapporté: genres de discours et discours rapporté*. Cadix : Presses de l'Université de Cadix. 612 p.
- LUCAS, Nadine (1991). Syntaxe discursive du japonais scientifique : à propos du thème. *II encontro nacional de professores universitários de língua, literatura e cultura japonesa*, São Paulo, Brésil : Université de Sao Paulo. pp. 77-98.
- LUCAS, N. (1993). "Syntaxe du paragraphe dans les articles scientifiques en japonais et en français" in *Parcours linguistiques de discours spécialisés*. Moirand. et al. (Eds). Berne, Paris : Peter Lang pp. 249-261.
- LUCAS, Nadine (2002). Le retour des idéogrammes: Unicode CJC vu du Japon. *Document numérique*, 6 (3-4), pp. 183-210.

- LUCAS, Nadine, Bruno CREMILLEUX & Leny TURMEL (2003). Signalling well-written academic articles in an English corpus by text-mining techniques. *UCREL technical papers* Vol. 16 special issue Proceedings Corpus Linguistics 2003, Lancaster : Lancaster University. pp. 465-474.
- LUCAS, Nadine & Bruno CREMILLEUX (2004). Fouille de textes hiérarchisée, appliquée à la détection de fautes. *Document numérique* vol 8. pp. 107-133.
- LUCAS, N. (2005) "Etude linguistique des procédés d'exposition dans un forum de discussion" Eric Bruillard & Mohamed Sidir (eds) *Symfonic* <http://archive-edutice.ccsd.cnrs.fr/>
- LUCAS, Nadine & Emmanuel GIGUET (2005) UniTHEM, un exemple de traitement linguistique à couverture multilingue *Conférence internationale sur le document électronique Cide8*, Beyrouth, 25-28 mai 2005. Paris : Europa. pp. 115-132.
- LUCAS, N. (2006) Textbooks as a research challenge for computational linguistics. In *Caught in the web or lost in the textbook?* E. Bruillard, B. Aamotsbakken, S. V. Knudsen & M. Horsley (Eds.), Paris: IARTEM, STEF et IUFM Basse-Normandie. pp. 49-60.
- LUCAS, Nadine, Mohamed SIDIR & Emmanuel GIGUET (2006) Analyse de forums dans la formation à distance. *Conférence internationale sur le document électronique Cide 9*. Fribourg 18-20 septembre 2006. Zreik & Vanoirbeek (eds) Paris, Europa, pp. 169-180.
- LUCAS, N. & E. GIGUET (2008). Robust adaptive discourse parsing for e-learning fora. *The 8th IEEE International Conference on Advanced Learning Technologies (ICALT 2008)*, Santander, Spain July 1-5, IEEE. pp. 730-732.
- LUCAS, N. (2009a). "Discourse parsing for text mining" chapter 12 in *Information retrieval in biomedicine: Natural language processing for knowledge integration* V. Prince & M. Roche (Eds), Hershey, PA, USA : IGI Global (Medical Information Science Reference). pp. 229-262.
- LUCAS, N. (2009b). "Le discours des géographes en situation académique, pédagogique et médiatique". In *Les discours universitaires: formes, pratiques, mutations*, Defaÿs *et al.* (Eds) Paris : l'Harmattan.
- MANN, William C., Christian MATTHIESSEN & Sandra A. THOMPSON (1989). *Rhetorical structure theory and text analysis* Information Sciences Institute. ISI/RR-89-242.
- MANN, W. C. & S. A. THOMPSON, Eds. (1992). *Discourse description. Diverse linguistic analyses of a fund-raising text*. Amsterdam, Philadelphia, John Benjamins. 399 p.
- MARCOCCIA, Michel (2003). "La communication médiatisée par ordinateur : problèmes de genres et de typologie", Journée d'étude *Les genres de l'oral* le 18 avril 2003 [en ligne], Université Lumière Lyon 2 <http://gric.univ-lyon2.fr/Equipe1/actes/journees_genre.htm>. (Consulté le 8 février 2009).
- MARCU, Daniel (1997). The Rhetorical parsing of Natural Language Texts. *35th annual meeting of Association for Computational Linguistics, eighth conference of the European chapter of the Association for Computational Linguistics*, Madrid, Spain, Association for Computational Linguistics. pp. 93-103.
- MARCU, D. (2000). "Extending a formal and computational Model of Rhetorical Structure Theory with Intensional Structures à la Grosz and Sidner" The 18th International Conference on Computational Linguistics COLING'00, Saarbrueken, July 31-August 4. pp. 523-529.
- MARFOUK, Hicham & Noël GILAIN (2001). *EDDAP2 Extraction de discours et résolution d'attributions* Caen : mémoire Université de Caen.
- MARTIN, James R. (1992). *English text: system and structure*. Philadelphia, Amsterdam : Johns Benjamins. 620 p.
- McCOY, K. F. & M. STRUBE (1999). Taking Time to Structure Discourse: Pronoun Generation Beyond Accessibility. *21th Annual Conference of the Cognitive Science Society*, Vancouver, Canada.
- MOIRAND, Sophie (1990). *Une grammaire des textes et des dialogues*. [3^e éd. 2000], Paris : Hachette. 160 p.
- MOIRAND Sophie (1992). Des choix méthodologiques pour une linguistique de discours comparative. *Langages*, 26ème année (105) pp. 28-41.
- MOIRAND, S., A. Ali Bouacha, J. C. Beacco & A. Collinot (Eds.) (1993). *Parcours linguistiques de discours spécialisés*. Berne, Paris, New York: Peter Lang.

- MOIRAND, S. (1997). Formes discursives de la diffusion des savoirs dans les médias. *Hermès* (21), 33-44.
- MOIRAND, Sophie (2003). "Quelles catégories descriptives pour la mise au jour des genres du discours ? " Journée d'étude *Les genres de l'oral* 18 avril 2003, Université Lumière Lyon 2. [en ligne] <http://gric.univ-lyon2.fr/Equipe1/actes/journees_genre.htm>. (consulté le 8 février 2009).
- MOIRAND Sophie (2006). Responsabilité et énonciation dans la presse quotidienne : questionnements sur les observables et catégories d'analyse. *Semen* 22 ; pp. 45-59. [en ligne 2007] (consulté le 11 février 2009) <http://semen.revues.org/document2798.html>
- MOIRAND Sophie (2007). *Les discours de la presse quotidienne. Observer, analyser, comprendre*. Paris : Presses Universitaires de France. 186 p.
- MORAND, Bernard (2004). *Logique de la conception. Figure de sémiotique générale d'après C. S. Pierce*. Paris : L'Harmattan. 289 p.
- MORAND, B. (2006). "Arguments pour une sémiotique de la conception" in *Intelligence et Innovation en conception de produits et services* B. Yannou & P. Deshayes (Ed.). Paris : L'Harmattan. pp. 101-126.
- MORAND, B. (à paraître). "Quels savoir-faire faut-il cultiver pour des candidats aux métiers de l'informatique" in *De la didactique de l'informatique aux enjeux didactiques des progiciels*, G.-L. Baron, E. Bruillard & L. O. Pochon (Eds.).
- MOREL, Mary-Annick & Laurent DANON-BOILEAU, Eds. (1998). *Grammaire de l'intonation. L'exemple du français oral*. Paris : Ophrys. 231 p.
- MOURAD, Ghassan & Jean-Pierre DESCLES (2004). "Identification et extraction automatique des informations citationnelles dans un texte." in *Le discours rapporté dans tous ses états: Question de frontières*. J. M. López-Muñoz et al. (Eds.). Paris : L'Harmattan. pp. 397-409.
- MUSLEA, Ion, Steven MINTON & Craig KNOBLOCK (2001). "Hierarchical wrapper induction for semistructured sources" *Journal of Autonomous Agents and Multi-Agent Systems* (4) pp. 93-114.
- MUSLEA, I., S. MINTON & C. KNOBLOCK (2002a). Active+ Semi-Supervised Learning = Robust Multi-View Learning. *International Conference on Machine Learning 19th ICML*. pp. 435-442.
- MUSLEA, I., S. MINTON & C. KNOBLOCK (2002b). Adaptive view validation: a case study on wrapper induction. *International Conference on Machine Learning 19th ICML*. pp. 443-450.
- MUSLEA, Ion, Steven MINTON & Craig KNOBLOCK (2003). Active Learning with Strong and Weak Views: A Case Study on Wrapper Induction. *IJCAI*. pp. 415-420.
- NAVARRO, Gonzalo & Mathieu RAFFINOT (2007). *Flexible pattern matching in strings: Practical on-line search algorithms for texts and biological sequences* (paperback ed.). Cambridge: Cambridge University Press. 232 p.
- NICOLLE, Anne (1996). L'expérimentation et l'intelligence artificielle. *Intellectica* (96/1) pp. 9-19.
- NICOLLE, Anne (2002). "Sciences de l'artificiel, modélisation et rationalité" *Revue d'intelligence artificielle* 16 (1-2). pp. 63-86.
- NICOLLE, Anne (2003). "Le continu, le discontinu et le discret en informatique", *Espaces-Temps, les cahiers*, n° 82-83 2003. pp. 97-109.
- NICOLLE, Anne (2006). Un modèle des traces pour l'interaction entre les processus. *Journées de Rochebrune sur les systèmes complexes Traces, Enigmes, Problèmes : émergence et construction du sens*, Paris : ENST ISSN 1242-5125. pp. 75-87.
- NILSSON, Jens, Joakim NIVRE & Johan HALL (2006) Tree Transformations for Inductive Dependency Parsing *45th Annual Meeting of the Association for Computational Linguistics*, pp. 968-975.
- NIVRE, Joakim (2006). "Two strategies for text parsing". In M. Suominen et al., (Eds.), *A man of measure: Festschrift in honour of Fred Karlsson on his 60th birthday*. Turku: The Linguistic Association of Finland. pp. 440-448.
- NIVRE, J., J. HALL, J. NILSSON, G. ERYIGIT & S. MARINOV (2006) Labeled Pseudo-Projective Dependency Parsing with Support Vector Machines *Tenth Conference on Computational Natural Language Learning (CoNLL)*, pp. 221-225.

- NIVRE, J. (2008) Sorting Out Dependency Parsing *6th International Conference on Natural Language Processing (GoTAL)*, Gothenburg. Springer, pp. 16-27.
- O'NUALLAIN, Sean (2002). *The search for mind*. 2nd ed. Bristol : Intellect Books. 282 p.
- ORLIKOWSKI Wanda J. & JoAnne YATES (1994). Genre Repertoire: The Structuring of Communicative Practices in Organizations *Administrative Science Quarterly*, 39 (4) pp. 541-574.
- PEDAUQUE, R. T. (2006). *Le document à la lumière du numérique*, Caen : C&F Editions. 218 p.
- PEREGRIN, Jaroslav (1995). Structural linguistics and formal linguistics. In *Travaux du cercle linguistique de Prague nouvelle série*, Amsterdam: John Benjamins. (Vol. 1, pp. 85-97).
- PERY-WOODLEY, Marie-Paule (2000). *Une pragmatique à fleur de texte: approche en corpus de l'organisation textuelle*. Mémoire HDR. Toulouse : Université Toulouse Le Mirail, 181 p.
- PIKE, Kenneth (1958). "On tagmemes née grammemes" *International Journal of American Linguistics* (24) pp. 273-278.
- PIKE, Kenneth (1967). *Language in Relation to a Unified Theory of The Structure of Human Behavior*. The Hague, Mouton.
- PINATEL, Pascalie (2003). *Coloriage thématique à l'intérieur d'un document: approche contextuelle* mémoire DESS Université de Caen.
- POGNAN, Patrice (1975). *Analyse morphosyntaxique automatique du discours scientifique tchèque*. Paris: Dunod - Association Jean-Favard pour le développement de la linguistique quantitative. 262 p.
- POGNAN, P. (2007). Forme et fonction en analyse automatique du tchèque. Calculabilité des langues slaves de l'ouest. *BULAG* 32 (Les langues slaves et le français : approches formelles dans les études contrastives) pp. 13-33.
- POPESCU Adrian, MOELLIC Pierre-Alain, KANELLOS Ioannis (2008). ThemExplorer: finding and browsing geo-referenced images. *International Workshop on Content-Based Multimedia Indexing CBMI 2008*, June 18-20, London, UK. pp. 576-853. doi : 10.1109/CBBI.2008.4564999
- POPESCU-BELIS, Andrei (1999). Evaluation numérique de la résolution de la référence : critiques et propositions. *TAL* 40 (2) pp. 117-146.
- PRASAD, Rashmi, Nikhil DINESH, Alan LEE, Aravind JOSHI & Bonnie WEBBER (2006). Attribution and its Annotation in the Penn Discourse TreeBank. *revue tal* 47 (2) pp. 43-64.
- PRINCE, Violaine (2008). L'influence de quelques grands domaines discursifs sur les méthodologies et les performances d'applications en traitement automatique des langues. *Praxiling*, 28 janvier 2008 Montpellier.
- PRINCE, Violaine & Yves KODRATOFF (2007). "Le Défi fouilles de textes: quels paradigmes pour la reconnaissance d'auteurs?" *Revue des Nouvelles Technologies de l'Information* E 10 pp. 1-14.
- PRINCE, Violaine & Alexandre LABADIE (2007). Text segmentation based on document understanding for information retrieval. *12th International Conference on Applications of Natural Language to Information Systems NLDB'07*. Berlin : Springer Vol 4592 Natural Language Processing and Information Systems pp. 295-304.
- PRINCE, Violaine & Mathieu ROCHE (Eds) (2009). *Information Retrieval in Biomedicine: Natural Language Processing for Knowledge Integration*. Hershey, PA : IGI Global (Medical Information Science Reference) 460 p.
- RASTIER, François (1989). *Sens et textualité*. Paris : Hachette. 286 p.
- RASTIER, François (2001). *Arts et sciences du texte*. Paris : Puf. 303 p.
- RASTIER, F. (2002). Enjeux épistémologiques de la linguistique de corpus. *Journées de Linguistique de Corpus*, Lorient, septembre 2002, G. Williams (ed) Rennes : Presses universitaires de Rennes. pp. 31-45.
- RASTIER, François (2006). "La structure en question" revue *texto net*: 1-9. <http://www.revue-texto.net/>

- REDEKER Gisela & Markus EGG (2006) Says who? On the treatment of speech attributions in discourse structure. In Sidner, C. *et al.* (eds) *Workshop Constraints in Discourse*, pp. 140-146.
- REITTER, David (2003). Rhetorical Analysis with Rich Feature Support Vector Models. Mémoire. Postdam, Germany : University of Potsdam. <http://www.david-reitter.com/compling/index.html>
- ROSIER, Laurence (1998a). "Discours grammatical et ponctuation: l'exemple du discours rapporté" in J.-M. Defaÿs et al. (Ed.). *A qui appartient la ponctuation?* Paris, Louvain la Neuve : Duculot. pp. 353-364.
- ROSIER, Laurence (1998b). *Le discours rapporté: histoire, théories, pratiques*. Paris, Bruxelles : Duculot. 325 p.
- ROSIER, Laurence (2002) La Presse et les modalités du discours rapporté: l'effet d'hyperréalisme du discours direct surmarqué. *L'information grammaticale*. 94. pp. 27-32.
- ROSIER, Laurence (2009). *Le discours rapporté*. Paris : Ophrys.
- SABAH, Gérard (1988). *L'intelligence artificielle et le langage: représentation des connaissances*. Paris :Hermès. 352 p.
- SABAH, Gérard (1989). *L'intelligence artificielle et le langage : processus de compréhension*. Paris : Hermès. 411 p.
- SAKUMA, Mayumi (1983). "Danraku to paragurafu" [Paragraphe logique et paragraphe typographique] *Nihongo kyôiku* 2 (2) pp. 21-31.
- SAKUMA, Mayumi. Ed. (1989). *Bunshô kôzô to yôyakubun no shosô [Structures textuelles et caractéristiques des phrases de résumé]*. Tôkyô, Kuroshio. 281 p.
- SALTON Gerard (1989). *Automatic text processing: the transformation, analysis and retrieval of information by computer*, Boston, MA : Addison-Wesley, Longman. 530 p.
- ŠAUMJAN (Saumjan ou Shaumyan), Sebastian K. (1965). *Strukturalnaja lingvistika*, Moskva : Nauka, traduit en anglais en 1971, *Principles of Structural linguistics*, The Hague : Mouton.
- SEKI, Yohei, Koji EGUCHI, Noriko KANDO & Masaki AONO (2006). Opinion-focused Summarization and its Analysis at DUC 2006. *Document Understanding Conference 2006 (DUC 2006)*, New York City, New York, USA, June 2006. pp. 122-130.
- SCIUTO, Giovanni (s.d.) La comunicazione www.giovannisciuto.it/la_comunicazione.htm visité le 23 février 2009.
- SGALL, Petr (1982) Automatic Understanding with a Linguistically Based Knowledge Representation. *5th European Conference on Artificial Intelligence ECAI* 82. pp. 240-243.
- SGALL, Petr (1987). "Prague functionalism an [sic] topic vs. focus" In *Functionalism in Linguistics*, Dirven, R. and V. Fried (eds.), Amsterdam : John Benjamins (Linguistic and Literary Studies in Eastern Europe 20) pp. 169-190.
- SGALL, Petr (1992). "Classical structuralism and present-day Praguian linguistics". In *Prospects for a New Structuralism*, Lieb, H-H. (ed.), Amsterdam : John Benjamins. pp. 75-90.
- SGALL, Petr (1995). "Formal and computational Linguistics in Prague" In *Prague Linguistic Circle Papers*, Hajičová, E., M. Červenka, O. Leška & P. Sgall (eds.) distrib. Amsterdam : John Benjamins. pp. 23-37.
- SGALL, Petr, Eva HAJICOVA & Eva BURANOVA (2003). "Topic-Focus articulation and degrees of salience in the Prague Dependency Treebank". In *Formal Approaches to Function in Grammar*, Carnie, A, H. Harley and M.A. Willie (eds.) Amsterdam : John Benjamins. pp. 165-177.
- SHAUMYAN, Sebastian K. (1971). *Principles of Structural linguistics*, The Hague : Mouton. 359 p.
- SIDDHARTHAN, Advaith (2003). Resolving pronouns robustly: Plumbing the depths of shallowness. *Workshop Computational Treatments of Anaphora 11th EACL*. pp. 7-14.
- SIDIR Mohamed, Nadine LUCAS & Emmanuel GIGUET (2006). De l'analyse du discours à l'analyse structurale des réseaux sociaux : une étude diachronique d'un forum éducatif. *revue STICEF* 13. pp. 289-316. mis en ligne 2007 <http://revuesticef.org>

- SIDIR, M. & N. LUCAS (2007). "L'écriture en réseaux d'un document numérique: de l'analyse de discours aux processus collaboratifs" in F. Papy (Ed.). *Pratiques et usages dans les bibliothèques numériques*. Paris : Hermès. pp. 269-291.
- SOULET, Arnaud (2006). *Un cadre générique de découverte de motifs sous contraintes fondées sur des primitives*. Thèse de l'Université de Caen.
- SOULET, Arnaud (2007). Résumer les contrastes par l'extraction récursive de motifs. *Conférence d'Apprentissage 2007 (CAp 2007)*, Grenoble. pp 339–354.
- SOULET, Arnaud & Bruno CREMILLEUX (2008). Adequate condensed representations of patterns. *Data Mining Knowledge Discovery*, 17. pp. 94-110.
- STIENNE, Nicolas & Nadine LUCAS (2003). Exploitation d'informations disponibles sur Internet et génération d'un portail multilingue sur le cinéma. *Conférence internationale sur le document électronique Cide 6*, In Faure et Madelaine (eds) *Document électronique dynamique*, Paris : Europia, pp. 239-255.
- SWIGGERS, Pierre (1997). *Histoire de la pensée linguistique*. Paris : Presses Universitaires de France. 312 p.
- TAMBA-MECZ, Irène (1988). *La sémantique*. Paris : Presses Universitaires de France. 127 p.
- TERAMURA, H., SAKUMA, M., SUGITO, K., & HIRAZAWA, S. (1990). *Kêsu sutadei nihongo no bunshô-danwa [Etudes de cas, phrases et discours en japonais]*. Tôkyô: Ofusha. 189 p.
- TEUFEL, Simone (1999). *Argumentative zoning* PhD thesis University of Edinburgh, Edinburgh. 352 p.
- TEUFEL, Simone, Jean CARLETTA & Marc MOENS (1999). An annotation scheme for discourse-level argumentation in research articles. *European Association for Computational Linguistics conference*. pp. 110-117.
- TRNKA, Bohumil (1964). On the linguistic sign and the multilevel organization of language. *Travaux linguistiques de Prague 1*. pp. 33-40. Reimpressum 1982 Fried (Ed) *Selected Papers in Structural Linguistics*, Mouton ; 2006 Hajicova & Sgall (Eds) en ligne <http://dlib.lib.cas.cz/3131/>
- TURMEL, Leny, Nadine LUCAS & Bruno CREMILLEUX (2003). Extraction d'associations pour la caractérisation de segments de textes en anglais avec et sans faute. *Cide 6*, In Faure et Madelaine (eds) *Document électronique dynamique*, Paris : Europia, pp. 221-237.
- VACHEK, Josef Ed. (1964). *A Prague school reader in linguistics*. Bloomington, Indiana, USA: Indiana University Press.
- Van DIJK, Teun A. (1972 b). *Some aspects of text grammars*. La Haye : Mouton. 377 p.
- Van DIJK, Teun A. (1973 a). "Grammaire textuelle et structures narratives" in C. Chabrol (Ed.) *Sémiotique narrative et textuelle*, Paris : Larousse. pp. 177-206.
- Van DIJK, Teun A. (1977). "Sentence topic and discourse topic" *Papers in Slavic Philology* (1). pp. 49-61.
- Van DIJK, Teun A. & Walter KINTSCH (1983). *Strategies of discourse comprehension*. New York : Academic Press.
- VAUQUOIS Bernard & Christian BOITET (1985) Automated translation at Grenoble University. *Computational Linguistics*, 11 (1). pp. 28–36.
- VEGA, José & Jacques VERGNE (2005). Mycatex - a language-independent term extractor. *Workshop EU Enlargement and Integration Workshop: Exploiting parallel corpora in up to 20 languages*, Arona, Italie. september 26-27.
- VERGNE, Jacques (1989). *Analyse morpho-syntaxique automatique sans dictionnaire*, Doctorat d'université. Université Paris 6.
- VERGNE, Jacques (1998). "Entre arbre de dépendance et ordre linéaire, les deux processus de transformation: linéarisation, puis reconstruction de l'arbre" *Cahiers de Grammaire* (23) pp. 95-136.
- VERGNE, Jacques (1999). *Étude et modélisation de la syntaxe des langues à l'aide de l'ordinateur : Analyse syntaxique automatique non combinatoire : Synthèse et Résultats* Mémoire d'habilitation à diriger des recherches, Université de Caen. 110 p.

- VERGNE, Jacques (2001). Analyse syntaxique automatique de langue: du combinatoire au calculatoire. *8^e conférence sur le traitement automatique des langues naturelles TALN 2001*, Tours, ATALA. pp. 15-29.
- VERGNE, Jacques (2002). Une méthode pour l'analyse descendante et calculatoire de corpus multilingues: application au calcul des relations sujet-verbe. *TALN 2002*, Batz 2002, ATALA. pp. 63-74.
- VERGNE, Jacques (2005). Une méthode indépendante des langues pour indexer les documents de l'internet par extraction de termes de structure contrôlée. *8^e colloque international sur le document électronique CIDE 8*, Paris : Europia. pp. 155-168.
- VIRBEL, Jacques (1985). Langage et métalangage dans le texte du point de vue de l'édition en informatique textuelle. *Cahiers de Grammaire* (10) pp. 5-72.
- YAMADA, Yoshio 山田孝雄 (1908). 日本文法論 *Nihon bunpôron [Traité de grammaire japonaise]* rééd. 1970, Tôkyô : Hôbunkan. 1141 p.
- YAMADA, Yoshio (1913). 奈良朝文法史 *Narachô bunpô shi [Histoire de la grammaire de l'époque de Nara]*. rééd. 1954, Tôkyô : Hôbunkan. 649 p.
- YAMADA, Yoshio (1913). 奈良朝文法史 *Heianchô bunpô shi [Histoire de la grammaire de l'époque de Heian]*. rééd. 1952, Tôkyô : Hôbunkan. 629 p.
- YAMADA, Yoshio (1922). 日本口語法講義 *Nihon kôgohô kôgi [Cours sur le japonais oral]*. rééd. 1970, Tôkyô : Hôbunkan. 398 p.
- YAMADA, Yoshio (1936). 日本文法学概論 *Nihon bunpôgaku gairon [Somme sur la grammaire japonaise] [Introduction to Japanese Grammar]*. Rééd. 1951, Tôkyô : Hôbunkan. 1174 p.
- YANGARBER, Roman (2003). Counter-Training in Discovery of Semantic Patterns. *41st Annual Meeting of the Association for Computational Linguistics: ACL-2003*, Sapporo, Japan. pp. 343-350.
- YANGARBER, Roman & Lauri JOKIPII (2005). Redundancy-based Correction of Automatically Extracted Facts. *Human Language Technology Conference/ Conference on Empirical Methods in Natural Language Processing: HLT/EMNLP-2005*, Vancouver, Canada. pp. 57-64.
- ZACKLAD, Manuel (2004). Processus de documentarisation dans les Documents pour l'Action (DOPA) : statut des annotations et technologies de la coopération associées. *Le numérique : Impact sur le cycle de vie du document pour une analyse interdisciplinaire*, 13-15 Octobre 2004, Montréal, disponible sur : <http://archivesic.ccsd.cnrs.fr/>
- ZERIDA, Nadia, Nadine LUCAS & Bruno CREMILLEUX (2006a). Combinaison de descripteurs linguistiques et de structure pour la fouille d'articles biomédicaux *Colloque international "Discours et Document"*. juin Caen Schedae 9-1 Presses universitaires de Caen, pp. 69-78.
- ZERIDA, N., N. LUCAS & B. CREMILLEUX (2006b). Combining Linguistic and Structural Descriptors for Mining Biomedical Literature *ACM Symposium on Document Engineering* 10-13 Octobre Amsterdam. ACM, 2006. pp. 62-64.
- ZERIDA, N., N. LUCAS & B. CREMILLEUX (2007). Exclusion-Inclusion based Text Categorization of biomedical articles. *ACM symposium on Document engineering*, Winnipeg, Manitoba, Canada August 28-31, 2007, ACM. pp. 202-204.

Sites

- Cercle linguistique de Prague - Linguistic Circle of Prague - Pražský lingvistický kroužek
<http://www.prague Linguistics.org> ou <http://www.cercledeprague.org>
- Dictionnaire Sens-agent <http://dictionnaire.sensagent.com>
- DITL : Dictionnaire international des termes littéraires / International Dictionary of Literary Terms
<http://www.ditl.info>
- Europa <http://europa.eu/languages/fr/home>

RST <http://www.sfu.ca/rst> pour le site général et <http://www.sfu.ca/rst/07french/index.html> pour la version française, due à Péry-Woodley et collaborateurs.

Silva rhetoricae, Brigham Young university, créé par Gideon O. Burton
<http://humanities.byu.edu/rhetoric/silva.htm>

Signo, Rimouski (Québec) <http://www.signosemio.com>.

Texte ! Paris <http://www.revue-texte.net/>

EXEMPLES

FDJ t20 Anonyme. Des traces de ricine dans des flacons gare de Lyon à Paris. PARIS (AFP), le 21-03-2003.
exemple 1 p. 30 repris exemple 3 p. 45
traitement illustré p. 69 Fig. 3.11

FSV4 Anonyme. Ferme la bouche quand tu manges! *Science et vie* 927 (1994) p. 32.
Exemple 2. p. 30 repris exemple 4 p. 47

FSV12 Gozzo, Jacques. Le point chaud de l'Afar sous surveillance *Le journal du CNRS*, septembre 2001, p. 25.
exemple 5 p. 48
traitement illustré p. 70 Fig. 3.12

FDJ8 Anonyme. FRANCE: Glavany plaide pour une nouvelle PAC où l'on produirait "mieux". Reuters, Paris 31/01/2001.
Exemple 3.1 p. 56 et 3.2 p. 57
traitement illustré p. 59 Fig. 3.3

FQJ41 Anonyme. Collectif budgétaire européen pour faire face à la crise du secteur bovin. Genève *La Tribune* / Reuters Business Briefing, 01/02/2001
Exemple 3.3 p. 57-58

ADJ174 Dobbyn, T. FAA to insist on checks of Boeing plane rudders. 15/06/1998 Reuters, Washington 05:44 p.m Eastern
Exemple 3.4 p. 60
Schématisation p. 61 Fig. 3.4
Traitement p. 61 Fig. 3.5

AJD38 O'Donnell, Kathie. USA: Retail interest in munis ho-hum despite stock woes. Reuters, New York 11/12/1997.
Exemple 3.5 et 3.6 p. 73