FILTRAGE SEMANTIQUE DE TEXTESPROBLEMES, CONCEPTION ET REALISATION D'UNE PLATE-FORME INFORMATIQUE

Jean-Luc Minel

Université Paris-Sorbonne

HABILITATION A DIRIGER DES RECHERCHES

DISCIPLINE : INFORMATIQUE ET SCIENCES HUMAINES

FILTRAGE SEMANTIQUE DE TEXTES PROBLEMES, CONCEPTION ET REALISATION D'UNE PLATE-FORME INFORMATIQUE

DOSSIER SCIENTIFIQUE
PRESENTE PAR

Jean-Luc Minel

CAMS - LaLIC
(Langage Logique Informatique et Cognition)
UMR 8557 - CNRS - EHESS- Université Paris-Sorbonne

Jury :

Michel Charolles, professeur à l'Université Paris III
Jean-Pierre Desclés (directeur), professeur à l'Université Paris-Sorbonne
Yves Jeanneret (rapporteur), professeur à l'Université Paris-Sorbonne
Guy Lapalme (rapporteur), professeur à l'Université de Montréal
Jean-Marie Pierrel (rapporteur), professeur à l'Université Henri Poincaré Nancy 1
Gérard Sabah, directeur de recherches au CNRS

Remerciements

Mes remerciements s'adressent tout d'abord à Jean-Pierre Desclés qui a su, par son accueil au sein de son équipe, par sa confiance et ses nombreux conseils, guider mes recherches dans le domaine de l'informatique linguistique.

Je remercie vivement les rapporteurs, Yves Jeanneret, Guy Lapalme et Jean-Marie Pierrel de m'avoir fait l'honneur d'évaluer ce travail ainsi que Michel Charolles et Gérard Sabah pour leur participation au jury.

J'ai trouvé au sein de l'équipe LaLIC un lieu propice aux échanges, aux discussions et aux collaborations scientifiques. Je remercie toutes celles et tous ceux, qui depuis mon arrivée en 1993, ont contribué à l'élaboration de ce travail.

La plate-forme ContextO a concrétisé de nombreux travaux sur l'exploration contextuelle menés dans l'équipe LaLIC. Le développement logiciel de cette plate-forme est le fruit d'une étroite collaboration avec Slim Ben Hazez et Gustavo Crispino. Qu'ils trouvent ici l'expression de ma sincère gratitude.

Enfin, la rédaction de ce mémoire s'est faite en grande partie à l'Université de la République à Montevideo, sur les rives du Rio de la Plata. Merci à Dina Wonsever et à Javier Couto, enseignant-chercheurs au département du Traitement Automatique du Langage Naturel de cette Université, pour leur accueil chaleureux et attentionné.

Table des matières

Préambule : Des manuscrits aux textes

Ce mémoire est l'aboutissement d'un cheminement qui m'a conduit de l'étude des manuscrits médiévaux jusqu'au filtrage sémantique d'informations dans les textes. Ma formation en informatique, plutôt axée sur l'étude des réseaux d'ordinateurs et des systèmes d'exploitation, me destinait plutôt à rejoindre des laboratoires de recherche du domaine des Sciences pour l'Ingénieur mais différentes rencontres avec des chercheurs des Sciences de l'Homme et de la Société, où mes compétences informatiques pouvaient trouver également des développements, m'ont orienté vers d'autres horizons scientifiques.

Mes premiers travaux de recherche, menés dans le cadre de l'Institut de Recherche et d'Histoire des Textes (IRHT), Unité Propre de Recherche du CNRS, m'ont fait découvrir la complexité de l'étude d'un objet, le livre médiéval. L'étude du livre médiéval nécessite la description d'une

multitude d'informations suivant différents points de vue : certains ont trait à l'objet matériel lui-même, son processus de fabrication et à sa décoration ; d'autres à son contenu, c'est-à-dire aux textes, traductions, commentaires et iconographies qui le composent ; d'autres, enfin, aux bibliothèques anciennes et aux possesseurs qui en assurèrent autrefois la conservation.

Au milieu des années soixante-dix, l'utilisation de l'informatique dans le domaine du livre médiéval en était encore à son tout début et était confrontée à d'importants obstacles. Sur le plan technique, les possibilités de codage des informations étaient particulièrement pauvres (absence de diacritiques et de caractères minuscules notamment) et les capacités de stockage de l'information étaient très faibles au regard des besoins. Sur le plan conceptuel, les recherches sur l'informatisation des données se situaient au carrefour des sciences de l'information, de la philologie, de l'histoire et de l'informatique, d'où la nécessité de faire cohabiter et, si possible, converger des savoirs multiples. Sous l'impulsion de Lucie Fossier, directrice adjointe de l'IRHT, une équipe de recherche réunissant ces diverses compétences s'est attelée à cette tâche. La base de données MEDIUM [10, 17, 18][1] en fut un des résultats les plus significatifs. En 1988, cette base de données sur le manuscrit médiéval, unique en son genre, décrivait plus de 40 000 manuscrits et était interrogeable par les réseaux de télécommunications[2].

L'étude des manuscrits et la construction de MEDIUM m'ont amené à réfléchir et à étudier le problème soulevé par la description de données non homogènes, non normalisées, d'interprétations multiples et parfois contradictoires. Par exemple, les sources historiques peuvent mentionner un même individu de manières extrêmement diverses, soit en raison de variantes orthographiques d'un scribe à l'autre, soit parce que les sources n'indiquent pas toutes les mêmes informations. Je me suis donc intéressé à la modélisation et à la représentation des connaissances d'un historien et d'un documentaliste [12, 13, 16].

Mes activités d'enseignement et de conseil menées en parallèle m'ont amené à collaborer avec l'équipe de recherche de la Direction Informatique de l'Ecole Nationale d'Administration. L'Intelligence Artificielle et, plus précisément les systèmes experts, faisaient alors irruption dans l'activité économique et dans l'administration française. Mes recherches menées avec les historiens de l'IRHT dans le domaine de l'ingénierie des connaissances, notamment en ce qui concerne la modélisation des informations incertaines, trouvèrent là, et plus particulièrement dans le domaine de la réglementation publique [6, 14], un nouveau domaine d'application mais aussi une nouvelle source de problèmes à résoudre. Réciproquement, les travaux menés avec les mathématiciens et économistes [5, 15] de l'équipe de recherche de l'ENA, notamment P. Lévine et J.-C. Pomerol, me permirent

[1] Ces numéros renvoient à la liste des publications (cf. p. 11-14).
[2] Depuis l'an 2000, cette base de données est accessible via Internet.

d'approfondir mes connaissances des modèles utilisés dans l'aide à la décision et plus généralement des méthodes de résolution de problème. Le croisement de ces différentes méthodes et outils de recherche fut sans aucun doute d'une exceptionnelle fécondité en m'amenant à constamment rechercher dans les disciplines connexes des concepts et des méthodes aptes à formaliser, complètement ou partiellement, la représentation des connaissances et les stratégies mises en œuvre pour appliquer ces connaissances.

L'étape suivante, qui s'intégrait dans la politique de valorisation de la recherche conduite par le CNRS, m'a conduit à participer à la modélisation du savoir dans un milieu extra-universitaire. Détaché du CNRS auprès d'un grand groupe industriel (les Ciments Français), j'ai mis en œuvre les méthodologies et les techniques issues des systèmes à base de connaissances. Ces travaux constituaient les prémisses de ce qui est devenu actuellement une activité de recherche à part entière, le « knowledge management ». Sur le plan scientifique, cette période m'a permis d'acquérir et de maîtriser les méthodes de conduite de projet et d'organisation rigoureuse nécessaires à la gestion de projets complexes et multidisciplinaires. En effet, la modélisation des savoirs dans ce domaine industriel faisait intervenir des chimistes, des mécaniciens, des économistes et des informaticiens.

Mon projet, lors de mon retour au sein du CNRS, était de capitaliser mes compétences dans un laboratoire dont les recherches étaient centrées sur l'intelligence artificielle. J'ai ainsi intégré le Laboratoire Formes et Intelligence Artificielle (LAFORIA) pour participer aux recherches sur l'aide à la décision dans l'équipe dirigée par Jean-Charles Pomerol. La direction du département des Sciences de l'Homme et de la Société me fit alors la proposition de travailler au Centre d'Analyse et de Mathématiques Sociales (CAMS). Au sein de ce laboratoire, qui affichait à l'époque clairement ses ambitions interdisciplinaires, J.-P. Desclés poursuivait, depuis déjà de nombreuses années, un ambitieux programme de recherche situé au carrefour de la linguistique, la logique, l'informatique et la cognition, et il souhaitait en renforcer la partie informatique. Bien que n'ayant pas de compétences particulières dans le traitement automatique des langues, j'ai néanmoins été attiré par la perspective de travailler sur des systèmes experts dans le domaine du résumé automatique.

Mon activité de recherche s'est ainsi déroulée au sein de l'équipe LaLIC où j'ai conçu l'architecture de la plate-forme Filtext [1, 7, 8, 19, 22] et guidé les différentes phases de sa réalisation en co-encadrant des doctorants et des étudiants du DEA MIASH. Ma contribution d'informaticien m'a permis d'enrichir la problématique générale du résumé automatique [27, 34] et du filtrage sémantique des textes [2, 20, 21]. Les résultats obtenus au travers de la plate-forme conceptuelle Filtext et de son instance logicielle ContextO, ainsi que les différentes thèses soutenues sur ce sujet depuis lors dans l'équipe LaLIC, démontrent que ce champ de recherche est propice à des recherches théoriques mais

aussi finalisées. L'informatique peut y être clairement articulée avec des descriptions empiriques au sein d'un modèle théorique. Par ailleurs, l'interdisciplinarité de l'équipe LaLIC m'a permis d'élargir mon champ de recherche à des travaux plus théoriques, notamment ceux portant sur le traitement informatique des schèmes sémantico-cognitifs et sur l'approche logique du langage.

Cette rapide description de ma trajectoire scientifique illustre la diversité de mes activités scientifiques qui, il faut aussi le souligner, ont souvent donné lieu à des enseignements dans différentes universités parisiennes. La transmission des connaissances qui nécessite leur mise en forme ainsi que le dialogue noué avec les étudiants constitue en effet à mes yeux un complément indispensable à l'activité de recherche. Reste une question : pourquoi soutenir une habilitation à diriger des recherches ? Il m'a semblé qu'il convenait d'aller au delà de la reconnaissance que m'apporte la publication d'ouvrages scientifiques. Cette habilitation me permettra d'institutionnaliser mon activité d'encadrement de jeunes chercheurs et de poser un jalon académique dans mon parcours professionnel quelque peu atypique. Mais il n'est pas dans mon intention de quitter le CNRS, ma « maison intellectuelle » ; mon statut d'ingénieur de recherche qui marie dans son intitulé et dans la pratique ingénierie et recherche correspond parfaitement à mes attentes.

Ce mémoire se divise en deux parties :

- La première partie se présente comme une rapide synthèse des systèmes de résumé automatique ; elle vise à montrer la genèse du projet Filtext, plate-forme d'ingénierie linguistique dédiée au filtrage sémantique d'informations dans des textes.
- La deuxième partie décrit plus spécifiquement nos propres travaux de recherche et d'encadrement nécessaires à l'élaboration de modèles de représentation des connaissances linguistiques, à la conception et à la réalisation de la plate-forme Filtext au sein d'une équipe pluridisciplinaire. Les travaux de recherche en informatique, et plus particulièrement en linguistique informatique, impliquent un travail en équipe ou s'expriment des compétences multiples. La linguistique informatique, sous la pression notamment des avancées technologiques issues du Web, doit passer du traitement de la phrase au traitement du texte. Ce changement d'échelle des observables nécessite non seulement une collaboration locale entre linguistes et informaticiens mais aussi une collaboration plus globale entre différentes équipes de recherches.

> *« Plus il y a d'incertitudes et d'informations à digérer, plus il est important de travailler avec les autres. Il faut pouvoir créer des équipes qui puissent se lancer des idées,..., je prends des idées aux autres, comme*

eux s'inspirent de moi. Il est nécessaire pour cela de travailler dans la confiance et le respect. »

Symposium de la créativité de Zermatt, déclaration du prix Nobel d'économie 1997, Myron Scholes

Enfin, il m'a semblé important de mettre à l'épreuve les modèles et la plate-forme informatique élaborés en répondant à des appels à propositions. C'est pourquoi je présente deux exemples d'utilisation de la plate-forme ContextO, le projet RAP et le projet[3] « Modèle d'exploration sémantique de textes guidé par les points de vue du lecteur ». Ces expérimentations devraient permettre de mieux cerner les modes de représentation des stratégies de parcours textuel. Elles nécessitent, entre autres, une collaboration étroite avec les chercheurs en sciences humaines et avec les chercheurs spécialisés dans l'étude « *des changements qui affectent actuellement les modes de médiatisation des processus d'information et de communication et en particulier le rôle joué dans cette redéfinition de l'économie médiatique par les dispositifs informatiques* »[4].

[3] Ce projet auquel collaborent le CEA, l'équipe LaLIC du CAMS, le LATTICE et le LIMSI est financé par l'Action Concertée Incitative Cognitique 2000.
[4] [JEA 01].

Liste des travaux

Ouvrages en préparation

• Livre

Minel J.-L. (2002). *Méthodes et outils informatiques pour le résumé et le filtrage automatique des textes*, parution prévue en juin 2002, 200 pages, Editions Hermès.

• Contribution

Desclés J.-P., **Minel J.-L**. (2002). *L'exploration contextuelle*, in *Interpréter en Contexte*, sous la direction de Francis Corblin, parution prévue en 2002, 40 pages, Editions Hermès.

Contribution à des ouvrages[5]

*[1] **Minel J.-L.,** J.-P. Desclés. (2000). *Résumé Automatique et Filtrage des textes*, in *Ingénierie des langues*, sous la direction de Jean Marie Pierrel, Paris, p. 253-270, Editions Hermès.

[2] Berri J., E. Cartier, J.-P. Desclés, A. Jackiewicz, **J.-L. Minel**. (1997). *Safir, système automatique de filtrage de textes*, Langages, Cognition et Texte, vol II., Université Hankuk des Langues étrangères, Université Paris-Sorbonne, Séoul & Paris, p. 3-16.

*[3] Guillaumont A., **J.-L. Minel**. (1992). *Medium* : *Une base de données au service des médiévistes* in *Recherche et Histoire des textes* : *Filmothèques, photothèques et techniques nouvelles*, (édité par G. Contamine), p. 111-119, Editions du Léopard d'or.

[4] Guillaumont A., **J.-L. Minel**. (1991). *Medium : Database for Medieval Manuscripts*, in *Bibliographic Access to Medieval And Renaissance Manuscripts*, (édité Wesley M. Stevens), p. 29-39, The Harworth Press.

[5] Lévine P., **J.-L. Minel**, J.-C. Pomerol, (éditeurs). (1991). *L'intelligence artificielle : philosophie science ou technologie*, 62 pages, n° 657, La Documentation Française.

*[6] Lévine P., **J.-L. Minel**. (1989). *A development Tool for Expert Systems in the field of regulations*, in *Expert Systems in Public Administration*, p. 111-118, Elsevier.

Articles dans des revues avec comité de lecture

*[7] **Minel J.-L.**, J-P. Desclés, E. Cartier, G. Crispino, S. Ben Hazez, A. Jackiewicz. (2001). Résumé automatique par filtrage sémantique d'informations dans des textes. Présentation de la plate-forme FilText, *Revue Technique et Science informatiques*, n° 3, p. 369-395, Hermès.

*[8] Wonsever D., **J.-L. Minel**. (2001). Contextual Rules for Text Analysis, Proceedings de Cicling 2001, *Lecture Notes in Computational Science*, p. 503-517.

[9] Guillaumont A., **J.-L. Minel**. (1991). Medium : Data base for Medieval Manuscripts, *Primary Sources and Original Works*, vol. 1, n° 3/4, p. 29-38, Haworth Press.

---

[5] Les publications précédées d'une étoile (*) figurent dans le dossier des travaux publiés.

[10] Guillaumont A., **J.-L. Minel**. (1990). Medium : Un exemple d'application en histoire des sciences, *in the Use of Computer in Cataloging Medieval Rennaissance Manuscript*, (édité par M. Folker & ali.), p. 57-64, Munich.

[11] Mathieu Y., **J.-L. Minel**. (1987). Prolog, un outil déclaratif, *Informatique et Sciences Humaines,* n° 74, p. 11-17, Paris.

[12] Bourlet C., **J.-L. Minel**. (1987). Artificial Intelligence and the Methodology of History, Manchester University Press.

[13] Bourlet C., **J.-L. Minel**. (1987). Un système déclaratif d'aide à l'identification des individus dans un corpus prosopograhique, *Informatique et Sciences Humaines*, n° 74, p. 49-59, Paris,

[14] Lévine P., **J.-L. Minel**, J.-C. Pomerol. (1986). Systèmes experts et Formation, *Revue Française d'Administration Publique,* n° 37, p.71-79.

[15] Lévine P., **J.-L. Minel**, J.-C. Pomerol. (1986). Utilisation des systèmes experts pour l'aide à la décision, Publication *AFCET*.

*[16] Bourlet C., L. Fossier, A. Guillaumont, **J.-L. Minel**. (1986). Construction d'un prototype de Système Expert dans le domaine historique, *Computer and Humanities*, vol 20, n° 9, p. 273-275, Paradigm Press Ospreg.

*[17] Guillaumont A., **J.-L. Minel**. (1986). Medium, Realities and Projects, *Computer and Humanities*, vol 20, n° 9, p. 269-271, Paradigm Press Ospreg.

[18] **Minel J.-L.,** A. Guillaumont., M.-P. Beaud. (1985). A medieval Manuscript Data Base. *Data Base in the Humanities and Social Sciences*, Paradigm Press Ospreg.

Communications dans les colloques internationaux avec comité de lecture et actes

*[19] Ben Hazez S., J-P Desclés, **J.-L. Minel**. (2001). Modèle d'exploration contextuelle pour l'analyse sémantique des textes, p. 73- 82**,** *TALN 2001,* Tours.

*[20] Ferret O., B. Grau, **J.-L. Minel**, S. Porhiel. (2001). Repérage de structures thématiques dans des textes, p. 163-172, *TALN 2001,* Tours.

[21] Mourad G., **J.-L. Minel**. (2000). Filtrage sémantique du texte, le cas de la citation, *CIDE 2000*, p. 41-55, Lyon.

*[22] Ben Hazez S., **J.-L. Minel**. (2000). Designing Tasks of Identification of Complex Patterns Used for Text Filtering, *RIAO'2000,* p. 1558-1567*,* Paris.

[23] Crispino G., J.-P. Desclés, S. Ben Hazez, **J.-L. Minel**. (1999). ContextO, un outil du projet Filtext orienté vers le filtrage sémantique de textes, *VEXTAL,* p. 361-368*,* Venise. Italie.

[24] Crispino G., J.-P. Desclés, S. Ben Hazez , **J.-L. Minel**. (1999). Architecture logicielle de Context, plate-forme d'ingénierie linguistique, *TALN 99,* p. 327-332*,* Cargèse.

[25] Crispino G., J.-P. Desclés, S. Ben Hazez , **J.-L. Minel**, G. Mourad. (1999). ContextO : una plataforma de ingeniería lingüística orientada al filtrado semántico de textos,

SEPLN, Procesamiento de Lenguaje Natural, Revista nº 25, p. 215-216*,* Lleida, Espagne*.*

[26] Battistelli D., J.-P. Desclés, C. Valliez, **J.-L. Minel**. (1997). Building a Sequence of Images from a Text, *NLPRS97*, p. 381-386, Phuket, Thailande.

*[27] **Minel J.-L.,** S. Nugier, G. Piat. (1997). How to Appreciate the Quality of Automatic Text Summarization*, 35th Annual Metting of the ACL, Workshop Intelligent Scalable Text Summarization,* p. 25-30, Madrid, Espagne.

[28] **Minel J.-L.**, S. Nugier, G. Piat. **(**1997). Comment apprécier la qualité des résumés automatiques de textes : les exemples des protocoles FAN et MLUCE et leurs résultats sur SERAPHIN, *1° JST FRANCIL 97*, p. 227-232, Avignon.

[29] Cartier E., J.-P. Desclés., A. Jackiewicz, **J.-L. Minel**. (1997). Filtrage automatique de textes : l'exemple des énoncés définitoires, *1° JST FRANCIL 97*, p. 183-189, Avignon.

[30] J.-P. Desclés, E. Cartier, A. Jackiewicz, **Minel J.-L**. (1997). Textual Processing and Contextual Exploration Method, *CONTEXT'97*, p. 189-197, Rio de Janeiro, Brésil.

*[31] Berri J., E. Cartier, J.-P. Desclés, A. Jackiewicz, **J.-L. Minel**. (1996). A linguistic method for text filtering, *SEPLN 96* , p.159-167, Séville, Espagne.

[32].J. Berri, E. Cartier, J.-P. Desclés, A. Jackiewicz, **Minel J.-L**. (1996). Filtrage Automatique de textes, *Natural Language Processing and Industrial Applications*, p. 28-35, Moncton, Canada.

[33] Berri, J., E. Cartier, J.-P. Desclés, A. Jackiewicz , **Minel J.-L**. (1996). SAFIR, système automatique de filtrage de textes, *Actes du colloque TALN'96*, p. 140-149, Marseille.

*[34] Berri J., D. Le Roux, D. Malrieu, **J.-L. Minel**. (1995). SERAPHIN, système à base de connaissances pour l'extraction de phrases, *JADT 95, 3°Journées internationales d'Analyse statistique des données textuelles,* p. 345-354, Rome, Italie.

[35] **Minel J.-L**., J. Berri, D. Le Roux, D. Malrieu. (1995). SERAPHIN : un système d'extraction automatique d'énoncés importants, *Actes des Journées de Génie Linguistique,* p. 409-419*,* Montpellier.

[36] Le Roux D., **J.-L. Minel**, J. Berri. (1994). SERAPHIN project : the industrial approach, *Cognitive Science in Industry.*, p. 275-284, Luxembourg.

[37] Guillaumont A., **Minel J.-L.** (1984). Medium : Base de données sur le manuscrit médiéval, Automatic Processing of Art History, data and documents, Pise, Italie.

Communications dans les colloques sans actes

[38] **Minel J.-L.** (1999). Présentation de la plate-forme ContextO, *Journée Outils pour le Traitement automatique des langues organisée par* l'*ATALA*, Paris.

[39] Desclés J-P, **J.-L. Minel**. (1997). Résumé automatique et filtrage*, Colloque sur les nouveaux paradigmes dans l'analyse du discours*, Université Paris-Sorbonne.

[40] Berri J., D. Le Roux, D. Malrieu, **J.-L. Minel**. (1995). SERAPHIN sentence extracting system for text skimming, *Text Analysis and Computers,* Manheim, Allemagne.

[41] Berri J., D. Le Roux , D. Malrieu, **J.-L. Minel**. (1995). SERAPHIN main sentences automatic extraction system, *Second Language Engineering Convention,* Londres, Grande Bretagne.

[42] **Minel J.-L.,** J. Berri, D. Le Roux, D. Malrieu. (1995). SERAPHIN un système d'extraction automatique d'énoncés importants, *Actes des Journées d'études de la Société Fançaise de Bibliométrie Appliquée,* Ile Rousse.

[43] **Minel J.-L.** (1988). Un Système d'aide au contrôle des prestations sociales, Convention informatique, CNIT, Paris.

[44] **Minel J.-L.,** A. Guillaumont. (1987). Conception, réalisation et exploitation d'une base de données, Colloque de Nimègue, Pays-Bas.

Articles et rapports internes

[45] **Minel J.-L**., J. Berri , J-P. Desclés, D. Malrieu. (1995). Le résumé automatique par exploration contextuelle, Rapport du CAMS 95/1, 25 pages.

[46] **Minel J.-L**. (1992). Génie Cognitif dans les systèmes de supervision, Rapport Technique. Centre de Recherche des Ciment Français, 35 pages, Guerville.

Organisation de colloques

Co-organisateur avec J.-P. Desclés (Université Paris-Sorbonne) et Ben Hamadou (Université de Sfax) de *RIFRA'98, Rencontre Internationale sur l'Extraction, le Filtrage et le Résumé Automatiques*, Sfax, Tunisie, 1998.

Organisation de journées d'études

Co-organisateur avec J-P Desclés (Université Paris-Sorbonne) d'une session de présentation sur les Nouvelles Technologies, organisée par le GRECO, Novembre 1997.

Co-organisateur avec J-P Desclés (Université Paris-Sorbonne) d'une *Journée d'étude sur le résumé de textes* dans le cadre des journées organisées par l'ATALA, Paris, Juin 1998.

Co-organisateur avec J-P Desclés (Université Paris-Sorbonne) et S. Chaudiron (Ministère de la Recherche, DISTNB) d'une *Journée d'étude sur le résumé et le filtrage automatique de textes,* Université Paris Sorbonne et Ministère de la Recherche, Paris, Mai 1996.

Journée d'étude sur "*SERAPHIN, système d'activité résumante*", co-organisée avec la DER de l'EDF, Paris, Mai 1996.

Conférences invités

Conférence à l'Université Paris-Sorbonne dans le cadre de la journée « Informatique et Sciences Humaines » organisée par l'Ecole Doctorale « Concepts et langages », juin 2001.

Conférence à l'Université Paris 3, séminaire « cadre et discours », février 2001.

Conférence au Département d'Informatique et de Recherche Opérationnlle (DIRO) de l'Université de Montréal, Novembre 2000.

Conférence à l'ENST Bretagne dans le cadre de la journée « Des industries de la langue à la fouille des textes », février 1999.

Première Partie

Résumé automatique : méthodes et évaluation

Table des matières

Chapitre 1

Résumé automatique : méthodes et évaluation

1. Introduction

Notre propos, dans cette partie, n'est pas de proposer un panorama des systèmes de résumé automatique[6], mais plutôt de montrer comment nos propres travaux de recherche sur le résumé automatique, à travers notamment l'encadrement de la thèse de J. Berri [BER 96a], nous ont amené à orienter notre propre problématique vers le filtrage sémantique et la réalisation de la plate-forme FiltexT. Nous commencerons d'abord par exposer brièvement comment la notion de résumé est généralement appréhendée par les chercheurs du domaine puis nous montrerons comment ont évolué les recherches dans ce domaine, en liaison avec les bouleversements qu'a connu, ces dernières années le traitement automatique du langage naturel.

Remarquons tout d'abord que la notion de résumé de textes, n'est pas neuve ; les premières traces de résumés ont été repérées, d'après [SOL 68, WIT 73][7], sur des tablettes de la civilisation sumérienne en Mésopotamie vers 3600 ans avant notre ère. Certaines tablettes rassemblent en effet les informations d'autres tablettes et on peut donc les considérer comme les premières traces écrites de résumés. Paradoxalement, malgré cet usage très ancien, le concept de résumé ne fait pas l'objet d'une définition très rigoureuse. De nombreux travaux [LER 92, SPA 93] en ont ainsi proposé des typologies, en fonction de leurs contenus mais aussi de leurs modes de production ; on trouve par exemple les définitions suivantes pour divers types de résumés envisagés :

[6] On trouvera dans [BER 96A], [MAS 98] et [MAN 99, 01] un tel panorama.

[7] Cités par [LEH 95]

- le *résumé informatif :* donne une information générale sur le contenu du texte en reprenant les éléments essentiels de celui-ci. Le résumé d'auteur ou « abstract » que l'on trouve au début d'un article scientifique est un bon exemple de cette catégorie de résumé ;
- le *résumé indicatif* : il couvre l'ensemble des thèmes développés dans le texte ;
- le *résumé des conclusions* : il a fait l'objet d 'une norme ISO et est défini comme « … un bref exposé dans un document (généralement placé à la fin de ce document) de ses découvertes et de ses conclusions caractéristiques et qui a pour but de compléter l'orientation du lecteur qui a étudié le texte précédent... » ;
- le *résumé critique* : il combine condensation du texte source et apport critique sur le contenu de ce texte ;
- le *résumé synthétique* : il synthétise le contenu de plusieurs textes ;
- le *résumé scolaire* : il obéit à des critères de fabrication précis [CHA 89], comme un taux de réduction normé, l'interdiction d'emprunts ou le recours systématique aux synonymes. En fait, ce type de résumé a pour fonction de vérifier la capacité d'un élève à reformuler les thèmes du texte source ; le résumé scolaire n'est donc pas destiné à être lu par un utilisateur potentiel, c'est un instrument de contrôle des connaissances.

Il a été proposé par ailleurs [MAS 98] une classification des résumés s'appuyant sur six critères : la concision, la couverture, la balance, la fidélité, la cohésion et la cohérence. Ces critères restent peu quantifiables puisque très liés à l'interprétation du lecteur. Il n'existe en fait pas de critères rigoureux, c'est à dire indépendants du lecteur, autres que des critères de mise en forme (utilisation d'un langage stéréotypé, nombre maximum de mots imposés, etc.), qui permettraient de catégoriser un résumé ; il est par exemple assez difficile d'établir une distinction très nette entre un résumé indicatif et informatif. Toutes les tentatives de typologie dissimulent finalement une difficulté théorique : nous ne savons pas définir à partir de critères linguistiques le résumé, car il n'existe pas non plus de définition, qui fasse l'unanimité, de ce que nous appelons communément un texte.

Nous allons montrer comment les systèmes de traitement automatique ont cherché à produire malgré tout des résumés, en dégageant ce qui nous semble être deux tendances actuelles des travaux menés dans ce domaine, à savoir, l'appel à des connaissances linguistiques d'une part, et d'autre part la volonté de mieux cibler les besoins des utilisateurs des résumés produits.

Chapitre 2

Méthodes de résumé automatique

Différentes méthodes ont été développées au cours des trente dernières années pour produire automatiquement un résumé à partir d'un texte. Ces méthodes peuvent être classées en deux groupes : les approches par compréhension et les approches par extraction. Cette dichotomie correspond d'ailleurs à l'évolution historique de ces systèmes de résumé automatique puisque le renouveau de ceux-ci, se caractérise entre autres, par l'abandon de l'approche par compréhension [SPA 99].

2.1. Méthodes fondées sur la compréhension

Un large courant de recherche, prenant ses racines dans l'étude de la cognition dont la préoccupation centrale est de caractériser les processus complexes de compréhension, considère l'activité résumante comme une activité qui doit nécessairement passer par la compréhension du texte source. L'activité résumante fournit ainsi un terrain d'expérimentation pratique qui permet de tester des modèles de compréhension et de représentation des connaissances.

Pour ce courant de recherche, résumer un texte se ramène à la succession de trois étapes fondamentales (cf. fig. 1) : le texte à traiter est un texte en langage naturel ; il peut aussi être le résultat d'une analyse syntaxique ou bien encore être constitué d'un ensemble de propositions. À partir de ces données, un module se charge de construire une représentation du texte. La forme de cette représentation varie selon les approches : il peut s'agir d'une représentation causale des événements du texte, démarche adoptée par R. Schank [SCH 75] pour qui la représentation des textes narratifs est une chaîne causale dont les nœuds correspondent aux événements du texte et les arcs représentent les relations causales ; pour d'autres auteurs [KIN 78, ALT 90], la représentation du texte est un graphe cohérent constitué par une séquence ordonnée de propositions de type prédicat-argument(s), les liens établis entre propositions correspondant aux arguments communs aux propositions en question. On pourrait aussi citer d'autres formes de représentation dans la tradition de l'Intelligence Artificielle

comme les frames [MIN 75], les scripts [SCH 77], les schémas [RUM 75], etc., (pour plus de détails voir par exemple [ALT 91]).

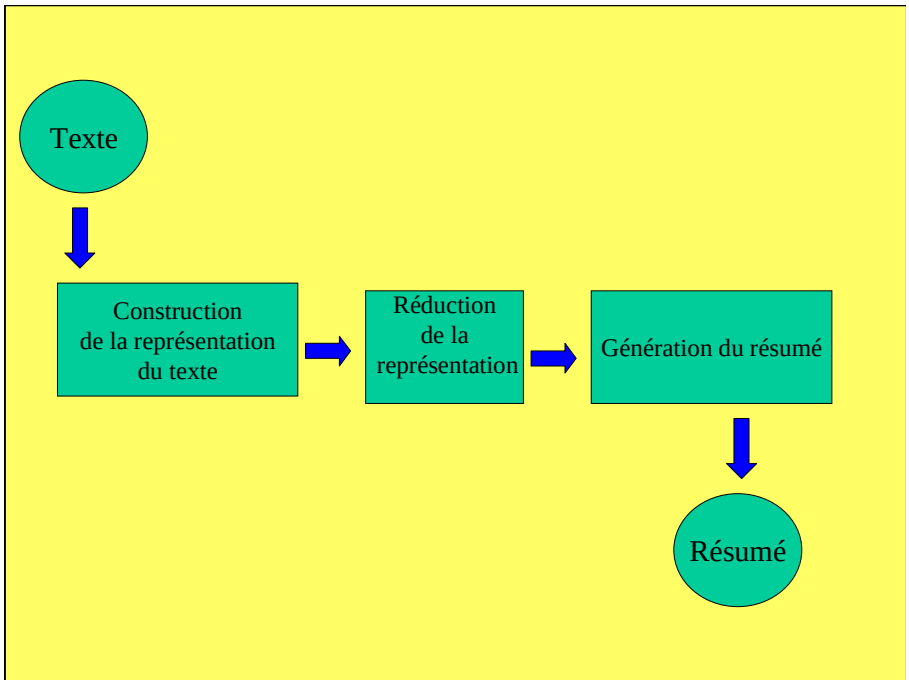


Figure 1. Comprendre pour résumer.

Une fois la représentation du texte construite, celle-ci est utilisée par un module qui procède à sa réduction au moyen d'une série d'opérations. Ces opérations de réduction, différentes selon les modèles, se fondent sur certaines hypothèses concernant l'importance des parties de la représentation à retenir pour le résumé final. Pour Schank (op. cit.), les événements les plus intéressants sont ceux qui correspondent à une succession narrative d'événements reliés par des liens causaux (cette succession est appelée le chemin critique de la représentation). Pour Kintsch et Van Dijk [KIN 75, 78, VAN 83], la représentation du résumé correspond à la macrostructure sémantique du texte qui caractérise un niveau global de la structure sémantique du discours : elle consiste en un réseau de propositions inter reliées et c'est cette structure qui capture la notion intuitive de ce qui est essentiel dans le discours. L'importance est évaluée quantitativement dans d'autres approches : par exemple, Alterman et Bookman [ALT 90] considèrent l'importance d'un événement comme une fonction du nombre de nœuds associés à cet événement dans la représentation ; il s'agit alors de fixer un seuil quantitatif d'importance à partir duquel certains événements les plus importants seront sélectionnés.

Le résultat obtenu à l'issue de cette étape est une représentation réduite aux parties les plus importantes de la représentation du texte initial. L'étape suivante consiste à générer un texte en langage naturel à partir de celle-ci. Ce texte est alors considéré comme le résumé du texte initial.

Afin de donner un aperçu plus détaillé de ces recherches, nous allons présenter brièvement deux approches : tout d'abord l'approche de Kintsch et van Dijk [KIN 75, 78, VAN 83], qui a marqué ce courant de recherches et qui a été très utilisée notamment par les psycholinguistes pour tester des hypothèses empiriques sur les capacités des sujets à résumer un texte ; puis l'approche de Alterman [ALT 90, 91] qui se situe dans la continuité des travaux de Kintsch et van Dijk, mais dans une perspective orientée Intelligence Artificielle, et dont la tentative a été concrétisée par l'implémentation effective d'un système de résumé nommé SSS.

2.1.1. Modèle de Kintsch et van Dijk

Le modèle de compréhension élaboré par Kintsch et van Dijk se situe essentiellement dans le courant de la psychologie cognitive ; il a pour vocation de décrire les opérations mentales sous-jacentes à l'activité de compréhension de textes.

La représentation macro-propositionnelle construite par un sujet lors de la compréhension d'un texte constitue le résumé du texte qui apparaît alors comme le produit automatique de l'activité de compréhension. Avant de décrire les différentes phases du traitement d'un texte, nous donnons un aperçu sur les idées directrices de ce modèle qui se caractérise par trois composantes principales ([pour plus de détails voir [DEN 85, FAY 89]) :

- la micro et la macrostructure sémantiques,
- le schéma du discours,
- et l'organisation de la mémoire en deux parties : la mémoire à court terme et la mémoire à long terme.

2.1.1.1 Micro et macrostructure

Pour Kintsch et van Dijk, la structure de surface d'un texte est un ensemble de propositions. Cet ensemble, appelé base de texte, est organisé par différentes relations sémantiques. Certaines relations sont explicitement exprimées dans la structure de surface, d'autres sont inférées pendant le processus d'interprétation.

La microstructure est le niveau local du discours ; c'est la structure qui comprend les propositions et leurs relations. La macrostructure constitue un niveau plus global du discours. Ce niveau se justifie par le fait que les propositions de la base de texte doivent être reliées au « topic » ou thème du discours. La macrostructure peut être vue comme un réseau de propositions, au sens de la logique des prédicats du premier ordre, associées entre elles en fonction de leur position hiérarchique. L'élaboration de ce réseau hiérarchisé s'obtient par la mise en œuvre, au cours même du traitement, d'un certain nombre d'opérations (appelées macro-opérations) : l'élimination des détails et

redondances, la substitution ou la généralisation d'éléments super-ordonnés à des listes d'objets ou d'événements, la construction et l'intégration des informations en un tout.

La macrostructure d'un texte ne retient finalement que les informations importantes du texte et les relations entre celles ci. Elle nécessite toutefois, pour être réalisée, le guidage par un schéma du discours, appelé aussi grammaire de texte.

2.1.1.2. Schéma du discours

Des exemples typiques de schémas sont la structure d'une histoire, la structure d'une argumentation ou bien la structure d'un article dans un domaine spécialisé. Une structure est spécifiée par un ensemble de catégories. Par exemple, le schéma du discours d'un article en psychologie expérimentale consiste en un ensemble de catégories : introduction, expérimentation, méthode, résultats, et enfin discussion.

2.1.1.3. Mémoires à court et à long terme

Lors du traitement d'un texte, le lecteur dispose de deux ressources : la mémoire à court terme et la mémoire à long terme, qui permettent toutes deux de stocker les propositions du texte.

La mémoire à court terme est une mémoire de travail. Elle a une capacité limitée à un nombre réduit de propositions qu'elle peut stocker. La valeur de ce nombre est fixée à partir de considérations empiriques. Les propositions retenues dans cette mémoire sont considérées comme importantes car elles jouent un rôle de pivot en assurant la cohérence entre propositions de deux cycles (cf. 2.1.1.4) adjacents au cours du traitement.

La mémoire à long terme, quant à elle, n'est pas limitée. Toutes les propositions qui ne sont pas retenues dans la mémoire à court terme sont stockées dans la mémoire à long terme. Dans certains cas, par exemple lorsque la tentative de relier les propositions d'un cycle donné avec les propositions de la mémoire à court terme échoue, une recherche est lancée pour trouver des liens avec les propositions de la mémoire à long terme.

2.1.1.4 Traitement d'un texte

Le traitement d'un texte est divisé en cycles. Des segments de texte, dont la taille correspond approximativement à la phrase sont traités à chaque cycle. Chaque cycle de traitement aboutit à la construction d'un ensemble limité de propositions organisé de manière hiérarchique en un graphe cohérent. Si l'on adopte la stratégie du "bord d'attaque", privilégiée par les auteurs, quelques propositions sont sélectionnées en fonction d'un double critère : leur hauteur dans la hiérarchie et leur récence. Ces propositions sont maintenues en mémoire à court terme pour assurer la liaison (par chevauchement d'arguments) entre les propositions de deux cycles adjacents. Si la cohérence inter

propositionnelle ne peut être établie, une recherche en mémoire à long terme intervient alors pour trouver, dans les propositions antérieurement construites, une proposition qui puisse assurer la liaison. En cas d'échec, une inférence doit être produite. Ce processus a pour résultat l'élaboration de la microstructure du texte ou base de texte [SPR 80]. Kintsch et Van Dijk précisent que cette base de texte est représentée sous la forme d'un ensemble de n-uplets de propositions du type prédicat arguments, qu'ils qualifient d'intuitive. Par ailleurs, ils distinguent la base de texte implicite de la base de texte explicite, censée comprendre toutes les présuppositions et implications réalisées par le lecteur[8]. La microstructure ainsi construite fait l'objet d'un second traitement qui s'effectue sous le contrôle du schéma du discours. Par application des règles de sélection, de généralisation et de condensation de l'information sémantique, ce second traitement aboutit à l'élaboration de la macrostructure sémantique qui correspond au résumé du texte, (voir [KIN 78] où un exemple est entièrement traité).

Ce modèle a subi deux évolutions. La première, proposée en 1983, consiste en une modification des structures de représentation et des processus qui permettent de passer d'un niveau à un autre. Trois niveaux de représentation sont distingués : une représentation de la surface du texte, une base de texte propositionnelle et un modèle de situation. Par ailleurs, le passage de la microstructure à la macrostructure n'est plus réalisé par des règles mais par la mise en œuvre de stratégies. La deuxième évolution, exposée dans [KIN 88] a conduit à un nouveau modèle appelé « construction-intégration ». Dans ce modèle, le processus de construction met en œuvre un système de production dans lequel les règles opèrent sur différents niveaux de représentation. On trouvera dans [EHR 93] une critique détaillée de ce modèle.

Comme on le voit, Kintsch et Van Dijk ont surtout cherché à construire un modèle qui explicite la construction d'un résumé sans chercher, mais tel n'était pas leur propos, à automatiser les procédures de construction. Remarquons par ailleurs que leur approche postule l'utilisation de connaissances tant linguistiques qu'encyclopédiques ces dernières se situant au delà des capacités des systèmes actuels de traitement informatique. Aussi, si plusieurs équipes ont cherché à implémenter ces modèles, aucune d'entre elles n'a réussi à exhiber un système opérationnel.

2.1.2. L'approche de Alterman

À la différence des recherches en psychologie cognitive qui s'intéressent à la modélisation des processus de compréhension d'un lecteur, les approches en intelligence artificielle tentent de modéliser, à l'aide de représentations des connaissances, la sémantique contenue dans le texte à résumer. L'approche développée par Alterman se situe dans ce courant. Pour cet auteur, l'activité résumante consiste à réduire le texte, appréhendé par le biais de représentations sémantiques

[8] Un exemple détaillé est donné par L. Sprenger-Charolles dans [SPR 80].

préalablement construites, en appliquant une série de réductions et de simplifications en ses points les plus importants.

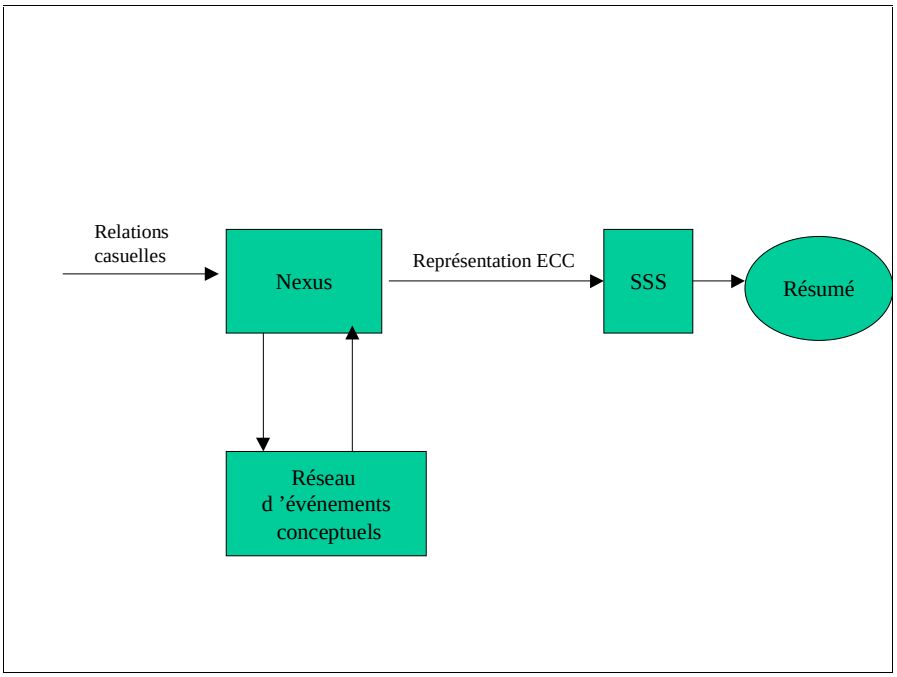


Figure 2. Résumer un texte [ALT 90]

La notion d'épaisseur d'un texte constitue la notion fondamentale mise en oeuvre dans le modèle d'activité résumante proposé par Alterman. L'épaisseur représente la richesse et la complexité des relations entre les événements d'un texte. Résumer un texte revient à simplifier (cf. fig. 2) la représentation d'un texte, c'est-à-dire à produire « un texte squelettique » à partir d'une représentation ECC (Event Concept Coherence pour *cohérence entre événements conceptuels)* fournie par un module appelé NEXUS.

Pour mesurer l'épaisseur d'un texte, une méthode quantitative est développée : elle s'appuie sur une mesure de la charge de travail sachant que Alterman postule que la compréhension d'un texte épais requiert une charge de travail plus importante qu'un texte moins épais.

2.1.2.1 La représentation ECC (cohérence entre événements conceptuels)

La représentation ECC est fournie par NEXUS qui est un module informatique dont l'entrée est une représentation casuelle (selon la théorie de Fillmore [FIL 68]) des événements décrits dans le texte. NEXUS recherche les intersections entre les instances des événements conceptuels pour construire une représentation ECC. Tout d'abord, il recherche un chemin[9] entre les événements

[9] Un chemin est une relation, éventuellement à travers plusieurs clauses (instances d'événements conceptuels), d'un même

conceptuels déjà introduits dans le texte[10] et l'événement conceptuel introduit par l'étape courante. Ensuite, si un chemin est trouvé, NEXUS propage ses connaissances dans le réseau des événements. Par exemple, soit la séquence « Jean lança la brique. Elle frappa Marie. ». NEXUS va traiter en entrée :

[LANCER (agent JEAN) (objet (BRIQUE (déterminant LA)))]
[FRAPPER (objet1 ELLE) (objet2 MARIE)]

Le chemin trouvé entre les deux concepts d'événements "LANCER" et "FRAPPER" est :

[conséquent LANCER MOUVEMENT ((même objet objet) ...)]
[antécédent FRAPPER MOUVEMENT ((ou (même objet1 objet) (même objet2 objet)) ...)]

Cette notation signifie : « comme conséquence au lancer d'un objet, cet objet est nécessairement en mouvement. Si deux objets entrent en collision (l'un frappe l'autre), il est nécessaire que l'un parmi les deux ait été en mouvement ».

[conséquent (LANCER (agent JEAN) (objet (BRIQUE (déterminant LA))))
(MOUVEMENT (objet (BRIQUE (déterminant LA)))]
[antécédent (FRAPPER (objet1 (BRIQUE (déterminant LA)))) (objet2 MARIE)
(MOUVEMENT (objet (BRIQUE (déterminant LA))))]

Cette notation peut être lue de la façon suivante : « comme conséquence au lancement de la brique par Jean, la brique est en mouvement. Un antécédent du fait que la brique frappa Marie est que la brique était en mouvement ».

A partir de cette recherche de chemin entre événements conceptuels, NEXUS forme des parties cohérentes dans la représentation, toutes constituées par des événements conceptuels reliés entre eux. Le tout forme un graphe qui représente l'interprétation du texte en entrée.

2.1.2.2. Etapes pour résumer un texte

Le processus de résumé consiste d'abord à délimiter les frontières et à extraire les événements conceptuels les plus généraux de la représentation produite par NEXUS. Un arbre conceptuel, forme un ensemble d'événements reliés qui dans la représentation ECC représente implicitement un concept délimité.

Soit I une interprétation ECC d'un texte T ; par hypothèse, les événements conceptuels les plus généraux de I sont les racines de chaque partie de I. Ces racines forment l'ensemble minimal couvrant toute l'interprétation et recouvrent par conséquent tout le texte. Deux exceptions s'opposent à l'extraction des racines des événements conceptuels : la première concerne les événements isolés que NEXUS n'a pas pu connecter avec un autre événement ; la seconde concerne les relations de

réseau ECC.

[10] Ces événements peuvent être explicites (introduits explicitement par l'utilisateur dans NEXUS) ou implicites (ajoutés par NEXUS lors de la construction de la représentation ECC pour des raisons de cohérence).

classe/sous-classe, le module SSS choisira alors la sous-classe car elle est considérée comme plus informative que la classe mère.

Le module SSS sélectionne ensuite, parmi l'ensemble des événements racines, ceux considérés comme étant les plus importants. Pour ce faire, SSS évalue l'importance d'un événement (explicite ou implicite) en mesurant le nombre d'inférences réalisées à partir d'un événement conceptuel donné[11]. Un événement est donc d'autant plus important que l'auteur du texte l'aura lié à d'autres événements. En d'autres termes, plus un événement est développé par l'auteur, plus il est considéré comme important par SSS puisque celui-ci est relié à un nombre important de nœuds dans la représentation ECC. Nous verrons que cette hypothèse a été reprise dans son principe dans les systèmes de résumé par extraction qui sont fondés sur le repérage des chaînes lexicales. Le module SSS reprend la liste des événements produits par la technique de délimitation et d'extraction et supprime les événements dont l'importance est inférieure à la moyenne des coefficients d'importance des événements.

Le processus de résumé revient donc à :

- générer la liste des événements conceptuels les plus généraux par délimitation et extraction ;
- déterminer l'importance relative des événements de cette liste en utilisant la mesure d'importance telle qu'elle a été décrite précédemment ;
- enlever de la liste les événements dont le coefficient d'importance est inférieur à la moyenne des coefficients d'importance.

Le résultat de ce processus est une représentation qui garantit la couverture du texte et la présence des événements les plus importants. À la fin de ce processus, la représentation résultante est traitée par un module de génération de textes, inclus dans le module SSS, qui produit des résumés en sortie.

Dans l'approche d'Alterman, La notion d'événement est centrale ; c'est sans doute la raison pour laquelle le système informatique réalisé résume surtout des textes courts composés d'une dizaine de phrases et décrivant une série d'événements factuels. D'une certaine manière, on peut le voir comme un des premiers systèmes d'extraction d'informations (cf. deuxième partie, §1.2).

2.1.3. Synthèse sur l'approche par compréhension

Les approches brièvement présentées précédemment considèrent l'activité résumante comme une activité de compréhension, celle-ci étant assimilée à une construction de représentations. Elles se caractérisent par la volonté de simuler un résumeur spécialiste d'un domaine donné qui lit un texte, le comprend, et en rédige un résumé. Dans un contexte finalisé, cette façon d'appréhender l'activité résumante est confrontée à trois problèmes majeurs qui sont à l'origine de leurs limites :

[11] L'importance est le nombre de nœuds associés à un événement conceptuel donné dans la représentation.

- Tout d'abord l'approche par compréhension reste très liée à un domaine particulier nécessitant donc des représentations et des connaissances spécifiques à ce domaine ; par exemple, elle a été utilisée pour traiter des textes courts, de un paragraphe à une page au maximum comme des dépêches d'agence de presse [DEJ 82].

- Le deuxième problème est plus général et concerne le traitement automatique du langage naturel : construire une représentation sémantique d'un texte est un travail qui nécessite de développer des modèles conceptuels, des ressources linguistiques et des outils informatiques qui, même si certains d'entre eux sont disponibles dans des laboratoires de recherche, n'ont pas atteint la maturité nécessaire à une industrialisation.

- Enfin, la représentation sémantique qu'il convient de construire doit refléter les relations importantes et celles qui le sont moins, entre les différentes parties d'un texte. Sans cette structure, un système de résumé sera dans l'incapacité de distinguer, lors de la phase de réduction, entre ce qui est important et ce qui ne l'est pas. Dans l'approche par compréhension, la notion d'importance est considérée du point de vue de l'auteur, alors que selon nous, elle doit être considérée du point de vue du lecteur du résumé.

Il faut enfin souligner que ces méthodes font implicitement l'hypothèse qu'un texte est toujours bien écrit, au sens où il respecte les règles de syntaxe et d'orthographe, et qu'il est bien structuré thématiquement. L'expérience montre combien cette hypothèse est forte, aussi bien pour les textes de rapport technique, qui ne sont bien souvent pas relus, que pour les textes de la presse quotidienne.

2.2. Méthodes par extraction

Face aux limites des méthodes par compréhension, un autre courant de recherche, que nous désignons sous le terme de « méthodes par extraction », a entrepris de contourner les difficultés précédemment décrites en évitant le processus de construction de représentation et de génération de textes. Ces méthodes mobilisent des ressources linguistiques beaucoup plus légères, ce qui leur permet de traiter des textes longs, de différents domaines et avec des temps de traitement acceptables pour l'utilisateur. Toutes ces méthodes partagent un certain nombre de caractéristiques, dont la chaîne de traitement générique est illustrée en figure 3 :

- elles sont fondées sur l'hypothèse qu'il existe, dans tout texte, des unités textuelles saillantes ; les unités textuelles considérées sont, en général, soit la phrase, soit un ensemble de phrases liées entre elles par des liaisons discursives, soit encore le paragraphe ;
- elles utilisent un algorithme de sélection fondé sur des connaissances statistiques ou linguistiques, ou encore sur des heuristiques combinant différents types de connaissances ; l'application de cet algorithme permet de sélectionner une liste d'unités textuelles ;

- elles construisent le résumé à partir de cette liste, en respectant l'ordre dans lequel les unités apparaissent dans le texte source tout en veillant à ne pas dépasser un certain nombre total d'unités textuelles. Ce seuil est souvent proportionnel à la taille du texte source, comme c'est en général le cas des résumés produits par des professionnels (sachant qu'un seuil de 20% est considéré comme une norme dans les sciences de l'information).
- certaines de ces méthodes, comme on le verra, cherchent à améliorer la lisibilité du résumé en contrôlant la cohérence et la cohésion de celui-ci.

Nous allons présenter différents algorithmes de sélection d'unités textuelles qui s'appuient sur des calculs de fréquences, sur des connaissances linguistiques, ou sur des heuristiques qui tentent de combiner plusieurs types de traitement.

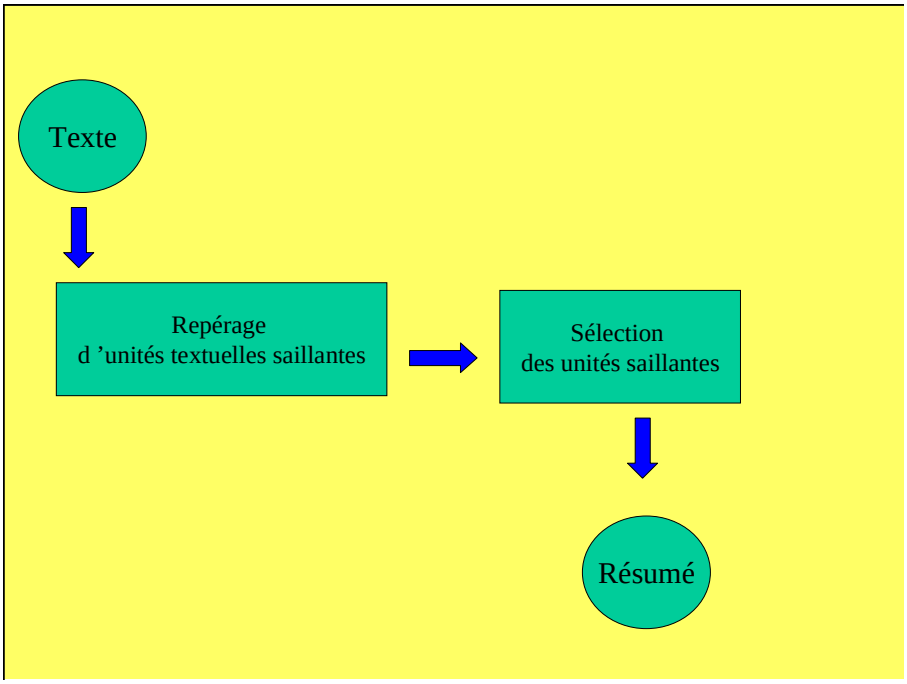


Figure 3. Extraire pour résumer.

2.2.1. Sélection d'unités textuelles par calcul de score

Ce type de méthode consiste à calculer un score S pour chaque unité textuelle, en général la phrase, puis à conserver les unités dont le score est supérieur à un certain seuil, ou à fixer un nombre absolu d'unités à garder en fonction d'un pourcentage de réduction. Le score le plus couramment utilisé est une fonction de la fréquence du mot dans le texte ; il est issu des techniques utilisées dans les sciences de l'information.

Ce score, appelé *tf*idf*, se calcule, pour chaque mot M du texte T à résumer, de la manière suivante :

$score(M) = F_{local} * Log (100 *N / F_{lglobal})$
F_{local} est la fréquence du mot M dans le texte T ;
$F_{lglobal}$ est la fréquence du mot M dans le corpus de référence ;
N le nombre de textes du corpus de référence ;

Un score est ensuite attribué à chaque unité textuelle par addition des scores de chacun des mots contenus dans celle-ci :

$$score (\text{unité textuelle}) = \sum score(M)$$

Ce calcul nécessite, d'une part, la constitution d'un corpus de référence et, d'autre part, la constitution d'une liste de « mots vides ». Il s'agit ainsi d 'éviter qu'un mot nécessairement fréquent dans un texte augmente le score de l'unité textuelle qui le contient. Il y a deux types de mots vides : ceux dont la présence vient du respect des règles syntaxiques, comme par exemple l'article indéfini « un » ou la préposition « de » ; et ceux qui dépendent du domaine traité par le texte, comme par exemple, le mot « résumé » qui a sûrement une fréquence importante dans ce chapitre. Il faut donc éviter que toutes les phrases qui le contiennent se voient attribuer un bon score ; les concepteurs de cette méthode essayent ainsi de normaliser le calcul. Cette entreprise s'avère délicate puisqu'il faut à la fois choisir un corpus de référence avec un spectre lexical le plus large possible pour pouvoir éviter les effets lexicaux dus au domaine (un terme comme « vache » est ainsi sûrement plus fréquent dans un texte traitant d'agriculture ou d'écologie que dans un texte de linguistique) et suffisamment concentré pour éviter d'introduire trop de termes polysémiques. Soulignons que la notion de mot vide, issue des sciences de l'information, n'a aucune justification linguistique : les pronoms, par exemple, sont souvent considérés comme des mots vides dans ces systèmes alors que ce sont souvent des marques anaphoriques qui devraient donc être prises en compte dans le calcul du score.

L'hypothèse sous-jacente à ce type de méthode est que l'importance d'une unité textuelle est une fonction des éléments lexicaux qu'elle contient, moyennant un correctif lié au domaine traité et aux usages syntaxiques, ou en d'autres termes que les phrases qui contiennent les mots les plus fréquents sont représentatives de la thématique du texte. Cette hypothèse amène plusieurs remarques :

- la première remarque concerne les pratiques stylistiques en cours. Il est remarquable, aussi bien dans les textes techniques que dans les articles de presse, de constater que l'auteur n'utilise presque jamais le même terme lexical pour désigner un même référent, déployant pour ce faire toutes les possibilités offertes par la rhétorique (utilisation de synonymes, de métaphores, etc.). Ainsi, dans un article d'un quotidien économique traitant de la fonction du dirigeant d'entreprise, nous avons relevé l'emploi du terme « dirigeant d 'entreprise », mais aussi ceux

de « décideur », de «skipper », d'«entrepreneur », de « capitaine d'industrie» et de « chef d'entreprise ».

- la deuxième remarque concerne le fait que le calcul ne tient pas compte de l'usage, très fréquent, des anaphores pronominales ; ainsi comme le montre l'exemple suivant, le score calculé pour la deuxième phrase, qui commence par le pronom *Elle* et qui ne prend donc pas en compte l'importance du syntagme nominal *génétique du comportement,* risque d'être inférieur au seuil prédéfini : [12]

> « Etonnamment, la génétique du comportement peut nous aider à affiner notre compréhension des multiples rapports au monde qu'entretiennent les hommes. Elle nous indique que si l'on veut comprendre la complexité du comportement humain, il faut au moins prendre en compte les interactions non-linéaires qui régissent les rapports entre l'environnement interne et l'environnement externe ». *La Recherche (numéro 311, page 107)*

De plus, même si l'on fait abstraction des problèmes soulevés par l'hypothèse sous-jacente à cette méthode, le mode de calcul du score fait apparaître plusieurs difficultés. Remarquons néanmoins que son principal avantage est qu'il est simple à effectuer : un outil de découpage en mots, même frustre, suffit et les temps de traitement sont donc excellents. Cependant, les inconvénients restent nombreux :

- Si aucun traitement morphosyntaxique n'est effectué, deux formes peuvent être considérées comme le même mot. Par exemple l'occurrence « avions » peut être considérée comme le pluriel du mot « avion », mais aussi comme la forme conjuguée du verbe « avoir ». Un étiqueteur morphosyntaxique permet de résoudre en partie ce problème, mais dégrade les performances, et soulève le problème du traitement des mots inconnus, comme les néologismes, ou les mots mal orthographiés, ou simplement empruntés à une langue étrangère ;

- Un deuxième inconvénient a trait au traitement des groupes nominaux. Ainsi le groupe nominal « effet de serre » sera décomposé en trois mots, avec le risque que le mot « effet » soit considéré comme un mot vide[13], ce qui diminuera d'autant le score de la phrase qui le contient. Là encore, il est possible d'utiliser un extracteur de groupes nominaux, sachant néanmoins qu'actuellement la performance de ces outils reste limitée ;

- Le dernier inconvénient concerne la prise en compte des différents niveaux de discours. Faut-il par exemple, considérer les citations comme faisant partie intégrante du texte ou au contraire

[12] Il est vrai que l'identification du référent d'une anaphore est un problème non résolu à l'heure actuelle. Ainsi dans cet exemple, un système de résolution d'anaphore identifierait au moins deux candidats potentiels, le terme *génétique du comportement* et le terme *compréhension,* en supposant qu'un tel système soit capable d'identifier les syntagmes nominaux.

[13] La locution « en effet », fréquente dans de nombreux textes techniques et scientifiques, est à l'origine de la présence du mot « effet » dans la liste des mots vides.

les ignorer ? Généralement, ces systèmes adoptent une position tranchée : soit ils ne prennent pas en compte les niveaux de discours du texte à résumer, considérant ainsi les citations comme partie intégrante du texte, soit ils suppriment systématiquement les citations directes. Voici un exemple, extrait d'un texte de *La Recherche* illustrant la difficulté d'avoir une position aussi tranchée sur ce problème (nous respectons les changements de police de caractères et la ponctuation du texte d'origine) :

> Un article récent de W. French Anderson, pédiatre et chercheur en thérapie génique, intitulé « Genetic engineering and our Humanness » repose sur la même erreur de façon plus flagrante. L'auteur se pose la question : *« Est-il possible d'altérer notre humanité à l'aide du génie génétique?* » Son raisonnement repose sur la distinction entre le corps, qui n'est pas *« ce qui fait d'un être humain un être unique par rapport à tous les autres* », et l'âme qui est « *la part spirituelle qui fait de chacun un être humain unique* ». Il conclut de ces prémisses : « *Ce qui constitue fondamentalement l'humanité de chacun... ne dépend pas de l'enveloppe corporelle. Or seule cette enveloppe peut être altérée. Il s'ensuit que le génie génétique ne peut altérer ce qui constitue l'humanité de chacun.* » En d'autres termes, Anderson soutient que nous n'avons aucune raison de craindre une altération de notre humanité puisque le génie génétique ne peut altérer ce qui fait de nous un être humain unique. Mais qu'entend-il par « altération de notre humanité »? « *Notre humanité*, explique-t-il, *est cette part " spirituelle " et " non quantifiable " de notre être qui fait que chaque être humain est unique par rapport aux autres. » La Recherche (N° 311, page 103)*

L'usage que fait l'auteur, pour construire son argumentation, des différents types de guillemets et de l'italique (pour citer un confrère) montre combien il est difficile d'exclure du calcul du score la citation elle-même, en admettant d'ailleurs qu'un logiciel de traitement automatique puisse être capable de démêler cet écheveau.

De notre point de vue, le principal défaut de cette méthode est plus général : le résumé est dans le meilleur des cas constitué des phrases représentatives de la thématique du texte, si l'on fait l'hypothèse que cette thématique peut se calculer uniquement à partir des éléments lexicaux qui composent le texte. Il n'est ainsi pas possible de prendre en compte des « actes discursifs ». Or, l'utilisateur d'un système d'activité résumante peut être intéressé par des informations qui ne relèvent pas de la thématique principale du texte.

2.2.2. Sélection d'unités textuelles fondée sur un calcul de similarité lexicale

Cette méthode s'appuie au départ sur les travaux de Salton [SAL 83] dans le domaine de la recherche d'information à l'aide de systèmes documentaires. Elle vise à pallier aux défauts de la méthode précédente en proposant d'exploiter la structure physique et thématique du texte. Un texte ou un document D_i est représenté comme un vecteur de termes pondérés, de la forme :

$$D_i = (d_{i1}, d_{i2}, .., d_{ik}, d_{it})$$

où les d_{ik} représentent les « poids » des termes T_k présents dans le document D_i. La première étape, le calcul du poids d'un terme, s'effectue généralement en appliquant la formule précédente (*tf*idf*), mais d'autres fonctions sont parfois mises en œuvre. L'étape suivante consiste à calculer un coefficient de similarité lexicale $Sim(D_i, D_j)$ entre les documents :

$$Sim(D_i, D_j) = \sum^{t}_{k=1} d_{ik}, d_{jk}$$

Ce coefficient de similarité est appliqué, non pas à des documents, mais à des paragraphes d'un même document [MIT 97, ABR 97], pour construire des résumés en calculant la similarité entre ces paragraphes. On obtient ainsi une matrice NxN de similarité, où N est le nombre de paragraphes du document, à partir de laquelle plusieurs stratégies de sélection de paragraphes sont possibles :

- la première stratégie consiste à rechercher le paragraphe qui possède le plus de liens de similarité avec les autres paragraphes, ce qui, d'après les concepteurs de la méthode, est un signe que ce paragraphe traite des principaux thèmes du texte. Ce processus est répété sur les paragraphes restants jusqu'à ce que l'on obtienne un résumé dont la taille n'excède pas un seuil déterminé.
- la deuxième stratégie vise à corriger un défaut de la stratégie précédente qui sélectionne des paragraphes fortement liés avec tous les autres paragraphes du texte. Ces paragraphes sélectionnés, qui forment le résumé, ne sont pas nécessairement liés entre eux, d'où le risque d'obtenir des résumés très peu cohérents. Pour améliorer la cohérence, le principe est de partir d'un paragraphe P_A, le premier ou encore celui qui possède le plus de connexions avec les autres paragraphes du texte, puis de choisir le paragraphe P_B qui possède le plus fort coefficient de similarité avec P_A. Le processus de sélection est répété jusqu'à ce que l'on obtienne un résumé de la longueur voulue. Les résumés ainsi obtenus sont plus cohérents mais ils ne couvrent alors que partiellement les thèmes du texte.

D'autres stratégies peuvent être élaborées à partir des deux précédentes : par exemple, en segmentant le texte, c'est à dire en regroupant des paragraphes et en obligeant le système à extraire au moins un paragraphe dans chaque segment.

Du point de vue des concepteurs de cette méthode, le choix du paragraphe comme unité textuelle saillante doit permettre a priori de résoudre, par contournement, les difficultés liées à la cohérence des résumés produits, en faisant l'hypothèse qu'un paragraphe possède par construction une cohérence thématique. Cependant, plusieurs travaux sur la notion de paragraphe [GRO 85, BES 88] ont montré que le découpage d'un texte en paragraphes obéissait à différents critères qui mélangent des règles discursives, des contraintes typographiques et des choix esthétiques. Par ailleurs, la notion de similarité lexicale utilisée comme critère de saillance et fondée sur un calcul de fréquence de formes, au mieux lemmatisées, sans prise en compte de la synonymie et de la co-référence, apparaît comme une hypothèse trop forte.

2.2.3. Sélection d'unités textuelles basée sur le repérage de phrases prototypiques

Les premiers travaux qui se sont éloignés des méthodes quantitatives décrites précédemment ont proposé une alternative au problème du résumé en intégrant de nouveaux critères de sélection. Ces travaux partent du constat que le texte produit par l'extraction des phrases selon des approches purement fréquentielles et lexicales contient, certes, un certain nombre de phrases représentatives du contenu thématique du document, mais que les techniques mises en oeuvre ne prennent pas en compte la manière dont l'auteur emploie ces éléments lexicaux.

L'approche par phrases prototypiques (*cue-phrases*) part d'observations effectuées sur des corpus de textes essentiellement scientifiques et techniques, et fait émerger des critères de sélection autres que ceux basés uniquement sur la fréquence des mots :

- certains mots ou expressions du texte peuvent indiquer l'importance des phrases, indépendamment de considérations purement fréquentielles. Par exemple, les phrases contenant des expressions du type « notre travail », « ce papier », « la présente recherche », etc., sont des marques placées par l'auteur pour présenter le thème de son article ;
- certaines expressions, en se référant à des passages précédents, font office de liens structurels entre les différentes parties d'un texte et peuvent donc être exploitées pour construire des résumés plus cohérents. C'est le cas d'expressions du type « présenté précédemment », « énoncé au-dessus », etc. ;
- la position des phrases dans un texte peut aussi être utilisée comme critère de sélection. Par exemple, les phrases de l'introduction, de la conclusion ou de certaines sections du texte ont une certaine importance par rapport à d'autres phrases du texte.

Les travaux de C. D. Paice [PAI 81, 90] sont les plus représentatifs de cette approche. L'auteur propose des critères d'importance, en partie à partir de l'observation d'expressions linguistiques du type de l'exemple de la figure 4. Les mots en gras se trouvent dans le schéma de l'expression de la figure 5.

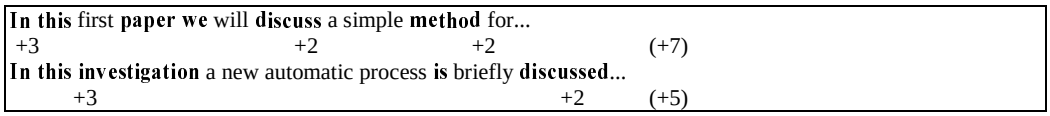


Figure 4. Exemples d'expressions importantes d'après [PAI 81]

À partir d'expressions similaires, l'auteur définit des règles qui se fondent sur deux types d'informations :

- la présence de mots ou de classes de mots. Ainsi, la classe de mots se référant à la discussion (« discuss words ») est constituée par les éléments suivants : *discuss-, introduce, present, examine, describe, review, report, outline, consider, investigate, explore, assess-, analy-, synthesi-, styd-, survey, ask, simplif-, deal-, trace(-2), cover(-2).* Les nombres entre parenthèses sont des poids qui indiquent la fréquence d'utilisation des mots qui leur sont associés par rapport au reste de la liste ;
- des schémas d'expressions (cf. fig. 5) construits à partir d'une généralisation des expressions linguistiques déjà observées. Les nombres entre parenthèses représentent le nombre de mots permis qui peuvent s'intercaler dans l'expression ; les nombres en exposant sont les poids associés à chaque mot ; les mots suivis par un point d'interrogation sont optionnels.

De cette manière, à chaque phrase est attribué un score résultant du cumul des poids figurant dans les schémas d'expressions. Après la phase de calcul des scores intervient la phase d'extraction des phrases ayant cumulé les plus grands scores. Celles-ci sont extraites avec les phrases adjacentes lorsqu'elles contiennent des références à des éléments externes.

Un premier système a été développé dans ce cadre [PAI 81] à l'Université de Lancaster. Il est composé d'une base de *cue phrases*, regroupées en classes, et d'un ensemble de règles de calcul d'importance ; certaines règles attribuent des poids négatifs en vue d'éliminer une phrase. Un deuxième système, GARP [PAI 90], utilise les techniques de résolution de la référence dans les phrases et résume exclusivement des textes techniques et scientifiques.

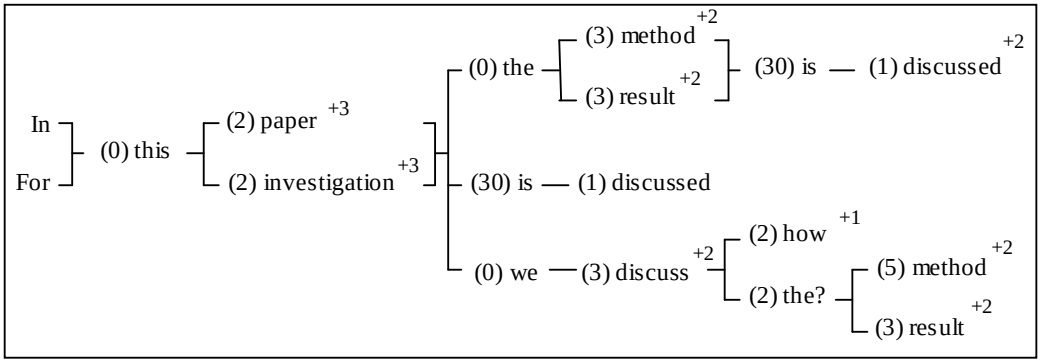


Figure 5. Exemple de schéma d'expression linguistique d'après [PAI 81]

A l'Université de Nancy 2, A. Lehmam [LEH 95] s'est inspiré de cette approche pour implémenter le système RAFI (Résumé automatique à fragments indicateurs). RAFI s'appuie sur un calcul de scores en fonction de la présence ou non de certains mots dans une phrase. Ces mots constituent le contenu de ce que l'auteur appelle « les fragments de phrases indicateurs ». Le système a comme entrée un texte sous forme de phrases séparées. La première étape consiste à rechercher dans chaque phrase un FPI (fragment de phrase indicateur) par consultation d'un thesaurus de FPI. Le système calcule ensuite le score de chaque phrase. La dernière phase consiste à sélectionner les phrases ayant eu les plus grands scores. Cette sélection se fait en fonction de la longueur finale de l'extrait, 25% étant le résumé type.

Ce qui confère à cette approche une certaine originalité, tient au fait qu'elle prend en compte des aspects purement linguistiques totalement absents dans les approches uniquement quantitatives. Des critères linguistiques sont ainsi définis pour l'extraction de phrases jugées importantes. Le problème de la cohérence de l'extrait final a été abordé par [PAI 90] qui propose quelques solutions pratiques, notamment en ce qui concerne la résolution des anaphores placées en début de phrases.

Son inconvénient majeur réside dans le fait que l'importance donnée aux phrases se réduit au calcul d'un score numérique. Or, aucune justification n'est fournie quant au calcul des coefficients d'importance donnés aux mots ou aux expressions prototypiques : comment évaluer, et en fonction de quels facteurs, ce coefficient ? Est-ce qu'à un mot (ou à une classe de mots), est toujours associé un même poids, indépendamment de l'expression où il (elle) se trouve? Par ailleurs, la quantification de l'importance par des nombres limite considérablement les possibilités de faire varier le contenu informatif de l'extrait final. Par exemple, dans le système RAFI, les expressions recherchées (les Fragments de Phrases Indicateurs) sont porteuses d'informations sur les intentions de l'auteur du texte. Ainsi le FPI, *Le but de la méthode discutée ici est de montrer comment* , peut être qualifié « d'annonce thématique » ; le fait de réduire cette information à un score numérique a pour conséquence une perte d'information importante : d'une part, parce que le principe qui consiste à effectuer une somme des

scores entraîne l'élimination des phrases qui ne contiennent qu'un seul FPI ; d'autre part, parce que le système ne peut pas s'adapter aux besoins spécifiques d'un utilisateur.

2.2.4. Sélection d'unités textuelles basée sur le repérage de chaînes lexicales

La cohésion du résumé reste une des principales difficultés auxquelles se heurtent les méthodes par extraction décrites précédemment. La cohésion d'un texte tient à plusieurs facteurs comme l'utilisation de termes sémantiquement liés, d 'anaphores, d'ellipses, etc. La cohésion lexicale peut être repérée en recherchant ce type d'éléments liés, les chaînes lexicales [SLA 75, 80, CHA 88, MOR 91]. L'approche de [BAR 97], à la suite des travaux de [STA 96] en informatique documentaire, se fonde sur l'hypothèse que les chaînes lexicales, représentatives à la fois des éléments thématiques du texte et de sa cohésion, sont des sources possibles pour la construction de résumés. La procédure de repérage employée est la suivante :

- sélectionner un ensemble de mots candidats. Les auteurs s'appuient sur WordNet [MIL 88, 90], réseau électronique de termes, qui fournit des liens entre les différentes entrées du dictionnaire (notamment les liens de synonymie qui regroupent les termes en « synset ») pour réaliser cette sélection. Les noms et les noms composés sont *a priori* des candidats.
- pour chaque mot candidat, rechercher une chaîne à laquelle ce mot est relié. Pour déterminer si un mot est sémantiquement lié à un autre, les auteurs recherchent s'il existe « un chemin » qui les relie dans WordNet. Trois types de liens sont définis : *très fort, fort* ou *standard.* Suivant le type de lien existant, la distance, calculée en nombre de mots qui sépare les occurrences des mots candidats dans le texte, est prise en compte. Ainsi, si deux mots ont un lien de type *très fort,* la distance n'intervient pas ; par contre, si ce lien est de type *standard*, ils ne doivent pas être distants de plus de trois phrases. Aucune argumentation, autre que celle liée à des observations empiriques, n'est fournie pour justifier ce choix.
- quand une chaîne est trouvée, insérer le mot dans la chaîne, sinon créer une nouvelle chaîne. Lorsqu'un nouveau mot est inséré, deux cas sont à considérer. Soit le mot est présent dans WordNet avec une seule signification et dans ce cas la chaîne intègre cette nouvelle information ; soit le mot apparaît dans WordNet avec plusieurs sens, le système duplique alors les chaînes existantes puis insère, dans chacune d'elles, le mot avec un de ses sens.

Les auteurs fournissent un exemple qui illustre l'apport d'une ressource linguistique du type de Wordnet tout en soulignant le type de difficultés qu'il convient de résoudre. Nous reproduisons en partie cet exemple ci-après.

Le texte traité est le suivant :

Mr Kenny is the person that invented an anaesthic machine which uses micro-computers to control the rate at which an anaesthic is

pumped into the blood. Such machines are nothing new. But his device uses two micro-computers to achieve much closer monitoring of the pump feeding the anaesthetic into the patient.

Le premier terme candidat, compte tenu que la seule catégorie morpho-syntaxique des noms intervient dans le calcul, est le mot « Mr. », ce qui entraîne la création d'un premier nœud N_1 qui contient les informations suivantes fournies par WordNet :

N_1 = [lex, « Mr », sense {mister, Mr.}]

Le terme candidat suivant est « person » qui a deux significations, « human being » et « grammatical category of pronouns ans verb forms », ce qui entraîne la création de deux nœuds, N_2 et N_3 :

N_2 = [lex, « person », sense { human being .}]

N_3 = [lex, « person », sense { grammatical category of pronouns ans verb forms }]

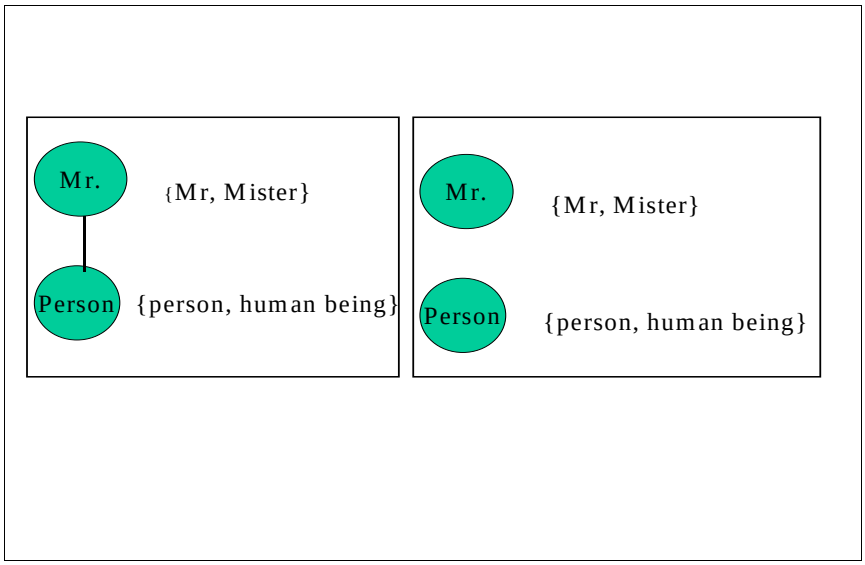


Figure 6. Exemple de composant composé de deux interprétations [BAR 97]

De plus, WordNet indique que le terme « person » avec la signification « human being » est lié par une relation R_1 forte au terme « Mr », ce qui permet de créer deux interprétations ; la première I_1, composée des nœuds N_1 et N_2 et de la relation R_1, la seconde I_2 , composée uniquement des nœuds N_1 et N_2. Ces deux interprétations, exclusives l'une de l'autre, forment un composant, que l'on peut représenter graphiquement (cf. fig. 6).

Le troisième terme candidat « anaesthetic » n'est relié à aucun des mots du premier composant créé ; il entraîne donc la création d'un nouveau composant qui ne contient qu'une seule interprétation.

Le terme « machine » possède cinq sens dont un, « an efficient person », est relié aux sens « person » et « Mr. ». Le terme « machine » est donc inséré dans le premier composant, ce qu'illustre la figure 7.

La suite du traitement permet de sélectionner l'interprétation la plus plausible en s'appuyant sur la force des relations existantes entre les différents nœuds. Ainsi après le traitement des termes « micro-computer », « device » et « pump », les interprétations les plus fortes, au sens où ce sont celles qui possèdent le nombre de nœuds le plus reliés, sont celles qu'illustrent la figure 8. L'interprétation qui contient le plus grand nombre de relations est finalement sélectionnée, les auteurs affirmant que ce critère est le reflet de la cohésion du texte.

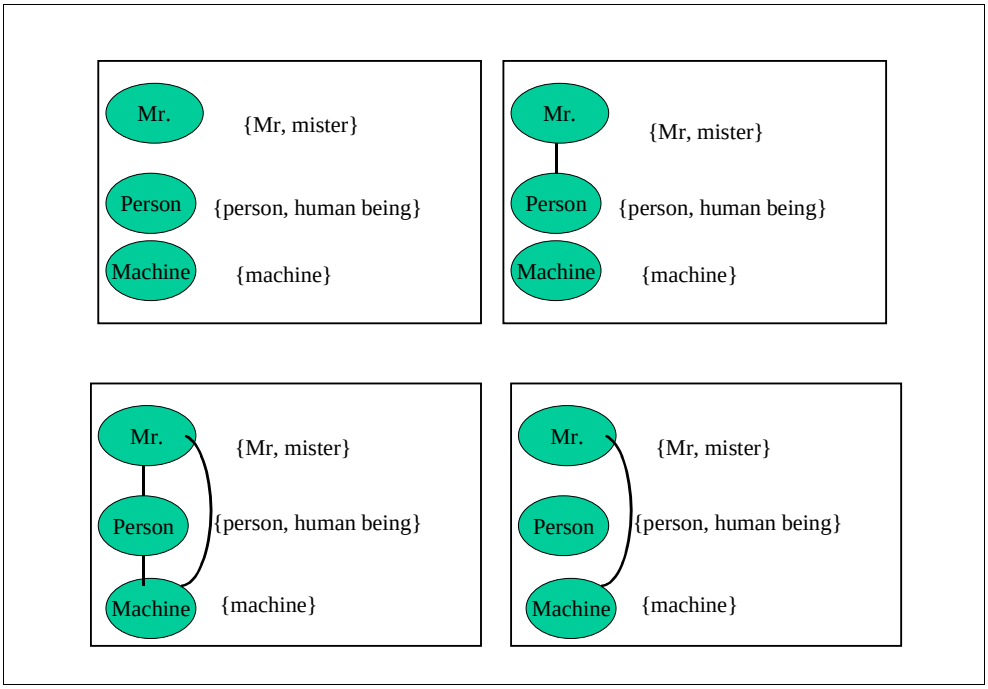


Figure 7. Exemple de composant composé de quatre interprétations [BAR 97]

On remarque que, pour le texte traité, ceci permet de choisir la signification adéquate pour le terme « machine ». En fait, le calcul du score d'une interprétation est défini comme la somme des scores d'une chaîne, et celui-ci est calculé à partir du poids des relations qui existent entre les termes d'une chaîne. Par exemple, le poids de la relation de synonymie est de 10, celui de l'antonymie est 7, et celui de l'hyperonymie est de 4. Tous ces poids ont été fixés expérimentalement. Les auteurs précisent qu'ils prennent aussi en compte les noms composés et que dans ce cas chaque élément du nom composé est relié aux autres termes.

A l'issue de ce processus, le système a identifié un ensemble de chaînes lexicales à partir desquelles il reste à construire un résumé. Pour ce faire, les auteurs proposent d'appliquer un calcul de scores pour classer les chaîne obtenues. Ce score, noté $S(L_i)$, est calculé de la manière suivante :

Score $S(L_i)$ = Longueur(L_i) * Homogénéité(L_i)

Longueur = Nombre d'occurrences de la chaîne
Homogénéité = 1- (Nombre d'occurrences distinctes / Longueur)

Les chaînes dont le score est supérieur à un seuil SE sont conservées. Le calcul du seuil s'effectue de la manière suivante :

SE = Moyenne (Scores) + 2 * DéviationStandard (Scores)

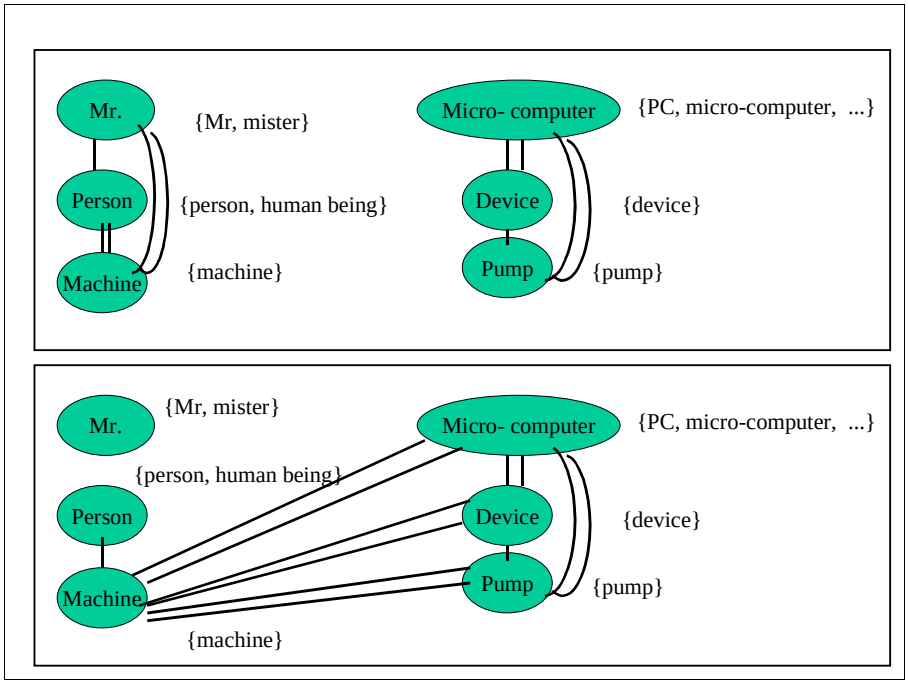


Figure 8. Exemple de composant après le traitement du terme pump [BAR 97]

Ces formules sont le résultat d'expérimentations effectuées sur une trentaine de textes. En moyenne, ces calculs permettent de sélectionner cinq chaînes lexicales pour des textes dont la taille est d'environ un millier de mots.

La dernière étape consiste à extraire des phrases du texte source à partir des chaînes lexicales conservées, le problème à résoudre étant qu'une chaîne lexicale contient un ensemble de termes qui peut apparaître dans plusieurs phrases. D'après les auteurs, les meilleurs résultats sont obtenus en appliquant la stratégie suivante:

- rechercher un segment textuel, c'est à dire un ensemble de phrases dans lequel la chaîne lexicale a une « forte densité ». La densité est définie comme le rapport entre le nombre de mots présents dans le segment textuel et le nombre de mots constitutifs de la chaîne ;
- extraire la phrase qui contient la première occurrence de la chaîne lexicale.

Comme on peut le voir cette approche se fonde sur des calculs, notamment le calcul du score d'un composant, qui n'ont d'autre justification que des ajustements empiriques, comme le soulignent d'ailleurs les auteurs. Néanmoins, elle illustre la volonté de mieux exploiter certaines ressources linguistiques, comme WordNet, en vue d'identifier les marques lexicales placées dans un texte.

2.2.5. Sélection par construction d'une structure pragmatique

Un certain nombre de travaux récents proposent des méthodes d'extraction automatique qui visent à construire une structure pragmatique (appelée aussi argumentative ou rhétorique) du texte en se basant sur des expressions linguistiques catégorisées qui expriment plusieurs types de relations entre phrases d'un texte. La représentation construite dénote les relations de dépendance entre les phrases d'un même texte. Ces relations peuvent être locales à deux phrases adjacentes, mais elles peuvent aussi être générales aux phrases d'une même section.

Ces travaux s'appuient sur l'hypothèse qu'il existe un niveau de représentation censé décrire le contenu global des textes, ce niveau de représentation pouvant être, par exemple, la macrostructure du modèle de Kintsch et Van Dijk (voir §2.1.1). Or, comme le souligne M. Charolles [CHA 91], les macro-règles du modèle de Kintsch et Van Dijk, qui permettent de passer du niveau micro-structurel au niveau macro-structurel, ne sont pas généralisables dans une perspective de résumé automatique car ces règles sont liées à une sémantique spécifique à un domaine. Contrairement aux approches utilisant le modèle de Kintsch et Van Dijk, d'autres approches, [RIN 94, MII 94, MAR 97] proposent s'appuyer essentiellement sur la sémantique de marqueurs qui engendrent des opérations [CHA 91] de :

- consécution (*donc, par conséquent, d'où*, ...);
- de correction (*mais*, ...);
- d'opposition (*pourtant*, *toutefois*, ...);
- de justification (*car, puisque*, ...);
- de confirmation (*en effet*, ...);
- d'explication (*parce que* ...);
- d'illustration (*par exemple*, ...);

Deux systèmes, que nous présentons ci-après, illustrent ce type d'approche, BREVIDOC de [MII 94] et celui de D. Marcu [MAR 97].

2.2.5.1. Système BREVIDOC

Le système BREVIDOC (Broadcatching System with an Essence Viewer for Retrieved Documents) [MII 94] est, comme le spécifient les auteurs, un système qui n'utilise que des connaissances linguistiques. Le système analyse la structure du texte et détermine les relations entre les paragraphes et les phrases du texte en s'appuyant sur des indices linguistiques : les connecteurs, les

anaphoriques et différents types d'expressions linguistiques. Le système comprend un module qui permet l'analyse de la structure du texte et un module de génération des résumés[14].

2.2.5.1.1. Module d'analyse de la structure du texte :

L'analyse de la structure est réalisée à l'aide de quatre composantes :

Analyse de l'organisation du document :

Cette analyse se fonde essentiellement sur les caractéristiques morphologiques des titres et des en-têtes et a pour objectif de construire une représentation du texte. Elle s'appuie sur des heuristiques du type : « les titres commencent par un nombre ou un symbole », « chaque titre est précédé et suivi par des sauts de ligne », etc. Les informations fournies sont les en-têtes de chaque section, les corps des sections (leurs étendues) et la position de chaque paragraphe dans sa section.

Analyse des phrases :

À ce niveau, une analyse morphologique est effectuée, puis une analyse syntaxique qui utilise un dictionnaire de 60 000 entrées. L'analyseur morphologique, spécifique à la langue japonaise où les mots ne sont pas séparés par des blancs, utilise les différentes possibilités de connections entre les mots afin d'éliminer les séquences morphologiques « douteuses ». L'analyseur syntaxique construit, à partir de la sortie précédente, une structure syntaxique. Les ambiguïtés lexicales sont levées en éliminant les catégories de mots qui ne permettent pas une combinaison cohérente.

Analyse de la structure rhétorique du texte :

Elle permet de traiter la structure interne de chaque section et la structure du texte en entier. La structure de chaque section est représentée par un arbre binaire reliant les phrases à l'aide de connecteurs (*donc, par conséquent,* ...) et d'expressions idiomatiques (*en outre, dans ce papier ... est décrit,* ...). Ces expressions et ces connecteurs sont classés en trente quatre catégories. La structure du texte est construite à partir des expressions rhétoriques. Elles sont détectées par des règles qui utilisent des patrons de surface. Selon ce principe, plusieurs structures peuvent être générées. La structure candidate est sélectionnée à l'aide de règles de préférence, en s'appuyant sur des pénalités associées aux relations entre phrases : par exemple la séquence de phrases ([Ph1 <*par exemple*> Ph2] <*alors*> Ph3] sera préférée à une séquence [Ph1 <*par exemple*> [Ph2 <*alors*> Ph3]] où Ph1, Ph2 et Ph3 sont des phrases qui sont reliées par es expressions indiquées entre les symboles « <> ».

Après une analyse de la structure intra-paragraphe, une analyse de la structure inter-paragraphe est effectuée : elle utilise les relations rhétoriques de la première phrase de chaque paragraphe. Par exemple, la structure du texte suivant est représentée par la figure 9 (les mots soulignés indiquent les connecteurs) :

[14]Il est implémenté sur une station de travail, et est utilisé essentiellement sur des textes techniques japonais de la revue mensuelle « Toshiba review » et des éditoriaux de « Asashi ».

paragraphe extrait d'un article de Rabiner L. R., Schafer R. W., 1978, "A zero-crossing rate which estimates the frequency of a speech signal", *Digital processing of speech signals*, Prentice-Hall, p 127.
1 : In the context of discretetime signals, zero-crossing is said to occur if successive samples have different algebraic signs.
2 : The rate at wich zero crossings occur is a simple measure of the frequency content of a signal.
3 : This is particularly true of narrow band signals.
4 : For example, a sinusoidal signal of frequency Fo sampled at a rate *F3*, has *F3/F3* samples per cycle of the sine wave.
5 : Each cycle has two zero crossings so that the long-term average rate of zero-crossings is $Z = 2Fo/F3$.
6 : Thus, the average zero-crossing rate gives a reasonable way to estimate the frequency of o sine wave.

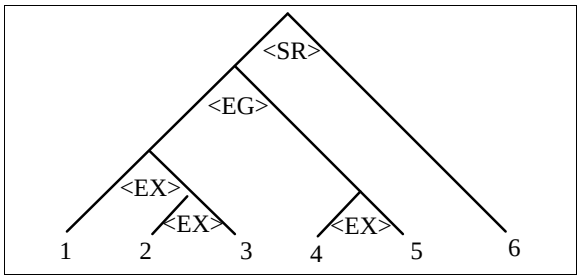


Figure 9. Structure arborescente du texte (repris de [MII 94])

<SR> est la relation "série". Exemples d'expression : *thus, then* ...
<EG> est la relation "exemple". Exemples d'expression : *for example, and so on* ...
<EX> est la relation "extension". Exemples d'expression : *this is, such* ...

Extraction des rôles sémantiques :

Les rôles sémantiques, tels qu'ils sont décrits par les auteurs, correspondent aux différents thèmes abordés dans un article et ils dépendent du type de document traité. Ainsi, un article de recherche inclut l'introduction et le propos de la recherche, l'approche adoptée et la conclusion, tandis qu'un article de journal contient une introduction et l'opinion de l'auteur.

Le système extrait les rôles sémantiques à l'aide de règles qui recherchent dans le texte les expressions linguistiques correspondant à ces rôles sémantiques. Par exemple, les expressions « ... has been achieved » et « .. has been confirmed » correspondent à des rôles sémantiques de conclusion. Une fois extraits, ces rôles sémantiques sont projetés sur les phrases avoisinantes en respectant la structure du texte. Ainsi, dans la séquence de phrases « The feature (le sujet) ... as follows. First ... Second ... », le rôle sémantique de *feature* (le sujet), extrait de la première phrase, est projeté sur les deuxième et troisième phrases.

2.2.5.1.2. Module dynamique de génération des résumés

Ce module génère un résumé en se basant sur l'importance relative des relations rhétoriques pour toutes les sections. Les relations sont catégorisées en *RightNucleus* (le nœud droit est plus important que le nœud gauche), *LeftNucleus* (inverse de la situation précédente) et *BothNucleus* (les deux nœuds

sont importants). Par exemple la relation correspondant à la séquence « thus *(*ainsi) ... » est la relation rhétorique de conclusion. Cette relation est catégorisée en *RightNucleus,* le nœud droit étant considéré comme plus important que le nœud gauche (la deuxième phrase résume la première).

2.2.5.2. Système proposé par D. Marcu

Les travaux de Marcu [MAR 97] se fondent explicitement sur la Rhetorical Structure Theory (RST) développée par [MAN 88] en vue de construire ce que l'auteur appelle un analyseur rhétorique (*rhetorical parser*). Cet analyseur considère le texte comme un ensemble d'unités élémentaires disjointes, reliées entre elles par deux types de relations, les relations *paratactic* et les relations *hypotactic.* Les relations *paratactic* lient des unités d'égale importance, alors que les relations *hypotactic* lient un *nucleus*, c'est à dire une unité considérée comme essentielle par l'auteur, avec un satellite, unité considérée comme non essentielle. L'analyseur va d'abord construire un arbre binaire, l'arbre RS, en appliquant l'algorithme suivant :

- Identifier l'ensemble D des marqueurs discursifs du texte T et l'ensemble U_T des unités textuelles élémentaires de T ;
- Identifier l'ensemble des relations potentielles R entre les éléments U_T ;
- Déterminer l'ensemble ValTrees, ensemble de tous les arbres RS valides sur T;
- Calculer l'arbre RS optimum.

L'identification de l'ensemble D s'appuie sur une analyse d'un corpus de 7900 fragments de textes, qui a permis de construire une base de 450 marqueurs discursifs. L'auteur fournit un exemple que nous reproduisons ci-après. Le texte analysé est le suivant :

> *[With its distant orbit [1]] [-50 percent farther from the sun than Earth -[2]] [and slim atmospheric blanket, [3]][Mars experiences frigid weather conditions. [4]] [Surface temperatures typically average about -60 degrees Celsuis (-76 degrees Farenheit) at the equator [5]] [and can dip to -123 degrees C near the poles . [6]] [Only the midday sun at tropical latitudes is warm enough to thaw ice on occasion, [7]] [but any liquid water formed in this way would evaporate almost instantly [8]] [because of the low atmospheric pressure. [9]]*
>
> *[Although the atmosphere holds a small amoount of water, [10]] [and water-ice clouds sometimes develop, [11]] [most Martian weather involves blowing dust of carbon dioxide. [12]] [Each winter, for example, ablizzard of frozen carbon dioxide rages over one pole, [13]] [and a few meters of this dry-ice snow accumulate [14]] [as previously frozen carbon dioxide evaporates from the opposite polar cap. [15]] [Yet even on the summer pole, [16]] [where the sun remains in the sky all day long, [17]] [temperatures never warm enough to melt frozen water. [18]]*

A partir d'une analyse de surface qui prend en compte certains marqueurs (essentiellement des connecteurs), le texte est découpé automatiquement en unités élémentaires. Si ces unités correspondent en général à des propositions, ce n'est pas toujours le cas. Ainsi, dans le texte présenté, les propositions identifiées par un analyste humain sont placées entre les signes [ ], alors que le système considérera les propositions 13, 14 et 15 comme une seule unité élémentaire (cf. fig. 10).

D. Marcu fait l'hypothèse que ces unités élémentaires sont totalement disjointes, qu'elles sont liées par des relations de type rhétorique et que ces relations sont généralement binaires. L'auteur n'explicite pas cette dernière hypothèse et souligne qu'elle est simplement vérifiée dans la plupart des textes. Puisqu'il peut exister plusieurs relations entre les unités élémentaires, la détection de ces relations binaires étiquetées sur le texte précédent conduit à la construction d'un ensemble d'arbres binaires.

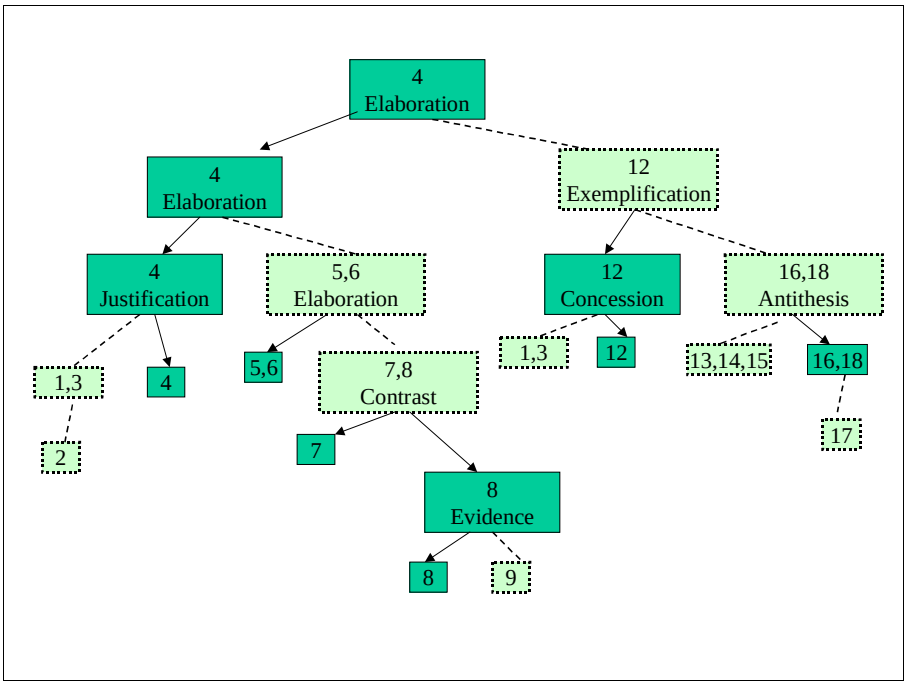


Figure 10. L'arbre rhétorique optimum calculé [MAR 97]

L'étape suivante consiste à choisir l'arbre optimum. Ce choix est effectué en faisant l'hypothèse que le meilleur arbre est celui qui est le plus déporté vers la droite, ce qui correspond à la même hypothèse que celle appliquée dans le système BREVIDOC. La figure 10 illustre l'arbre choisi pour le texte traité. Les nombres placés dans les rectangles correspondent aux numéros des unités élémentaires du texte tandis que les étiquettes comme « Elaboration » ou « Justification » correspondent aux types de relations rhétoriques. Les nucleus sont représentés par des rectangles en traits pleins et les satellites par des rectangles en traits pointillés. Un lien entre un nœud et un nucleus satellite est représenté par

une flèche en trait plein et un lien entre un nœud et un satellite subordonné par une flèche en trait pointillé.

La construction du résumé consiste alors à choisir dans l'arbre sélectionné les unités élémentaires qui sont considérées comme saillantes. Par construction, l'unité élémentaire qui se trouve à la racine est l'unité la plus saillante. Si la taille du résumé n'est pas atteinte, les nœuds fils sont sélectionnés, et ainsi de suite jusqu'à obtenir un résumé de taille adéquate. Dans l'exemple traité, les unités choisies sont en premier lieu l'unité 4, en second lieu l'unité 12, puis les unités 5 , 6, 16 et 18, et ainsi de suite.

Il convient de souligner que Mann et Thompson, fondateurs de la RST, ont insisté à plusieurs reprises le caractère interprétatif de la RST. Ils précisent que deux analystes produiront des analyses différentes. Le fait de considérer qu'une relation existe entre deux unités ne s'appuie pas uniquement sur des marques lexicales, mais aussi sur la sémantique du discours analysé :

> *« ... RST provides a general way to describe the relations among clauses in a text, whether or not they are grammaticlly or lexically signalled. Thus, RST is a useful framework for relating the meanings of conjunctions, the grammar of clause combining, and non-signalled parataxis... »*
>
> [MAN 88] p. 244.

> « *We and others have had the experience of giving the same text to several analysts, who then created differing analyses, sometimes more than on from an individual analyst. Thes are several qualitatively different kinds of multiplicity :*
>
> 1. *Boundary judgements - reults of forcing borderline case into categories.*
> 2. *Text Structure Ambiguity - comparable to many other varieties of linguistic ambiguity.*
> 3. *Simultaneous Analyses - multiple compatible analyses.*
> 4. *Difference Between Analysts - especially, differing plausibility judgements. »*
>
> [MAN 88] p. 265

On peut donc dire que D. Marcu va au-delà des intentions des fondateurs de la RST. Néanmoins, cette approche caractérise bien les tendances actuelles, à savoir la volonté de prendre en compte des structures de discours en s'appuyant sur des connaissances linguistiques, tout en évitant la construction de représentations trop complexes. On peut cependant regretter que l'auteur n'exploite pas mieux les valeurs des relations rhétoriques pour produire des résumés ciblés en fonction des besoins de l'utilisateur.

2.2.6. SERAPHIN, un système de sélection fondée sur un étiquetage sémantique

Le système SERAPHIN [LER 94, BER 95a, 95b, 96a] a été développé par l'équipe LaLIC du Centre d'Analyse et de Mathématiques Sociales (CAMS) en collaboration avec la Direction des Études et des Recherches (DER) de la société EDF. A l'origine, SERAPHIN était un système qui avait pour objet d'extraire automatiquement d'un texte un ensemble de phrases pertinentes pour un utilisateur; il s'insérait dans le contexte général d'automatisation de la chaîne de traitement (cf. fig. 11) et de manipulation des documents textuels à EDF-DER.

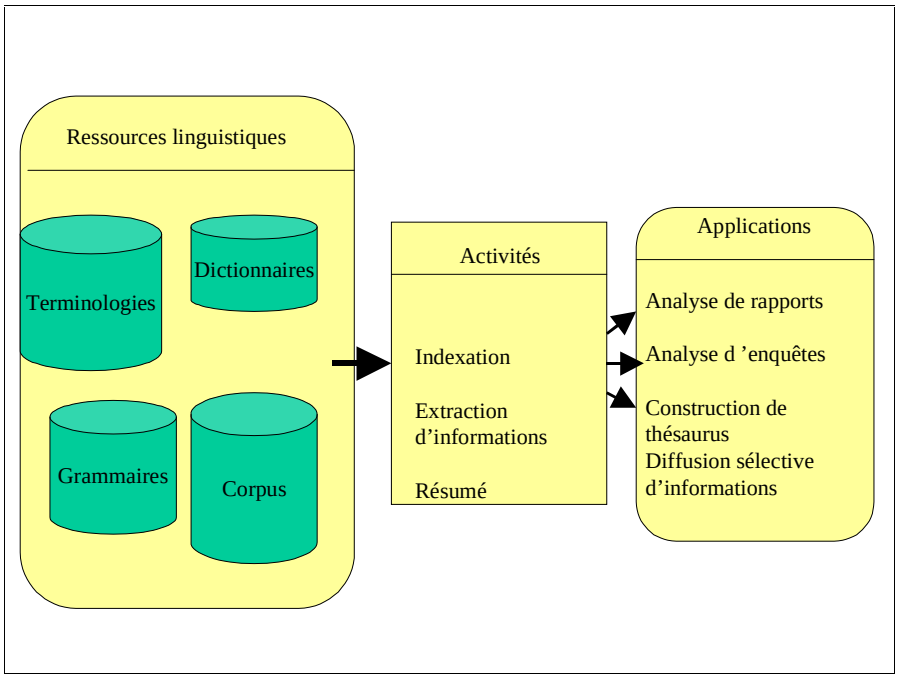


Figure 11. Contexte de développement de SERAPHIN d'après [LER 92]

SERAPHIN devait donc satisfaire à des contraintes de type industriel :

- les textes à résumer traitaient de domaines très divers : l'effet de serre, le véhicule électrique, le réchauffement des climats, la sémiologie, etc. Cette diversité des domaines ne permettait pas d'envisager la construction d'ontologies ou de ressources lexicales spécialisées ;
- les textes à analyser étaient issus d'origines diverses, journaux spécialisés, rapports techniques, articles scientifiques ; ils présentaient par conséquent une structuration physique très variable, mais cette structuration n'était pas indiquée explicitement puisque certains textes étaient numérisés par un passage au lecteur optique ;
- l'extrait fourni par le système devait être lisible par un lecteur intéressé par le thème dont traitait le texte source, mais l'extrait n'avait pas pour finalité d'être publié en l'état ; il était donc possible de tolérer des ruptures thématiques ;

- l'utilisation du système devait être la plus interactive possible, ce qui impliquait de produire un résumé en quelques minutes.

L'approche développée dans SERAPHIN est fondée sur la méthode d'exploration contextuelle [DES 88, 91, 97a, 97b]. Cette méthode vise à identifier les connaissances linguistiques en les restituant dans leurs contextes et en les organisant en tâches spécialisées. Elle présente l'avantage de rendre le travail linguistique relativement indépendant de son implémentation informatique mais aussi d'articuler effectivement dans une même architecture logicielle, les deux types de démarche. Pour appliquer la méthode d'exploration contextuelle, le linguiste doit accumuler des marqueurs linguistiques ; ces marqueurs pouvant être de simples unités lexicales comme le verbe « *présenter* » ou des unités composées comme « *les lignes suivantes»*. Ces marqueurs linguistiques, lorsqu'on les identifie dans un texte constituent des indices de ces catégories. Certains indices sont plus importants que d'autres. Ils sont appelés indicateurs. Le linguiste doit ensuite spécifier, sous forme de règles d'exploration contextuelle, les dépendances contextuelles entre ces indices. Dans l'exemple ci-après, la co-présence de l'indicateur « *Cet article »* et de l'indice *« brosser »,* sous réserve que ces marques apparaissent dans le premier paragraphe[15], expriment une *Annonce Thématique*.

> *Exemple : **Cet article brosse** un portrait du paysage électrique chinois actuel ainsi que des insuffisances et incertitudes qui l'entourent, puis **décrit** la stratégie que EdF a choisi de mettre en œuvre dans ce contexte. (Revue de l'Energie)*

Nous expliquerons plus en détail, dans la deuxième partie de ce document, la méthode d'exploration contextuelle, et nous montrons ci-dessous comment cette méthode a été utilisée dans le projet SERAPHIN.

A l'origine du projet, les concepteurs avaient exprimé des règles d'exploration contextuelle qui pouvaient produire deux actions : « *conserver »* la phrase ou « *éliminer »* la phrase. Le résumé était ensuite construit par extraction dans le texte source des phrases qui avaient le statut « *conserver »*. Les résultats des premiers tests et les recherches linguistiques effectuées en parallèle [CAR 97, JAC 98] ont amené les auteurs à élaborer des règles d'exploration contextuelle dont l'objectif était d'attribuer des étiquettes aux phrases du texte source, et à développer des stratégies de sélection des phrases à extraire à partir des étiquettes attribuées. L'ensemble de ces travaux a donné lieu au développement du système SAFIR [BER 96b].

L'architecture du système, identique pour SERAPHIN et SAFIR, se compose d'un module de reconnaissances des éléments structuraux du texte, d'un module d'étiquetage qui utilise les connaissances linguistiques, c'est à dire la base d'indicateurs, d'indices et les règles d'exploration contextuelle, et d'un dernier module qui construit le résumé. Cette architecture s'appuie sur les concepts de systèmes à base de connaissances et nécessite un moteur d'inférence en chaînage avant. Il

[15] La position d'une marque dans le texte est considérée aussi comme un indice.

revient à C. Jouis d'avoir mis le premier en œuvre ce type d'architecture dans le système SEEK [JOU 93].

Le module de reconnaissances des éléments structuraux segmente le texte source en paragraphes et en phrases. Cette segmentation est réalisée en appliquant des règles heuristiques qui s'appuie sur l'analyse de la ponctuation. Les résultats de ce découpage se sont montrés, à l'usage, insuffisants, notamment pour des textes journalistiques où les auteurs utilisent abondamment des citations. La reconnaissance des éléments structuraux du texte, à savoir, les sections, les paragraphes et les phrases, est utilisée pour construire un modèle du texte. Il faut préciser que dans SERAPHIN comme dans SAFIR, la reconnaissances des sections n'est pas automatique.

Le module d'étiquetage sémantique se décompose en deux sous-systèmes. Le premier a en charge la reconnaissance lexicale, limitée aux seuls indices et indicateurs, soit pour ces deux systèmes un ensemble d'environ trois milles marqueurs organisés en listes. Cette reconnaissance donne lieu à la construction d'une base de faits. Chaque élément de cette base de faits est un objet valué qui décrit les informations nécessaires à l'application des règles d'exploration contextuelle :

- les coordonnées de l'unité lexicale dans le texte, c'est-à-dire sa position dans la phrase, la position de la phrase dans la paragraphe, la position du paragraphe dans la section et la position de la section dans le texte ;
- la forme de l'unité lexicale ;
- la classe d'exploration contextuelle de l'unité lexicale, c'est à dire le nom symbolique de la liste à laquelle appartient la forme.

Le deuxième sous-système applique les règles d'exploration contextuelle. Ces règles sont d'abord écrites par le linguiste qui utilise un langage de spécification très proche des règles de production classiques comme l'illustre la figure 12.

Règle *Centhe1.1*

Soit le contexte d'une phrase P

Si l'on rencontre dans P un élément X appartenant à la liste *#Centhe1*

Et si l'on rencontre dans P un élément X appartenant à la liste *#Centhe1.2.3* dans les cinq positions qui suivent X

Alors attribuer à P l'étiquette « ENONCE_SOULIGNEMENT »

Figure 12. Règle d'exploration contextuelle exprimée dans le langage de spécification [BER 96a]

Ces règles sont ensuite traduites (cf. fig. 13) manuellement dans le langage SMECI, moteur d'inférence réalisé en LISP par la société ILOG. Sans entrer dans le détail de ce processus de réécriture, il convient néanmoins de souligner que l'expression de ce type de règle nécessite des connaissances en informatique, notamment pendant la phase de test des règles.

```
Règle Centhe1.1
Soit    *s une Section
Soit    *p une Phrase parmi les SecPhrases de *s
Soit    *x une Unite-lexicale parmi les Phunites de *p
Soit    *y une Unite-lexicale parmi les Phunites de *p
Si      Ulmarqueur^*x = #Centhe1.1
et      Ulmarqueur^*x = #Centhe1.2.3
et      Ulplace^*x < Ulplace^*y
et      Ulplace^*y < Ulplace^*x + 5
Alors   ajouter # 'ENONCE-SOULIGNEMENT' aux phconserver^*
Action
        #(sortie-ul *p 'Centhe1.1 *x *y)
finregle
```

Figure 13. Règle d'exploration contextuelle exprimée en SMECI [BER 96a]

Les règles sont considérées comme indépendantes ; par conséquent l'ordre de leur déclenchement est indifférent. Pour les concepteurs du système, ce principe est plus facilement maîtrisable lors de l'écriture des règles par un non-spécialiste, en évitant notamment les problèmes de manipulation d'arbres de décision, et de plus il correspond mieux à l'hypothèse selon laquelle certaines marques sémantiques ne sont pas exclusives entre elles. Par exemple, la présence d'une négation dans une phrase conclusive n'implique pas que cette phrase ne soit pas par ailleurs une « conclusion » comme l'illustre l'exemple suivant :

> *On peut donc en conclure que l'effet de serre est la cause principale du réchauffement de l'atmosphère.*
> *On ne peut donc pas en conclure que l'effet de serre est la cause principale du réchauffement de l'atmosphère.*

Chaque règle d'exploration contextuelle permet d'attribuer une étiquette aux phrases qui contiennent les indicateurs et les indices pertinents et contribue ainsi à produire une structure hiérarchisée « décorée » par des informations sémantiques. Les concepteurs ont défini au total une vingtaine d'étiquettes. Nous allons présenter celles qui sont le fruit d'un travail de recherche

linguistique le plus abouti et qui concernent : l'annonce et la récapitulation thématique, les annonces définitoires et l'expression de la causalité. Nous reprenons la présentation faite par les auteurs de ces travaux.

Annonce thématique

L'étiquette « annonce thématique » [CAR 96] est attribuée aux phrases exprimant le sujet, le thème d'un segment textuel quelconque, ou explicitant une prédication défendue dans un tel segment. Il s'agit d'une information requise pour l'intelligibilité du texte. L'annonce thématique par excellence est le plan du document, exprimé en tête de texte et/ou, plus rarement, en conclusion du texte. Ces énoncés sont reconnus par la co-présence d'un déictique (« le présent document », « nous »...) ou d'une formulation impersonnelle (« il faut... », « il est utile de... »), d'un présentatif (« commencer l'étude », « présenter », « expliquerons », « montrerons ») et de marqueurs d'intégration linéaire; des contraintes transphrastiques permettent d'extraire aussi les énoncés reliés. Dans les cas où le texte ne contient pas de plan de document explicité en tête ou repris en fin de document, ni de titres et de sous-titres, les annonces thématiques sont recherchées au fil du texte. Trois formes d'annonce locale sont ainsi identifiées : à l'aide des mêmes marqueurs déictiques ou impersonnels et présentatifs que pour l'annonce globale, à l'aide d'une question directe ou indirecte ou par l'entremise d'un soulignement .

Récapitulation, conclusion thématique

L'étiquette « récapitulation/conclusion thématique» [CAR 96] est attribuée aux phrases explicitant les enseignements et conclusions généraux du texte ; il s'agit là encore d'une information capitale puisqu'elle correspond à ce qu'il faudra retenir de l'argumentation de l'auteur. Cette classe d'informations comporte deux sous-classes : les récapitulations et les conclusions. Les énoncés récapitulatifs sont aisément identifiables au moyen de locutions *: pour nous résumer..., nous pouvons récapituler/résumer en disant..., en résumé, en guise de récapitulation...,* cependant, la lourdeur même de ces expressions fait qu'elles sont assez rares. Les énoncés conclusifs comprennent deux types principaux de marqueurs, dont les uns sont non ambigus et les autres fortement ambigus. Nous en donnons quelques exemples[16] :

> *(1)* ***Il faudrait donc*** *utiliser toutes les énergies disponibles car pour empêcher l'effet de serre il faut que l'emploi des énergies qui le favorise soit limité de façon que le CO2 qu'elles produisent ne dépasse pas ce qui peut être résorbé par le cycle du carbone.(...)*
>
> *(2)* ***Notre deuxième conclusion****, est que, à cause de l'effet de serre, l'intérêt de développer l'électronucléaire est devenu évident à un certain nombre d'hommes politiques, d'industriels et de scientifiques de disciplines diverses. (...)*

16 Tous les exemples de cet article proviennent du corpus utilisé dans le projet SERAPHIN, identifiés par le nom de l'auteur. Les indicateurs sont indiqués par une police en gras et les indices par le souligné. Les positions de la phrase et du paragraphe qui contiennent ces marqueurs sont aussi des indices.

(3) ***Donc****, pour que le développement de I'electronucléaire ait une influence significative,* ***il faudra*** *qu'il soit très important. Ceci est* ***notre troisième conclusion.****(...)*

(4) ***De toute façon*** *il sera évidemment nécessaire de freiner l'augmentation de la consommation d'énergie puisque les réserves de combustibles fossiles sont limitées: elles représentent quelques dizaines d'années pour le pétrole, 60 à 100 ans pour le gaz naturel et plusieurs siècles pour le charbon.(...) (BERTIN)*

Pour lever l'ambiguïté des connecteurs conclusifs (1, 3) ou reformulatifs (4), des contraintes requièrent la co-présence d'un marqueur de soulignement (2) ou d'un modal aléthique (1,3,4) et précisent la position du connecteur dans la phrase (1,3).

Définitions

Les énoncés définitoires (5, 6) constituent un autre type d'information recherché. Des règles, élaborées par E. Cartier [CAR 97] permettent d'extraire les différentes formulations d'une définition, ainsi que l'énoncé converse, la dénomination (7) :

(5) (a) Vapeur d'eau, gaz carbonique, monoxyde de carbone, méthane, chlorofluorocarbures, oxydes d'azote et ozone sont ce que l'on appelle communément des "gaz a effet de serre". (b) Sous ce vocable sont regroupés les gaz qui laissent passer le rayonnement solaire incident mais qui absorbent les rayonnements infrarouges de grande longueur d'onde renvoyés par la surface de la Terre, les empêchant ainsi de s'échapper vers l'espace. (LAMBERT)

(6) L'effet de serre est un phénomène naturel : la couche supérieure de l'atmosphère, composée d'eau et de gaz, absorbe, comme la vitre d'une serre, une partie des rayons infrarouges émis par la Terre. (NOYER)

A côté de ce type d'énoncés, un ensemble de règles qui, sur la base d'un groupe nominal du titre, extraient certaines prédications à son propos, présentent l'intérêt de pallier à l'absence des annonces thématiques classiques.

Argumentation causale et argumentation par la cause

La sélection de phrases est également fondée sur un autre type d'informations l'information causale [JAC 98]. Plus précisément, elle est fondée sur l'identification de données causales exprimées dans deux contextes argumentatifs particuliers : l'argumentation causale et l'argumentation par la cause. Dans le premier cas, la cause participe à l'expression et à la construction d'un savoir nouveau (le lien causal est vu comme une thèse restant à confirmer ou à infirmer). Dans le deuxième cas, elle joue le rôle d'argument, particulièrement pertinent car fondé sur le réel, pour justifier des choix et évaluations effectués, ou pour légitimer des objectifs et projets futurs. Dans les deux cas, le raisonnement causal fournit une information précieuse car synthétique et directement exploitable, qui enrichit la connaissance et sert de guide pour l'action.

Dans l'argumentation causale, l'identification du ce contexte est fondée sur des indices (emploi du conditionnel, de verbes modaux, etc.) exprimant le caractère hypothétique, possible, démontré,

largement admis, ou encore certain de l'information causale présente dans la phrase. L'exemple (8) illustre ce type d'organisation :

> *(8) Selon l'UNICEF, l'écotaxe **aurait** un **effet** pervers **sur** l'économie : elle **entraînerait** une baisse de la compétitivité et de la capacité à créer des emplois.*

Dans l'argumentation par la cause, le contexte qui dépasse habituellement le cadre de la phrase, contient des indices relatifs, d'une part à l'évaluation des conséquences (réelles ou possibles) et d'autre part, à l'action envisagée. L'information causale exprimée n'est plus centrale, elle sert à montrer que cette action est possible et à montrer comment l'opérer (cf. exemple (9)).

> *(9) Les risques de changement climatique **consécutif à** un accroissement de l'effet de serre ont conduit la France à se fixer un objectif volontariste de prévention des émissions de gaz à effet de serre et à proposer un accord international sur les moyens de prévention.*

2.2.7. Synthèse sur les méthodes par extraction

La diversité des méthodologies (et des algorithmes correspondants) employées pour identifier et classer les unités textuelles, en général les phrases, d'un texte en vue de produire un résumé est révélateur d'une difficulté théorique, que l'on peut résumer de la manière suivante : il n'existe pas de critères reconnus, et par conséquent de procédure de décision, qui permettent de caractériser la saillance d'une unité textuelle dans un texte.

Un autre constat est relatif au principe même d'extraction : une unité textuelle n'est jamais indépendante des autres unités textuelles qui composent un texte. Comme l'ont fait remarquer de nombreux linguistes, et comme nous l'avons déjà nous-même souligné, un texte obéit à des principes de cohérence et de cohésion marqués lexicalement, grammaticalement et structurellement. Toute rupture de ces liens sémantiques et structurels entraîne non seulement une perte d'informations mais aussi une falsification du contenu du texte. Le choix du paragraphe comme unité textuelle, qui constitue une tentative pour réduire ces distorsions, n'a pas donné lieu à des résultats plus convaincants.

Le principe des méthodes par extraction doit donc être considéré comme un moyen qui permet à la fois de réduire la complexité du problème cognitif que constitue l'activité résumante, et de contourner les problèmes linguistiques inhérents aux méthodes basées sur la compréhension du texte.

Il convient aussi de souligner que les systèmes développés ces dernières années ont en général chercher à mieux prendre en compte les connaissances linguistiques. La problématique du résumé automatique a ainsi glissé des sciences de l'information aux sciences du langage.

Chapitre 3

Evaluer les systèmes de résumé automatique

L'évaluation de résumés a souvent été abordée sous l'angle de l'informatique documentaire [SAL 89] en utilisant notamment des critères comme le rappel et la précision. Rappelons que ces deux mesures sont calculées, pour une requête d'un utilisateur qui cherche des documents dans un fonds documentaire, à partir des trois paramètres suivants :

P, nombre de documents non pertinents fournis par le système ;
Q, nombre de documents pertinents fournis par le système ;
R, nombre de documents pertinents présents dans le fonds documentaire et non fournis par le système.

Le rappel et la précision sont alors calculés comme suit :

Rappel = Q / (Q + R)
Précision = Q / (P + Q)

Pour appliquer ce mode d'évaluation aux systèmes de résumé automatique, on considère non plus le nombre de documents pertinents, mais le nombre de phrases pertinentes, c'est-à-dire les phrases qui devraient être placées dans le résumé.

Le principal défaut de ces critères est de postuler l'existence d'un résumé type, construit avec des phrases extraites du texte source, qui pourrait être utilisé comme référence. Or plusieurs expériences remettent en cause cette hypothèse. Par exemple, Teufel [TEU 97] de l'Université de d'Edimbourg a constitué un corpus de 202 articles en anglais, accompagnés de leurs résumés-auteurs, choisis dans le domaine du traitement automatique du langage naturel, et extraits d'un site WEB. La longueur moyenne d'un article est de 210 phrases, et celle du résumé de 4,7 phrases. Chaque article a été codé semi-automatiquement, au format SGML, et nettoyé de ses figures, tableaux, équations et références bibliographiques. Deux paramètres, appelés *Gold Standard A et B* , ont été élaborés en vue de quantifier la présence d'une phrase dans le texte source et dans le résumé de celui-ci :

- *Gold Standard A* (Alignement) quantifie les phrases qui figurent à la fois dans le résumé de l'auteur et dans le texte source et qui sont strictement identiques ;

- *Gold Standard B* (Jugement Humain) quantifie les phrases qui figurent à la fois dans le résumé de l'auteur et dans le texte source mais qui présentent des différences, tout en étant jugées similaires par un évaluateur humain. Les auteurs donnent l'exemple suivant de deux phrases jugées comme alignées, au sens du *Gold Standard B* :

  Phrase présente dans le résumé auteur : *In understanding a reference, an agent determines his conficdence in its adequacy as a means of identifying the referent*.
  Phrase du texte source *: An agent understands a reference once he is confident in the adaquacy of its (inferred) plan as a means of identifying the referent*.

Le résultat est le suivant : seules 17,8 % des phrases des résumés auteurs ont pu être alignées (*Gold Standard B)* avec les phrases du texte source, et seulement 3% des phrases étaient identiques (*Gold Standard A*).

Une autre expérience réalisée par [KUP 95) faisait état d'un score de 79 % (*Gold Standard A)*, mais dans cette expérience les résumés avaient été rédigés par des professionnels (il ne s'agissait donc plus de résumés-auteurs proprement dits). Enfin une évaluation réalisée par [RAT 61], citée par Kupiec, montre des résultats encore plus révélateurs. Ainsi le taux de recouvrement (calculé en nombre de phrases identiques qui apparaissent dans des résumés d'un même texte) entre des résumés réalisés par quatre résumeurs professionnels est de 25% ; le taux de recouvrement entre deux résumés d'un même texte, mais effectué à 6 mois d'intervalle, par le même résumeur, est de 55%.

Néanmoins, l'évaluation étant devenue une quasi-obligation dans le domaine du traitement automatique du langage naturel [SPA 96], tous les concepteurs de système de résumé décrivent des procédures d'évaluation. En dehors de deux expérimentations que nous allons décrire, ces évaluations sont souvent conduites sur des corpus extrêmement peu significatifs. Ainsi la taille des textes qui constituent l'échantillon dépasse rarement une page et le volume de l'échantillon se réduit à une dizaine de textes. L'explication de ceci tient généralement, en dehors du problème soulevé précédemment, au coût de mise en œuvre d'une procédure d'évaluation et à la difficulté de définir un protocole adéquat. Enfin, remarquons que dernièrement certains protocoles d'évaluation se sont focalisés sur la mesure des caractéristiques spécifiques d'un résumé produit automatiquement comme la fidélité [SAG 20] ou la pertinence thématique [MAN 98].

Nous allons décrire deux protocoles qui ont été proposés en vue de fournir un cadre d'évaluation générique, celui mis en place conjointement par le GRETS[17] et l'équipe LaLIC du CAMS, pour évaluer les résultats du système SERAPHIN [MIN 97] et celui proposé par le Defense Advanced Research Projects Agency (DARPA) décrit dans [HAN 97] dans le cadre du programme TIPSTER[18].

[17] Le GRETS est un département à la Direction des Etudes et des Recherches (DER) de l'EDF.
[18] Cette description s'appuie aussi sur la présentation qu'en fait N. Masson dans (Masson 98).

3.1. Protocole FAN du GRETS et de l'équipe LaLIC

Ce protocole vise à évaluer la qualité d'un résumé indépendamment du texte source et de son contenu informatif. L'évaluation est réalisée par des jurés, non spécialistes des domaines traités, qui lisent les résumés sans avoir eu connaissance des textes sources. La grille d'évaluation comporte 4 critères que nous décrirons ci-après, en les illustrant par des exemples fournis par les auteurs du protocole.

Critère 1 : Nombre d'Anaphores Privées de Référent

Puisqu'un résumé est construit à partir des phrases du texte source, il est possible qu'une phrase contienne une anaphore dont le référent n'appartienne pas à la phrase extraite. Selon les auteurs du protocole, ce critère semble déterminant pour la lisibilité du résumé ; néanmoins plusieurs problèmes d'évaluation ont été soulevés par les jurés. En voici un exemple :

> *On peut penser que ces premières relations entre les acteurs EDF et le riverain ont constitué un précédent créant un terrain peu propice à l'établissement de relations sereines entre EDF et la population locale riveraine de la ligne*.(Texte N°9)

Le terme potentiellement anaphorique « *ces »* peut s'interpréter comme référant soit à des relations décrites précédemment dans le texte, soit à un exposé chronologique qui prend sa signification à la lecture du texte complet et en partie à la fin de la phrase en question. Dans le texte source, c'est la deuxième interprétation qui est valide, mais le juré n'a pas les moyens de lever l'incertitude.

Critère 2: Rupture des segments textuels organisés par les Marqueurs d'Intégration Linéaire

Ce critère s'appuie sur les différents travaux de linguistique textuelle [CHA 89, ADA 90] qui ont souligné l'intérêt d'un repérage de marqueurs linguistiques pour identifier les organisations discursives qui dépassent la phrase. L'exemple ci-dessous illustre une rupture dans la séquence des marqueurs d'intégration linéaires utilisés par l'auteur pour organiser son argumentation.

> *Quels sont les atouts du charbon qui font que ce scénario de croissance se crédibilise ? En premier lieu, les réserves. [..] [..] En troisième lieu l'économie du charbon.*

Critère 3 : Présence de phrases "tautologiques".

Une phrase est considérée comme tautologique si son apport informationnel est complètement indépendant du texte source, comme dans l'exemple suivant :

> *Prédire l'avenir est un exercice difficile et incertain. (Texte N° 4)*

Ce critère s'est avéré être beaucoup trop dépendant de la connaissance que le lecteur a du domaine traité par le texte, et les résultats du protocole ont montré qu'il n'était pas pertinent.

Critère 4 : Lisibilité du résumé.

Ce critère dont les valeurs sont *Très mauvaise, Médiocre, Bonne,* et *Très Bonne* est une appréciation globale du résumé. Bien que très subjectives, les notes attribuées par les deux jurés, dans l'évaluation du système SERAPHIN, divergeaient peu.

3.2. Protocole MLUCE du GRETS et de l'équipe LaLIC

L'ambition de ce protocole est de faire évaluer la qualité des résumés automatiques par des utilisateurs potentiels de ceux-ci. Ici, la qualité du résumé ne sera pas définie de manière absolue mais en fonction de l'utilisation que veut en faire le lecteur. Par exemple, si un lecteur cherche l'idée « prouvée » dans un texte, il aura besoin d'en comprendre l'enchaînement argumentatif ; en revanche, s'il souhaite seulement relever la cooccurrence de deux thèmes, cet enchaînement n'est plus indispensable. C'est pourquoi la mesure de « qualité » dépend de l'objectif espéré du résumé [RAT 61]. Il faut donc, au préalable, définir un (ou plusieurs) type(s) d'utilisation de ces sorties et pour chacun, une mesure adéquate de la "distance" entre le texte source et son résumé. Le protocole MLUCE cherche alors à mesurer comment un résumé automatique répond à ces deux objectifs. Deux applications des résumés automatiques de textes ont été retenues.

Application 1 : le résumé est un outil permettant de décider de la lecture ou non d'un texte source.

Pour la première application, les critères définis dans MLUCE visent à évaluer l'intérêt du résumé en tant qu'instrument de décision adéquat pour un lecteur : ils doivent permettre de juger si le résumé contient les informations nécessaires pour décider ou non de lire le texte source. Pour cela il faut :

- Pouvoir identifier le domaine ou la nature du texte source. Chaque lecteur remplit deux grilles (une pour le texte source, une pour son résumé) dans lesquelles figurent les domaines ou natures de textes : scientifique ou technique, politique, sociologique, polémique, général, prospectif, rétrospectif, état des lieux ou état de l'art.
- Vérifier la présence des idées essentielles. Chaque lecteur surligne dans le texte les idées qu'il considère comme essentielles, puis vérifie qu'elles sont présentes dans le résumé. Les résultats du surlignage des « idées essentielles » dans le texte source et de l'indication faite par le lecteur des "idées parasites" présentes dans le résumé, sont regroupés pour définir un indicateur de « proximité ». Cet indicateur est calculé de la manière suivante :

  - un résumé est *proche du texte* si plus de 75% des phrases le constituant sont parmi les idées essentielles (surlignées) et moins de 10% sont des idées parasites ;
  - un résumé est *assez proche du texte* si entre 50% et 75% des phrases le constituant sont parmi les idées essentielles et moins de 10% sont des idées parasites ;

- un résumé est *assez éloigné du texte* si entre 25% et 50% des phrases le constituant sont parmi les idées essentielles et moins de 10% sont des idées parasites ;
- résumé est *éloigné du texte* dans les autres cas.

- Eviter les idées parasites. Chaque lecteur indique les phrases du résumé R_i qui ne devraient pas figurer dans R_i et les phrases du résumé R_i coupées du contexte (idées essentielles tronquées).

Application 2 : le résumé est un support de rédaction, d'une synthèse d'un document écrit.

Deux critères complètent les deux premiers critères définis précédemment :

- repérer l'enchaînement des idées. Chaque lecteur remplit deux grilles (une pour le texte source, une pour son résumé) dans lesquelles figurent les enchaînements argumentatifs suivants : *cause implique conséquence, conséquence induit cause, proposition de solution, du particulier vers le général, du général vers le particulier, juxtaposition de faits motivée, énumération de faits, confrontation.* Puis il énonce l'idée « prouvée » dans chacun des documents lus.
- Évaluer si le résumé est *clair, assez clair, peu clair, incompréhensible*.

Soulignons que la mise en place de ces deux protocoles pour l'évaluation du système SERAPHIN sur un corpus de 27 textes, dont la taille variaient entre 3 et 12 pages, a nécessité huit mois de travail et a mobilisé 6 jurés.

3.3. Protocole de la DARPA

Le protocole se décompose en deux tâches :

- Une tâche de catégorisation dont la fonction est de vérifier que le résumé peut être utilisé comme instrument de routage automatique vers une catégorie de lecteurs. Les jurés doivent évaluer si, à la lecture du résumé, ils peuvent déterminer les thèmes abordés dans le texte source.
- Une tâche de recherche d'informations ; à partir d'une requête utilisateur sur un thème donné, parmi cinq possibles, le système doit produire un résumé du texte dont la pertinence est évaluée.

Pour ces deux tâches, des critères de mesure ont été définis : un critère quantitatif qui mesure le temps de décision du lecteur pour effectuer la catégorisation ou pour décider de la pertinence et un critère qualitatif qui permet aux jurés d'évaluer le contenu du résumé comparativement au texte source. Ce protocole a été testé sur des documents issus des tests TREC (Text Retrieval Conference) [HAR 96] avec des résumés dont la taille était fixée à 10% du texte source.

3.3. Conclusion

Comme le montrent en partie les protocoles mis en place pour évaluer les systèmes de résumé, les recherches actuelles s'éloignent de la volonté de produire des résumés standards. Elles reposent plutôt sur une meilleure prise en compte des besoins de l'utilisateur, rompant ainsi avec les règles en usage dans les sciences de l'information qui ont cherché à imposer le résumé standard ou résumé-auteur. Ce type de résumé, comme son nom l'indique, considère le texte du point de vue de son auteur et par conséquent cherche à présenter au lecteur les thèmes généraux abordés dans le texte. A l'extrême, et il en est de fait souvent ainsi, le résumé-auteur d'un article scientifique est d'une telle généralité qu'il n'apporte aucune information sur le contenu réel du texte. En fait, en cherchant à répondre à tous les utilisateurs potentiels, ce résumé ne satisfait aucun utilisateur. Au contraire, un résumé construit comme une réponse spécifique aux besoins d'un lecteur se focalisera sur des thèmes ou sur des segments textuels qui contiennent les informations recherchées.

La problématique du résumé s'est ainsi déplacée depuis quelques années vers la recherche d'une adéquation entre l'expression d'une requête d'un lecteur et l'identification d'informations dans un texte. Une information n'est pas importante en soi, mais uniquement relativement à l'attente d'un lecteur ; c'est ce que nous appelons le ***filtrage.*** Le problème reste néanmoins difficile puisque d'une part, il faut fournir au lecteur des outils plus puissants que ceux généralement utilisés en informatique documentaire (utilisation d'opérateurs booléens et de descripteurs) et d'autre part, le système doit être capable d'identifier certaines informations sémantiques contenues dans le texte.

Un autre axe de recherche consiste à ne plus considérer le résumé comme indépendant du texte dont il est issu. L'informatique et plus généralement les outils du multimédia fournissent en effet des fonctionnalités qui permettent d'offrir au lecteur les moyens de naviguer entre le résumé et le texte. Plutôt que de chercher à produire un résumé autonome en abordant des problèmes comme la résolution des anaphores, le repérage des liens de cohésion et de cohérence, l'objectif se déplace vers la production d'un texte réduit aux informations jugées saillantes pour le lecteur, et vers la construction de liens qui permettent au lecteur, au vu des informations partielles qui lui sont présentées, de fouiller, à la demande, le texte source.

Ce sont ces deux axes de recherche, d'une part le filtrage, et le développement d'outils inter-actifs de fouille de textes, qui nous ont amené à concevoir la plate-forme Filtext.

Deuxième partie

Conception d’une plate-forme d’ingénierie linguistique dédiée au filtrage sémantique d’informations dans des textes

Table des matières

Chapitre 1

Les besoins en filtrage d'information

1.1 Introduction

Comme nous l'avons montré dans le chapitre précédent, non seulement les évaluations réalisées sur certains systèmes de résumé automatique [HAN 97, MIN 97, JIN 98] mais aussi les travaux menés en collaboration avec les résumeurs professionnels [END 95] y compris ceux menés avec les résumés produits par ces professionnels [SAG 98], ont montré la difficulté à réaliser des résumés standards, c'est-à-dire construits sans tenir compte des besoins des utilisateurs.

A l'origine de ce constat, vient sans doute en premier lieu le fait qu'il n'existe pas de critères précis [SPA 93] pour déterminer ce qu'est un bon résumé. Par ailleurs, l'activité résumante des humains [FAY 89] a été fort peu étudiée du point de vue psycholinguistique. Par exemple, le résumé scolaire, qui vise à tester les capacités de paraphrasage et de synthèse des élèves, n'est pas conçu et organisé de la même façon que le résumé d'auteur. Les résumés sont aussi très différents selon les utilisateurs auxquels ils sont destinés. Ainsi, on ne produira pas le même résumé d'un article scientifique innovant si l'on doit adresser ce résumé à la direction générale, au service des brevets pour consultation juridique, au laboratoire de développement, aux services de presse grand public... Les résumés dépendent également des types de textes. On ne résume pas de la même façon un texte narratif, un article scientifique relatif à une science expérimentale, un article d'une science théorique ou d'un domaine spéculatif, des articles juridiques, etc. Il n'y a donc pas de résumé idéal qui serait indépendant des demandes des utilisateurs et des types de textes.

Cette expérience du résumé automatique nous a donc amenés à élargir le champ de nos recherches en visant non plus de simples résumés automatiques non ciblés, mais des *systèmes automatiques de filtrage d'informations,* adaptés aux besoins spécifiques d'une tâche d'identification. Divers travaux menés parallèlement dans l'équipe LALIC depuis plusieurs années nous ont en effet permis d'identifier une même problématique qui relève du filtrage d'informations. Nous montrerons ci-dessous sur

quelques exemples, qui correspondent à des travaux de recherches aboutis ou à des thèses en cours, le type de besoins qui relève selon nous de la problématique du filtrage d'informations.

Premier exemple :

La possibilité de travailler sur des textes numérisés permet à un chercheur d'aborder la pensée d'un auteur en recherchant dans ses œuvres des expressions linguistiques spécifiques. Ainsi, P. Gauvain [GAU 01] travaille sur la définition du changement dans l'œuvre d'Aristote. Rechercher dans l'œuvre de cet auteur les occurrences lexicales du terme « changement », ce que permettrait un outil de recherche de concordances comme un KWIC, générerait trop de bruit. Par contre, la possibilité de filtrer les phrases qui sont des définitions du changement[19] répond aux attentes de ce chercheur et plus généralement aux besoins des philologues, des philosophes, etc. Un outil de filtrage devrait donc être à même d'extraire des œuvres d'Aristote les phrases suivantes :

> « **Tout changement** va d'un terme à un autre (c'est aussi ce que [en grec] montre, le mot en effet il exprime une succession, c'est-à-dire la distinction d'un antérieur et d'un postérieur). » [Aristote, *Physique* V, 1, 224 b 35, traduction H. Carteron, *Les Belles Lettres.*]
>
> « Tout mouvement est un **changement**. » [Aristote, *Physique* V, 1, 225 a 34 , traduction H. Carteron, *Les Belles Lettres.*]
>
> « **Tout changement** va d'un terme à un autre terme, aussi bien le changement dans la contradiction que le changement dans les contraires. » [Aristote, *Physique* VI, 10, 241 a 27, traduction H. Carteron, *Les Belles Lettres*.]

Deuxième exemple :

Le traitement des lettres de réclamation (cf. fig. 1) dans une grande entreprise [NAV 01a, 01b] nécessite dans un premier temps le traitement du type de phrase suivant :

> « Je me permets de solliciter auprès de vous des délais de paiement car je rencontre actuellement des difficultés financières passagères ».

Il convient d'une part, d'identifier le type de requête, marqué dans cet exemple par l'expression linguistique « Je me permets de solliciter », et d'autre part d'identifier le motif de la requête les « délais de paiement ». Dans un deuxième temps, il faut rechercher dans la mémoire de rédaction si un même type de réclamation a déjà été traité, et si c'est le cas, proposer en modèle la lettre de réponse correspondante.

[19] Ces exemples de phrases m'ont été fournis par P. Gauvain, doctorant au CAMS.

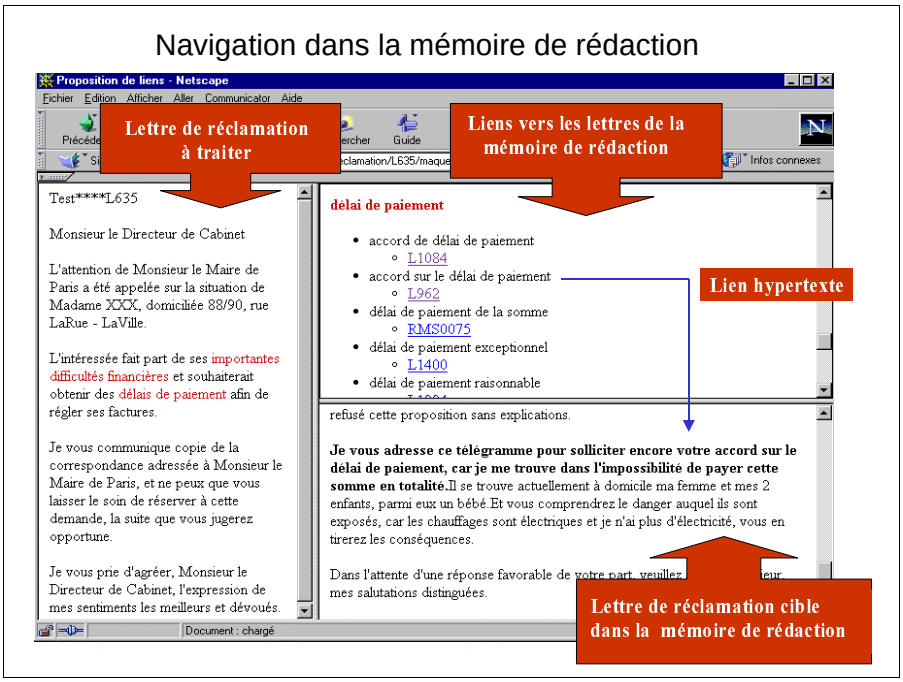


Figure 1. Traitement des lettres de réclamation [NAV 01]

Troisième exemple :

Les textes journalistiques ou scientifiques font souvent appel à des citations pour étayer une argumentation, ou pour définir un terme. Dans ce cas, la citation accompagne les segments textuels extraits ou filtrés qui constituent le résumé dans un cadre de recherche d'informations ciblée. Une citation peut comporter toutes sortes d'informations. Si on étudie les marqueurs de la définition [CAR 97] (exemple (1) citation indirecte et (2) citation directe) ou les marqueurs de la causalité [JAC 98] (exemples (3) et (4)), on se rend compte que la citation en soi peut être la définition d'un terme, constituer une relation de causalité, correspondre à un soulignement par l'auteur par des procédés d'argumentation, etc. Dans les exemples (3) et (4), les marqueurs de causalité sont en ***gras italique***. Les **gras** Times et **PETITES CAPITALES GRAS** sont les marqueurs de la citation[20].

> 1) Le même **AUTEUR** ***définit***, par opposition, le conflit comme une situation dans laquelle les parties sont conscientes de l'incompatibilité de positions futures potentielles et dans laquelle chaque partie désire occuper une position qui est incompatible avec les désirs de l'autre.
>
> 2) ***Définissons***, avec **JULIEN FREUND**, la politique **:** "l'activité sociale qui se propose d'assurer par la force, généralement fondée sur le droit, la sécurité extérieure et la concorde intérieure d'une unité politique particulière..."
>
> 3) On a entendu, plus tard , lors des entretiens avec la fillette **: « MON** père m'angoissait quand il rentrait. Il ne disait pas un mot. **»** Et c'était vrai **!** Mais la

[20] Ces exemples sont extraits des travaux de G. Mourad [MOU 00], doctorant au CAMS.

> fillette ***établissait un rapport de causalité entre*** son père ***et*** son angoisse **: « IL** m'angoissait dès qu'il rentrait », alors que son angoisse ***trouvait son origine dans*** l'enfance de la mère. (B. Cyrulnik, Sous le signe du lien, 1989)

> 4) Ancien **PROFESSEUR** à l'université du Rwanda, **PIERRE ERNY,** dans Clés pour comprendre le calvaire du peuple (2), ***s'interroge*** sur les ethnies, la révolution de 1959, puis la montée de l'extrémisme Hutu et de la rébellion tutsie. ***Pour lui****,* les offensives répétées des combattants du front patriotique rwandais, à partir de 1990, ***ont été*** **«** la ***cause*** déclenchante **»** du climat de violence et de psychose qui ***a débouché sur*** la guerre. (corpus « Le Monde Diplomatique »)

Comme on le voit sur ces exemples, la recherche de citations nécessite le repérage de marques spécifiques, mais cette recherche peut aussi être combinée avec le repérage d'autres expressions linguistiques, ce qui permet d'établir une typologie des citations [MOU 00].

Quatrième exemple :

L'extraction de connaissances à partir de textes. Un très grand nombre de textes sont analysés par des experts humains en vue d'acquérir puis d'organiser les connaissances décrites dans un texte. C'est le cas par exemple des analyses d'interviews d'experts dans lequel il est nécessaire de repérer à la fois les concepts utilisés par l'expert et les relations qui lient ces concepts entre eux. Les systèmes SEEK [JOU 93] et SEEK-JAVA [LEP 00] sont deux exemples de systèmes dédiés au repérage des relations statiques, ainsi que le système COATIS [GAR 98] qui repère l'organisation causale des actions. Ces systèmes doivent résoudre deux types de problèmes.

Le premier problème concerne l'identification des concepts qui ressort de la terminologie et plus précisément, bien qu'il ne se réduise pas à cette tâche, de l'identification des syntagmes nominaux présents dans le texte. Un outil comme LEXTER [BOU 94], qui s'appuie sur le repérage de patrons morpho-syntaxiques et un apprentissage endogène pour résoudre, en partie, le délicat problème de la délimitation des frontières des syntagmes nominaux, constitue une aide appréciable. Une autre approche, comme celle du système IOTA [CHI 86] fondée sur des techniques de classification, permet de construire automatiquement le thésaurus d'un domaine à partir d'un corpus de textes. Les résultats fournis par ces systèmes sont des listes de syntagmes nominaux. Par exemple sur le texte suivant, un logiciel comme LEXTER fournit la liste présentée dans le tableau 1.

Le deuxième problème à résoudre est de retrouver certaines relations qui lient les syntagmes nominaux entre eux. Dans le cas du texte précédent, que l'on peut qualifier de texte descriptif, une première étape est, par exemple, d'identifier les relations statiques entre les syntagmes nominaux identifiés précédemment. Ce sont des relations [DES 87] de localisation, d'ingrédience, d'attribution, etc. La représentation graphique (cf. fig. 2) de ces relations fournit une aide substantielle à l'analyste qui désire appréhender le réseau conceptuel des connaissances mis en place par l'expert du domaine.

Le Myosotis des Alpes a les caractéristiques suivantes : - *plante atteignant 30 cm de hauteur, à tiges couvertes de poils rudes ;* - *feuilles caulinaires sessiles, ovales à linéaires ;* - *fleurs atteignant 7 mm au calice au moment de la fructification, s'amincissant au pédoncule ;* - *corolle largement étalée de 9 mm de diamètre ;* - *floraison de mai à septembre.* *On trouve le Myosotis des Alpes dans les éboulis, les pelouses lacunaires, au-dessus de 1500 m d'altitude. Le Myosotis des Alpes est une espèce voisine des myosotis des bois, que l'on trouve en prairies, brousses et forêts d'altitude, jusque vers 1500 m.*

Myosotis des Alpes
tiges couvertes de poils rudes
feuilles caulinaires sessiles

Tableau 1. Exemples de syntagmes nominaux extraits par LEXTER.

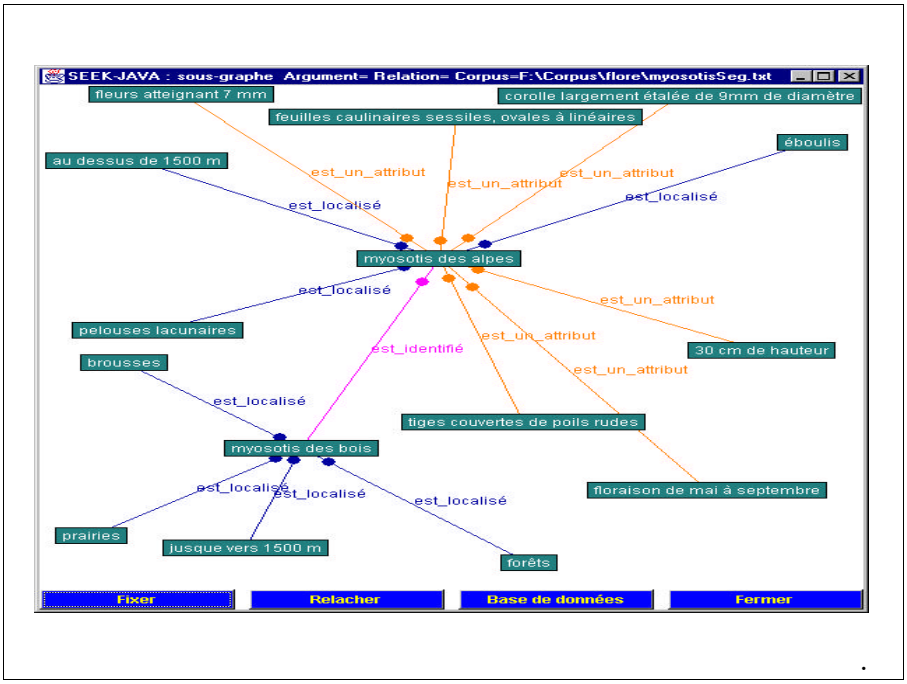


Figure 2. Exemples de relations statiques [LEP 00]

Ces divers exemples nous ont permis d'illustrer une problématique commune : comment identifier dans des textes, indépendamment des domaines traités, certaines des relations organisatrices des connaissances exprimées par l'auteur. Nous identifions cette problématique comme relevant du filtrage d'informations. Avant de présenter les choix conceptuels et techniques qui nous ont permis de

répondre à cette problématique nous voudrions montrer auparavant en quoi ceux-ci se différencient radicalement de la problématique de l'extraction d'informations.

1.2. Un contre-exemple au filtrage d'informations : l'extraction d'informations

Comme le soulignent T. Poibeau et A. Nazarenko dans [POI 99], c'est en partie l'échec des systèmes de compréhension automatique de textes du point de vue de leur généricité et de leur efficacité qui a amené les différentes équipes de recherche à s'orienter vers des systèmes d'extraction d'informations[21]. A l'origine, l'objectif des systèmes d'extraction d'informations était le repérage d'informations factuelles dans des textes. Mais l'organisation des conférences « Messages Understanting Conferences » (MUC)[22] a profondément modifié cette problématique. En effet, ces conférences, en imposant un protocole d'évaluation aux systèmes mis en compétition ont, de fait, introduit plusieurs biais :

Le premier biais vient du type de textes sur lesquels sont évalués les systèmes. En effet, chaque conférence fournit un corpus de référence dans un domaine spécialisé. Il s'agit de dépêches d'agence de presse qui décrivent des faits de terrorisme pour les conférences MUC3 et MUC4, des faits économiques pour la conférence MUC5, etc. Le fait de travailler sur des textes d'un domaine oriente nécessairement les travaux de recherche vers l'acquisition de termes lexicaux spécialisés, c'est-à-dire vers l'étude d'un sous-langage.

Le deuxième biais vient du principe même de la méthode d'évaluation. Les organisateurs définissent pour chaque campagne MUC un formulaire standard. Ce formulaire est composé de différents attributs qui doivent être remplis par les systèmes d'extraction. La figure 3 donne un exemple du formulaire de la campagne MUC4 où il s'agissait d'analyser des textes qui décrivaient des attentats terroristes. Comme on le constate rapidement à la lecture de la première colonne, les attributs du formulaire ne font référence qu'aux « actants » et non aux prédications elle-mêmes.[23] Par exemple, l'analyse de la phrase suivante :

> *« Garcia Alvarado, âgé de 56 ans, a été tué par l'explosion d'une bombe placée sur son véhicule par une guérilla urbaine alors qu'il était arrêté à une intersection dans les faubourgs de San Salvador.»*

permet de compléter la deuxième colonne du formulaire. La prédication « a été tué par » se trouve ainsi réduite à ses conséquences, a savoir la mort d'un des actants, et les relations causales sous-jacentes sont totalement ignorées. Ce point est fondamental de notre point de vue, puisqu'il illustre

[21] Le système TACITUS [HOB 92] en est un exemple prototypique. Il a été développé par le Stanford Research Institute (SRI) pour la compréhension automatique de textes puis utilisé par le SRI pour faire de l'extraction d'informations.
[22] La première conférence MUC1 a été organisée en 1987 et la dernière MUC7 en 1998.

parfaitement la différence entre l'extraction d'informations et le filtrage sémantique d'informations tel que nous l'envisageons.

Incident : Date	
Incident : Lieu	El Salvador : San Salvador
Incident : Type	bombe
Auteur : Organisation	
Auteur : Type	guérilla urbaine
Cible Physique : Description	véhicule
Cible Humaine : Nom	Garcia Alvarado
Victime : Description	Garcia Alvarado
Victime : Effets : Morts **Blessés** **Indemne**	Garcia Alvarado

Figure 3. Formulaire de sortie de MUC3

Le dernier biais tient au fait que l'évaluation est calculée en utilisant des critères comme le taux de rappel et de précision, indépendamment de tout utilisateur et de toute interactivité avec celui-ci.

Ces différents biais ont conduit toutes les équipes à utiliser le même type d'approche et la même technologie : la constitution de bases de patrons morpho-syntaxiques et l'utilisation de cascades de transducteurs.

Le choix de l'approche par patrons morpho-syntaxiques est la conséquence des deux premiers biais. Puisque les textes traitent d'un seul domaine, il est plausible de vouloir recenser tous les termes lexicaux qu'un rédacteur peut utiliser, dans ce domaine, pour référer une « date », un « lieu », un « auteur », etc. Comme le souligne très bien T. Poibeau dans [POI 99], « le processus d'analyse ne part plus du texte, il est guidé par la connaissance à priori des informations recherchées». Les équipes de recherche ont ainsi développé des outils d'identification de groupes nominaux et d'entités nommées. Remarquons aussi que la structure du texte ne joue aucun rôle, ce qui signifie que l'ordre dans lequel les phrases apparaissent dans le texte pourrait être modifié sans que cela influe sur le résultat obtenu. Ceci est la conséquence d'une part, du choix du type de texte, focalisé sur des événements, et d'autre part, du choix d'un formulaire comme sortie du système. Ce choix d'un formulaire qui supprime toute référence à la chronologie des événements ne serait plus possible si les résultats à fournir avaient au contraire privilégié l'identification des phases saillantes du texte, organisées chronologiquement, comme le montre D. Battistelli dans [BAT 00].

La non prise en compte de l'utilisateur final et l'aspect mono-tâche est à l'origine de l'utilisation d'une technologie à base de cascades de transducteurs. Ces outils sont sans aucun doute les plus

[23] Cet exemple est tiré de [POI 99].

performants en terme de rapidité de traitement mais ils ont aussi, de notre point de vue, deux graves défauts.

D'abord, ils ne sont pas réutilisables sans l'intervention de leurs concepteurs ; par exemple, le système FASTUS comprend 95 patrons qui s'appliquent uniquement sur les textes d'acte de terrorisme. Cette multiplication des patrons, sans organisation linguistiquement fondée, les rend peu accessibles. Ensuite, ils occultent totalement le problème de la modélisation des données linguistiques ; à propos toujours du système FASTUS [APP 93], les concepteurs proposent la notion de « trigger word » ou d'amorce, pour déclencher les patrons d'extraction. Mais le choix de l'amorce se fonde uniquement sur des critères fréquentiels ; c'est donc un choix technologique qui vise à optimiser la reconnaissance des patrons dans un texte, et non un choix linguistique comme nous le défendons ci-après dans la méthode d'exploration contextuelle. Le texte à traiter est ainsi considéré comme une simple chaîne de caractères, sans structure. Même si il est possible d'intégrer dans les patrons morpho-syntaxiques la recherche de balises de structure de type XML pour prendre en compte la structure du texte, cela nécessite de la part du concepteur des patrons une gymnastique de l'esprit peu commune. Par exemple, construire un patron qui recherche une expression dans le dernier paragraphe de l'avant dernière section du texte n'est pas un exercice trivial. De même, exprimer dans un patron, une condition du type « SI dans la première phrase du paragraphe précédent, le patron P1 a été reconnu, ET SI cette phrase contient l'expression E1 Alors exécuter Action A» nécessite la construction d'une cascade de transducteurs difficilement gérable. Nous reviendrons sur ce problème, fondamental, lorsque nous présenterons les langages d'expression des règles d'exploration contextuelle qui ont été définis dans le cadre de la plate-forme Filtext.

Les systèmes d'extraction d'informations constituent selon nous un contre-exemple au filtrage d'informations, au sens où ils imbriquent les connaissances linguistiques dans le système informatique, excluant ainsi toute possibilité d'enrichissement incrémental du système d'une part et toute adaptation à des aux besoins spécifiques d'un utilisateur.

C'est dans le but de répondre aux finalités spécifiques du filtrage d'informations que nous avons participé au développement de la méthode d'exploration contextuelle et à la conception d'une plate-forme conceptuelle, Filtext, et logicielle, ContextO. Nous allons d'abord présenter cette méthode, en insistant sur les principes linguistiques qui la sous-tendent puis nous décrirons l'architecture du système informatique qui permet de la mettre en oeuvre.

Chapitre 2

La méthode d’exploration contextuelle

2.1. Méthodologie : la méthode d'exploration contextuelle

La méthode d’exploration contextuelle vise à se donner les moyens d’accéder au contenu sémantique des textes pour mieux les cibler et en extraire des séquences particulièrement pertinentes. Elle est issue d’une réflexion initiale sur le traitement informatique des valeurs aspecto-temporelles dans les langues avec une première réalisation informatique SECAT [DES 91a] pour tous les temps du passé indicatif en français. La méthode a été ensuite généralisée, en tant que système de décision, en tenant compte des informations présentes dans le contexte textuel pour un calcul des valeurs sémantiques relevant de différentes tâches. Il ne s’agit ni de l’utilisation de mots clés, ni d’une simple analyse distributionnelle. Cette méthode met en effet en jeu des processus inférentiels [DES 97b] qui sont déclenchés par l’identification d’indicateurs linguistiques relatifs à un champ grammatical ou discursif précis. C’est en ce sens que ces indicateurs deviennent des marqueurs de valeurs sémantiques.

Il est ainsi possible de dégager plusieurs principes fondateurs, énoncés au départ par J-P. Desclés [DES 91a, 93] et qui se sont affinés à mesure que différents travaux de recherches linguistiques et informatiques ont mis en oeuvre cette méthode.

2.1.1. Principes de l'exploration contextuelle

Un premier principe consiste à n'exploiter que les connaissances linguistiques présentes dans les textes, ce qui implique que des connaissances encyclopédiques, telles que des ontologies, des classifications conceptuelles, des réseaux sémantiques, etc., ne sont jamais utilisées. Ce principe constitue une des hypothèses cognitives que formulées par J-P Desclés [DES 91b] et que l'on peut illustrer par le constat suivant : un lecteur non spécialiste d'un domaine est tout à fait capable

d'identifier dans un texte certaines relations organisatrices de la connaissance ainsi que les organisations textuelles mises en place par l'auteur. Jusqu'à présent, faute de temps et de moyens, aucune expérience psycholinguistique n'a pu être élaborée en vue de confirmer ou d'infirmer cette hypothèse. Le projet « Modèle d'exploration sémantique de textes guidé par les points de vue du lecteur »[24] auquel nous participons nous offrira peut-être l'opportunité de construire un tel protocole.

Le deuxième principe repose sur la conception du langage qui est celle de J.-P. Desclés ; cette conception découle d'un modèle du traitement du langage naturel différent du modèle traditionnel. Dans ce dernier les différentes étapes, traitement lexical et morpho-syntaxique, analyse syntaxique et enfin analyse sémantique s'enchaînent séquentiellement. Les défauts de ce modèle ont été abondamment soulignés ; voici par exemple un commentaire de P. Zweigenbaum [ZWE 97] qui dans le cadre du projet HELENE, a mis en oeuvre une telle architecture. Il résume à la fois la position de principe adoptée et les problèmes soulevés par celle-ci :

> *« L'ambiguïté des énoncés induit un non-déterminisme qui pose des problèmes de complexité à l'analyse. Cette ambiguïté peut être constitutive, mais aussi artificielle, créée par le découpage en modules d'analyse et par leur chaînage. Ce découpage est pourtant souhaitable pour des raisons méthodologiques et pratiques : à la fois pour isoler et mieux étudier les problèmes qui se posent, et pour rendre le développement et la mise au point des algorithmes et des connaissances concernés. Par ailleurs, la contribution effective des différents niveaux d'analyse à la compréhension d'un texte est difficile à cerner. Déterminer une stratégie d'analyse optimale est de ce fait malaisé. »*

Contrairement, donc, à cette approche traditionnelle du TALN, J.-P. Desclés propose de déterminer d'abord la carte sémantique qui correspond à la tâche de traitement que l'on désire automatiser.

[24] Ce projet d'une durée de deux ans et est piloté par le LIMSI. Il réunit des chercheurs du LIMSI, de l'équipe LATTICE, du CEA et de l'équipe LaLIC du CAMS. Il est financé par le Ministère de la Recherche dans le cadre de l'action Cognitique 2000.

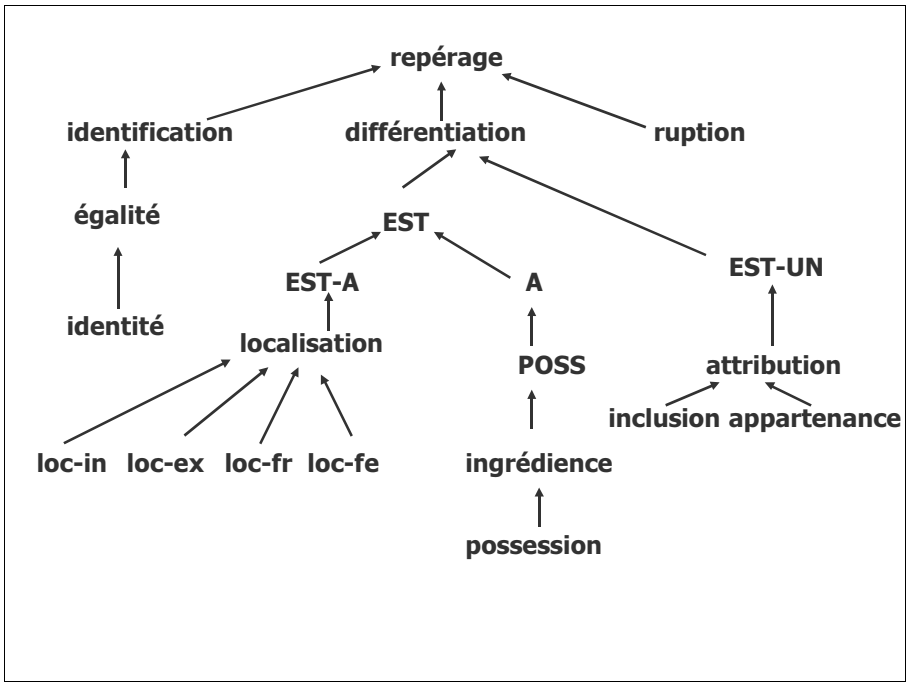


Figure 4. Un exemple de carte sémantique, le réseau des relateurs de repérage [DES 87]

Le travail linguistique de construction de cette carte sémantique consiste à identifier l'ensemble des valeurs sémantiques et à organiser ces valeurs sémantiques dans un réseau organisé. Le réseau des relateurs de repérage [DES 87] (cf. fig. 4) correspondant à la tâche d'identification des relations statiques est un premier exemple de carte sémantique. Un autre exemple de carte sémantique (cf. fig.5) nous est fourni par la thèse de Sabrina Baldo sur les valeurs du modal Would [BAL 00].

Ces deux exemples montrent l'importance du travail théorique à réaliser avant tout projet d'automatisation de la tâche et soulèvent une question essentielle : est-il toujours possible de construire une telle carte sémantique ? L'expérience acquise dans l'exploration contextuelle ne nous permet pas de répondre avec certitude à cette question. Par contre, nous avons pu constater qu'il est possible de s'appuyer sur cette méthode sans avoir au préalable élaboré une carte du domaine sémantique traité. Une voie intermédiaire consiste ainsi à se limiter à identifier des valeurs sémantiques, sans prétendre que ces valeurs couvrent exhaustivement le domaine et sans chercher à construire un réseau entre les concepts identifiés. C'est par exemple l'approche choisie par D. Garcia [GAR 98] et A. Jackiewicz [JAC 98] pour la causalité, J. Berri [BER 96a] pour le résumé automatique. Nous discuterons dans la suite de ce document des conséquences d'un tel choix.

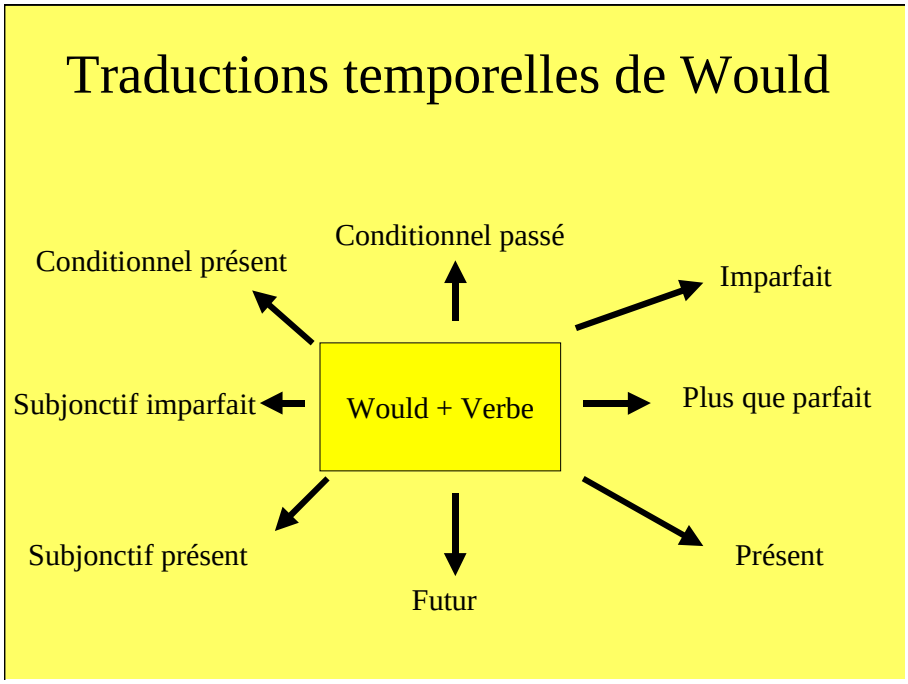


Figure 5. Un exemple de carte sémantique, les traductions du modal Would [BAL 00]

Les valeurs sémantiques ayant été identifiées et organisées, le linguiste s'engage alors dans la collecte des marqueurs discursifs explicites (morphèmes, mots, expressions et locutions...) qui expriment potentiellement ces valeurs (cf. fig. 6). Ces marqueurs, que l'on peut appréhender comme des déclencheurs d'un processus inférentiel [DES 97b], sont appelés les *indicateurs (ou indices déclencheurs)* linguistiques relatifs à un champ grammatical ou discursif précis. C'est en ce sens que ces indices deviennent des marqueurs de valeurs sémantiques car ils sont « des indices de quelque chose de non directement observable » [DES 01b]. Ainsi, la présence de ces indices permet, par un raisonnement de type abductif [DES 01a] (cf. fig. 7), de remonter à la signification qu'ils encodent.

Comme exemple de champ grammatical, donnons celui qui couvre l'identification des valeurs aspecto-temporelles associées aux morphèmes des temps grammaticaux du français. Pour le champ du discours, mentionnons par exemple les indicateurs discursifs des annonces thématiques, des expressions définitoires, des relations entre concepts, des relations de causalité, des relations temporelles entre événements, etc. Le linguiste regroupe ces indicateurs dans des classes d'indicateurs en fonction de critères sémantiques (relation de synonymie) et syntaxiques.

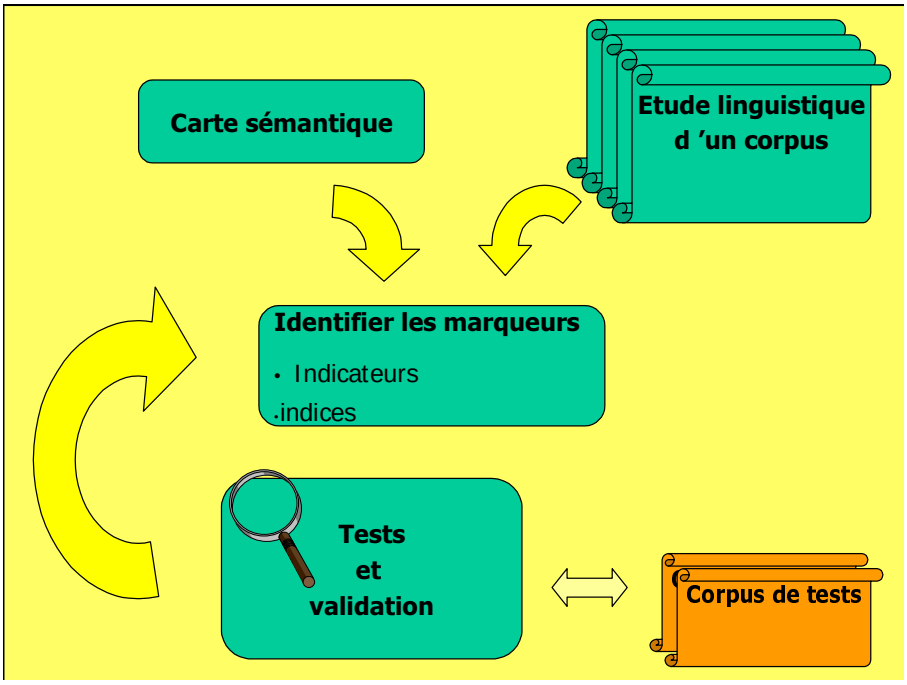


Figure 6. Cycle de vie de l'exploration contextuelle

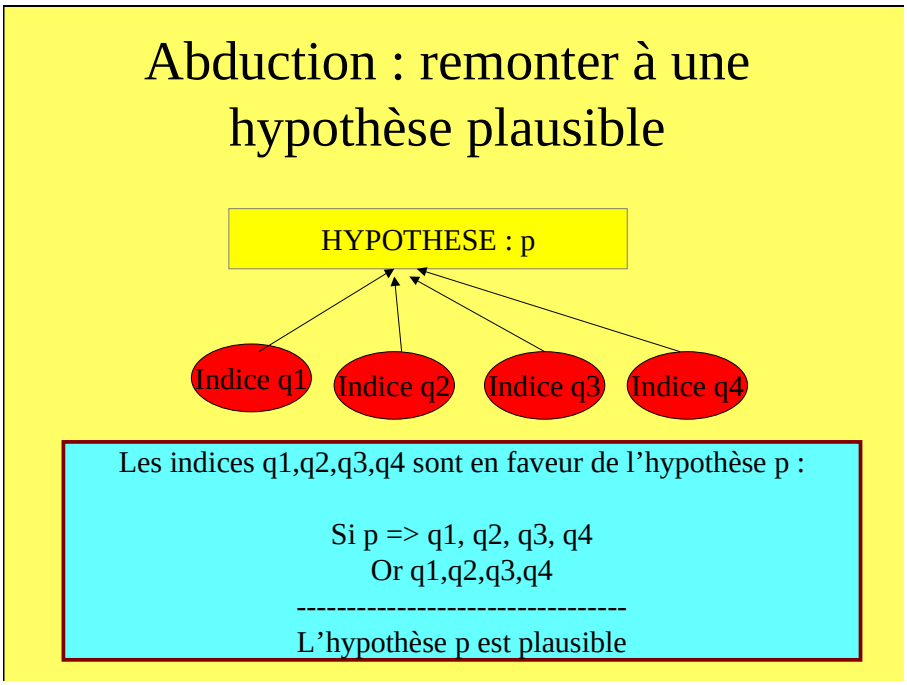


Figure 7. Indices et raisonnement par abduction [DES 01b]

L'identification d'un marqueur (grammatical ou discursif) n'est cependant pas suffisante pour déterminer complètement la valeur sémantique de ce marqueur. En effet, un indicateur linguistique est rarement un marqueur univoque d'une valeur sémantique unique. Le rapport entre signifiants et

signifiés n'est pas bijectif dans les langues, tout particulièrement pour les champs grammaticaux et discursifs. La plupart des marqueurs sont polysémiques. Ayant donc identifié une occurrence de marqueur sous la forme d'un indicateur répertorié, il faut, dans un deuxième temps, explorer le contexte de cette occurrence pour rechercher d'autres *indices* linguistiques, sous la forme d'occurrences d'indices complémentaires, qui viendront : i) soit lever l'indétermination sémantique attachée *a priori* au marqueur analysé, permettant ainsi de déterminer une valeur sémantique sous la forme par exemple d'une étiquette sémantique attribuée à une segment linguistique (syntagme, phrase, paragraphe selon les cas) ; ii) soit invalider les hypothèses sémantiques qui pouvaient être envisagées à propos du marqueur analysé dans son contexte. Ces indices sont eux aussi regroupés dans des classes, appelées classes d'indices, en fonction de critères sémantiques (relation de synonymie) et syntaxiques. Insistons sur le fait qu'un indice n'est pas nécessairement lexical, mais qu'il peut être :

- *ponctuationnel* ; dans ce cas, cet indice correspond à la présence d'une marque de ponctuation spécifique confirme la valeur de l'indicateur ;

- *dispositionnel* ; comme par exemple lorsque l'indice indique la place de la phrase, qui contient l'indicateur, dans le texte ;

- *structurel* ; dans ce cas, l'indice est un élément structurel du texte comme un titre, une énumération, etc. ;

- *discursif* ; dans ce cas, l'indice précise un type d'acte discursif comme une définition, une conclusion, etc.

Remarquons que M.P. Woodley dans [PER 00] propose une conception très proche, mais qu'elle ne fait pas de distinction entre indicateurs et indices :

> *« L'approche de la notion de marqueur développée (...) peut se résumer comme suit :*
> *– les marqueurs sont des configurations de traits régulièrement associées à une fonction ou à un objet textuel ;*
> *– ces configurations de traits caractéristiques regroupent des traits lexico-syntaxiques, typographiques, dispositionnels, ponctuationnels ;*
> *– l'équivalence fonctionnelle entre des traits "discursifs" et des traits "visuels " débouche sur l'affirmation du statut linguistique de ces derniers, le texte écrit étant un objet visuel ;*
> *– les marqueurs ne sont pas donnés mais doivent faire l'objet d'une procédure de découverte en corpus. » p. 117-118*

Nous allons illustrer ces différentes notions sur un exemple d'indétermination sémantique liée à un marqueur grammatical, le morphème de l'imparfait *-ait*, facilement identifiable[25].

[25] Cet exemple est un de ceux donnés par Jean-Pierre Desclés.

Prenons le segment textuel (ici, une proposition) :

... le lendemain, il démissionnait...

En dehors de tout contexte, deux valeurs référentielles contradictoires peuvent être attribuées à cette proposition : soit [il a effectivement démissionné] (valeur dite de «nouvel état» associée à l'imparfait), soit [il n'a pas démissionné] (valeur «irréelle» associée également à l'imparfait). Selon les contextes, il est possible d'identifier des indices linguistiques complémentaires qui contribueront à lever l'indétermination. Considérons l'insertion de la proposition *le lendemain, il démissionnait* dans deux contextes différents, avec les inférences qui s'en déduisent :

(1) *De nombreuses voix arrivèrent très rapidement pour soutenir sa proposition. Pourtant, le lendemain, il démissionnait...*
Il a effectivement démissionné (valeur de «nouvel état»)

(2) *Sans les nombreuses voix qui arrivèrent très rapidement pour soutenir sa proposition, le lendemain, il démissionnait...*
Il n'a pas démissionné (valeur d'«irréel»)

Les mots *pourtant* et *sans* ainsi que la ponctuation (un point, une virgule) sont autant d'indices linguistiques qui, combinés avec le morphème d'imparfait, orientent l'interprétation vers la valeur de «nouvel état» ou la valeur «irréelle». Une règle, associée au calcul des valeurs aspectuelles attachées à l'imparfait français, indiquera que la présence de ces indices permet de lever l'indétermination sémantique.

L'exploration du contexte est décrite à l'aide de règles (cf. fig. 8), appelées *règles d'exploration contextuelle*. L'exploration contextuelle est donc gouvernée par un ensemble de règles qui, pour une classe d'indicateur donnée et une décision à prendre, recherchent d'autres classes d'indices explicites dans le contexte textuel de l'indicateur. Formellement, une règle R_k est composée d'une classe d'indicateur K, d'un ensemble fini de couples (I_p, C_p) où I_p représente la pième classe d'indices à rechercher dans le contexte linguistique[26] C_p, et d'une décision D_k.

$$R_k = [K, \{I_p, C_p\}, D_k]$$

Remarquons que le terme de règle porte l'empreinte du contexte informatique dans lequel la méthode d'exploration contextuelle a été originellement conçue. En effet, le système SECAT [DES 91a], premier système opérationnel qui a mis en oeuvre la méthode d'exploration contextuelle, a été réalisé dans le modèle dominant de l'époque, à savoir les systèmes experts, généralisés par la suite par les systèmes à base de connaissances. Ces systèmes ont permis d'introduire la distinction entre la représentation des connaissances sous une forme déclarative et le système informatique, le moteur d'inférence qui applique ces connaissances. Le formalisme de représentation des connaissances le plus

[26] Dans la suite de ce mémoire, nous emploierons, sauf mention contraire, le terme de contexte au sens de contexte linguistique. Nous n'avons pas repris le terme de co-texte qui englobe des éléments sémiotiques.

utilisé à l'époque était le formalisme déclaratif des règles de production. Dans un premier temps, de nombreux avantages furent portés au crédit de cette représentation, notamment dans le domaine du processus industriel [LEV 90] ou de la réglementation [LEV 89]. Dans un deuxième temps, les difficultés rencontrées, notamment dans la gestion de la non-monotonie et de la granularité des connaissances représentées, firent que ce modèle connût un certain déclin. Reste que le principe de la déclarativité des connaissances indépendamment de leur mise en œuvre informatique constitue un acquis important.

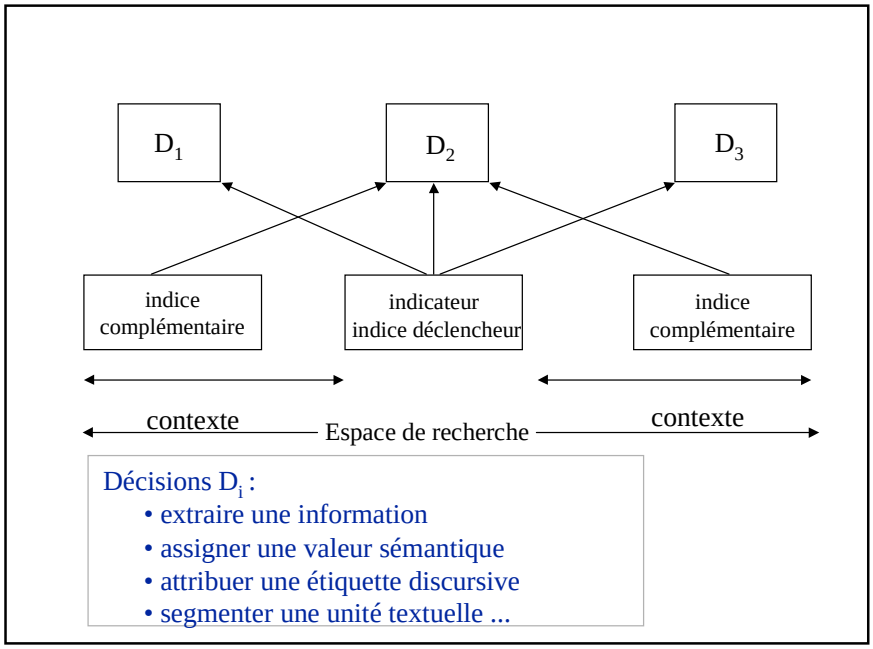


Figure 8. Les éléments d'une règle d'exploration contextuelle

Le travail de réflexion mené depuis plusieurs années dans l'équipe LaLIC a permis de préciser la notion de règle d'exploration contextuelle indépendamment de son origine technologique. Le terme de règle d'exploration contextuelle doit donc être compris comme une expression déclarative exprimée dans un langage spécifique, dont dispose le linguiste pour décrire une configuration textuelle dans laquelle une classe d'indicateurs K est représentative d'une des significations déclarées dans la carte sémantique. Une configuration textuelle est la réalisation dans un texte d'un des agencements possibles exprimés dans la règle. Il se caractérise donc, par la présence, d'une part, d'un indicateur qui appartient à la classe K et des indices I_p qui apparaissent dans les contextes C_p de l'indicateur. Le terme de règle d'exploration contextuelle ne doit donc plus être associé avec la notion de règle de production, et nous verrons qu'au fil du temps plusieurs langages ont été élaborés pour les décrire. Nous voudrions plutôt insister ici sur la notion de contexte qui fait partie du vocabulaire de l'exploration contextuelle, au même titre qu'indicateur et indice. Le fait d'avoir à préciser

explicitement, sous la forme d'une partie du texte, le contexte de chaque indice est un élément qui distingue l'exploration contextuelle de l'utilisation de patrons morpho-syntaxiques ou de meta-schémas, comme nous allons l'illustrer ci-après.

Voici un exemple de patron morpho-syntaxique emprunté au système FASTUS[27] :

<OrganisationCriminelle> a attaqué <CiblePhysique> de <CibleHumaine> à <Lieu> >Date> au moyen de <Arme> où les termes placés entre les signes <, > renvoient à des ensembles lexicaux.

Dans ce type d'expression, la seule possibilité pour exprimer un contexte réside dans la séquentialité des éléments qui composent ce patron. Pour exprimer une condition structurelle, comme par exemple le fait que ce patron doive apparaître en début de paragraphe, cela nécessite de placer la balise de début de paragraphe « <p> » dans le patron comme suit[28],

<p><OrganisationCriminelle> a attaqué <CiblePhysique> de <CibleHumaine> à <Lieu> >Date> au moyen de <Arme>.

Une autre possibilité consiste à utiliser des schémas (l'expression de schéma est emprunté a [PER 00]). Ce type de procédé a été utiliser pour reconnaître des définitions (cf. tableau 2). La première colonne du tableau indique le nom du schéma et le nom d'une opération de transformation que nous ne commenterons pas ici. Les autres colonnes précisent les noms des classes distributionnelles qui forment le schéma [PER 00] :

Nc : un nom classifieur (hyperonyme)

V= : soit la copule, soit un verbe appartenant à une classe restreinte

Vp : {permettre, servir à, avoir pour effet, être utlisé pour, ..}

Dans cette présentation, le symbole « § » indique que le schéma doit apparaître en début de paragraphe, mais en l'absence d'un langage qui puisse exprimer plus formellement la notion de contexte, l'auteur a placé ce symbole dans la même colonne que la classe « Nc », alors même qu'il exprime le contexte de toute l'expression.

[27] Cité par [POI 99]

[28] C'est aussi la solution adoptée dans le système INTEXT [SIL 93].

	Nc1	Nn	V=	Nc2	Vp	SV
SE 0 réduction	§ La commande	**Distance**	est	un analyseur lexico-statistique.	Elle permet de	comparer statisti-quement les lexiques de deux sous-textes quelconques d'un corpus
SR1 réduction N*c1*		§ Le **filtre**	est	un patron de fouille	qui permet de	définir ....
SR1' réduction V= N*c2*	§ L'analyseur	**COMPARAISON**			permet de	marquer ...
SR2 réduction N*c1*, V= N*c2*		§ **CARACTERISER**			permet de	préciser le fonctionn-ement du journal
SR1'' réduction V= N*c2* V*p*	§ L'analyseur	**SEGMENTATION**				découpe ...
SR2' réduc. N*c1*,V= N*c2*, V*p*		§ **APPLIQUER**				lance ...

Tableau 2. Les schémas de définition [PER 00]

Le travail de conceptualisation sur le langage d'expression des règles d'exploration contextuelle a justement visé, en se fondant sur les concepts initiaux, à proposer un langage qui permette au linguiste d'exprimer les règles d'exploration contextuelle indépendamment de toute implémentation informatique. Ce langage a connu différents développements dans le cadre de thèses actuellement en cours dans l'équipe ; mais dès l'origine de nos travaux de recherche, nous avons essayé d'élaborer un langage dans lequel on puisse manipuler les objets de la linguistique textuelle, tels que les titres, paragraphes, phrases, énumérations, mais aussi, citations, cadres de discours, enchaînements argumentatifs, etc. Il reste évidement beaucoup de travail à faire, certaines notions de linguistique textuelle sont encore difficilement automatisables ; néanmoins les résultats obtenus nous ont permis de mieux cerner, et de mieux répondre, aux besoins de la linguistique textuelle.

2.2. Langages d'expression des règles d'exploration contextuelle : de Ljava à Ltext

2.2.1. Langage de description des indicateurs et des indices

Avant de décrire les différents langages qui ont été élaborés pour exprimer les règles d'exploration contextuelle, il est nécessaire de présenter comment sont décrits indicateurs et indices. Un premier type de description a été conçu par S. Ben Hazez [BEN 99, 02] dans le cadre de sa thèse. Le formalisme proposé, élaboré à partir de l'étude critique des réalisations précédentes [JOU 93, BER

95b], s'émancipe de tout langage informatique tout en permettant une implémentation facilitant la gestion de ces données linguistiques. Il permet au linguiste de construire sa base de données linguistiques en décrivant : les tâches, les indicateurs ou les indices pertinents et le nom des règles d'exploration contextuelles associées. Nous allons illustrer son utilisation par quelques exemples.

Le linguiste déclare tout d'abord les formes lexicales significatives (cf. tableau 3) indices ou indicateurs confondus, qu'il organise en classes[29] non nécessairement disjointes : les formes lexicales sont donc déclarées en extension. Cette déclaration implique l'existence d'un logiciel [BEN 99, 00] de découpage du texte en unités lexicales d'une part, mais aussi la reconnaissance d'unités composés comme les dates, les nombres, les entités nommées, etc.

Bien que des outils d'aide à l'acquisition permettent de produire automatiquement toutes les formes fléchies ou dérivées, le linguiste doit souvent ne retenir que certaines de ces formes fléchies, car pour une tâche donnée, seules certaines flexions d'un verbe sont significatives. Ainsi, si le but est de rechercher les annonces thématiques d'un article scientifique, le verbe *présenter* est significatif seulement lorsqu'il est employé à l'indicatif présent ou au futur, à la première personne du singulier ou du pluriel. Il est aussi possible de déclarer des combinaisons de classes (dernière ligne du tableau). Ces combinaisons permettent de déclarer des lexies (ou locutions autonomes) ; par exemple, la déclaration de *il + &être1 + &importance* permet au linguiste de déclarer des lexies du type *il est primordial ; il est particulièrement important ; il est,…, essentiel ; etc.*

Forme	Nom de Classe
essentiel	&importance
qui suivent	&partie_document1
chapitre	&partie_document2
lignes	&partie_document3
présente	&verbe_présentatif
présentons	&verbe_présentatif
présenterai	&verbe_présentatif
présenterons	&verbe_présentatif
Il + &être1 + &importance	&soulignement

Tableau 3. Déclaration des formes lexicales [BEN 99]

Dans un deuxième temps (cf. . tableau 4), ces formes sont déclarées comme des indices ou des indicateurs et associées à une tâche d'identification Rappelons qu'un indicateur linguistique est un marqueur linguistique d'une valeur sémantique jugée pertinente pour une tâche donnée et qu'un indice permet de résoudre, en contexte, l'éventuelle polysémie de l'indicateur. Une tâche a pour finalité de regrouper des règles d'exploration contextuelle et correspond généralement à un processus d'étiquetage sémantique d'un segment textuel précisé. Il reste donc à déclarer le nom des règles qui doivent être déclenchées pour cette tâche ce qu'illustre le tableau 5. La première ligne déclare une

[29] Chaque classe est identifiée par un nom précédé du caractère & .

règle de nom *RCenthe1001* de la tâche *résumé*, déclenchée par une occurrence, dans une phrase du texte à analyser, d'un indicateur de la classe *&partie_document2*. Cette règle attribue l'étiquette *Thematique_2* à la phrase considérée.

Nom de Classe	Type	Nom de Tâche
&partie_document1	indice	résumé
&verbe_présentatif	indicateur	résumé
&partie_document2	indicateur	résumé

Tableau 4. Déclaration des indices et indicateurs [BEN 99]

D'une manière plus générale, une règle peut déclencher différents types de décision. Des outils d'aide à la gestion de la cohérence et à l'intégration des connaissances issus d'autres travaux linguistiques permettent de répondre à l'objectif d'une acquisition incrémentale et capitalisable des connaissances. Le linguiste peut ainsi connaître quelles sont les règles qui sont déclenchées par un indicateur donné, quelles sont les étiquettes attribuées par un ensemble d'indicateurs, etc.

Nom de la Règle	Etiquette attribuée	Segment Textuel	Nom de Tâche	Nom de Classe
RCenthe1001	Thematique_2	phrase	résumé	&partie_document2
RCenthe112	Thematique_2	phrase	résumé	&verbe_présentatif

Tableau 5. Déclaration des noms règles [BEN 99]

Ce formalisme a été totalement spécifié dans un modèle objet et implémenté en JAVA. Il s'appuie sur un système de gestion de base de données relationnelles [BEN 99, 01] pour la gestion de la persistance des données. Les règles doivent ensuite écrites dans le langage Ljava que nous présentons ci-dessous

2.2.2. Langage Ljava

Le langage Ljava a représenté la première tentative pour exprimer plus formellement le langage d'expression des règles d'exploration contextuelle. Il est le fruit, sous ma direction, des travaux de S. Ben Hazez et de G. Crispino dans le cadre de leur thèse [CRI 99a, CRI 99c, BEN 02, CRI 02, MIN 01], à partir d'une première ébauche que j'avais proposé à partir des travaux réalisés dans le cadre du projet SERAPHIN [BER 95b].

Les règles d'exploration contextuelle y sont exprimées dans un langage formel de type déclaratif. Ce langage est centré sur la notion d'espace de recherche, c'est-à-dire un segment textuel déterminé à partir de l'indicateur, espace dans lequel les indices complémentaires doivent être recherchés. L'intérêt de cette notion, comme nous l'avons souligné précédemment, est qu'elle permet au linguiste de construire simplement un espace sans que celui-ci soit nécessairement formé de phrases contiguës dans le texte. Il faut souligner que les règles d'exploration contextuelle sont exprimées par le linguiste

au vu des observables que constituent les textes. Pour une même tâche, les règles doivent être indépendantes les unes des autres, contrainte qui n'a pas soulevé de difficultés jusqu'à présent.

Chaque règle comprend une partie *Entête*, une partie *Déclaration d'un Espace de Recherche E*, une partie *Condition* et une partie *Action* qui n'est exécutée que si la partie *Condition* est vérifiée :

- La partie *Entête* permet d'indiquer le nom à la règle mais surtout d'assigner la classe de l'indicateur qui déclenchera cette règle. Cette classe est unique. Des commentaires peuvent être placés dans cet *Entête* en vue de documenter la règle.

- La partie *Déclaration d'un Espace de Recherche E* permet de construire un segment textuel, le contexte de recherche, en appliquant différentes opérations sur la structure du texte construite par le moteur d'exploration contextuelle. Il est possible de construire plusieurs espaces de recherche dans une même déclaration, à l'aide des opérations suivantes :
    - **Voisinage (p)** : spécifie un espace de recherche qui commence à partir du p-ième token précédant l'indicateur jusqu'au p-ième token suivant ce dernier.
    - **Voisinage (p1, p2)** : spécifie un espace de recherche qui commence à partir du p1-ème token précédant l'indicateur jusqu'au p2-ième token suivant ce dernier.
    - **Gauche (p) :** spécifie un espace de recherche qui commence à partir du p-ème token précédant l'indicateur jusqu'au token qui précède immédiatement ce dernier.
    - **Droite (p) :** spécifie un espace de recherche qui commence à partir du token qui suit immédiatement l'indicateur jusqu'au p-ième token suivant ce dernier.
    - **Antérieures ()** : spécifie un espace de recherche qui commence à partir du début de la phrase contenant l'indicateur jusqu'au token qui précède immédiatement ce dernier.
    - **Suivantes ()** : spécifie un espace de recherche qui commence à partir du token qui suit immédiatement l'indicateur jusqu'à la fin de la phrase contenant ce dernier.
    - **Phrase parent()** : représente la phrase qui contient l'indicateur.

Dans le langage Ljava, le contexte de recherche est toujours exprimé par rapport à l'indicateur et sa portée est au maximum celui de la phrase dans laquelle cet indicateur est présent. En conséquence, des prédicats prédéfinis permettent d'exprimer des conditions sur les agencements entre les indices ainsi que sur la position des éléments structuraux du texte (titre, section, paragraphe, phrase). Ces prédicats sont utilisés dans la partie *Condition* de la règle. En vue d'offrir au concepteur de la règle une vue complète des éléments qui doivent être pris en compte dans la règle, les classes d'indices, qui vont être recherchées dans les contextes de l'indicateur, sont aussi déclarées dans cette partie de la règle.

- La partie *Condition* explicite les conditions que doivent vérifier les indicateurs et les indices complémentaires. Le langage propose les prédicats suivants qui peuvent être précédés d'un opérateur de négation :
    - **Il_existe_un_indice *x* appartenant_à *C* tel_que :** vérifie l'existence d'un indice *x* de la classe *C* dans le contexte *E* pré-déclaré (dans la partie *Déclaration*). L'opérateur *OU* peut-être utilisé pour exprimer une disjonction entre les noms de classes.
    - **Précède(x, y, p):** ce prédicat permet d'exprimer une condition d'ordre entre les indices. Il indique que l'indice *x* doit précédé l'indice *y* de p tokens. Ces deux indices *x* et *y* doivent nécessairement avoir été préalablement déclarés dans un prédicat du type *Il_existe_un_indice.*
    - **Position(s):** ce prédicat permet d'exprimer une condition sur la position *p* d'un des éléments du modèle du texte (sections, paragraphes, phrases , titres) qui contient l'indicateur déclenchant la règle. L'expression *s* est une expression symbolique où la position *p* est comparée aux éléments structuraux du texte. Ceci permet d'exprimer une condition du type, « la phrase où apparaît l'indicateur appartient au premier paragraphe de la dernière section ».

- La partie *Action* indique le type d'actions réalisées par la règle. Actuellement, deux actions possibles sont : **Attribuer(e)** à un segment textuel ou **Déclencher(R)**. L'action **Attribuer(e)** attribue l'étiquette de nom *e* à la phrase qui contient l'indicateur qui a déclenché cette règle. Le nom de l'étiquette n'est pas spécifié dans la règle, puisqu'il est déclaré dans le système de gestion des connaissances linguistiques. L'action **Déclencher(R)** déclenche l'exécution de la règle R.

Nous allons illustrer les possibilités offertes par ce langage sur différents exemples extraits des travaux de recherche des membres de l'équipe LaLIC.

La figure 9 présente un exemple de règle utilisée dans la tâche de résumé automatique [MIN 01]. La règle de nom *Rthematique* est attribuée à la tâche *Thématique* et est déclenchée si un indicateur appartenant à la classe *&verbe_presentatif*, qui regroupe des formes lexicales des verbes présentatifs, est présent dans le contexte *E1.* Ce contexte s'étend, sur dix tokens, de part et d'autre de l'indicateur. Deux classes d'indice *&partie_document3* et *&partie_document1* sont déclarées. Une première *condition* exprime qu'une occurrence d'un indice de la classe *&partie_document3,* qui regroupe des locutions comme « dans les lignes, etc. », doit être présente dans le contexte spécifié. Une seconde *condition* exprime qu'une occurrence d'un indice de la classe *&partie_document1,* qui regroupe des locutions comme « *qui suivent, suivantes, ci-près, etc.* », doit aussi être présente dans ce contexte. Le prédicat *Precede* précise une contrainte sur l'agencement entre ces deux indices. La partie *Action*

indique qu'une étiquette, dans le cas présent « *Annonce Thematique* », est attribuée à la phrase en question.

Nom de la règle : Rthematique **;**
Tâche déclenchante : Thématique ;
Commentaire : capte une occurrence du type : *Dans les lignes qui suivent … nous présentons …*
Classe de l'Indicateur : &verbe_presentatif ;
E1 := Créer_espace(voisinage,10)
C1:= &partie_document3
C2:= &partie_document1
Condition : Il_existe_un_indice x appartenant_à E1 tel_que classe_de x appartient_a (C1) ;
Condition : Il_existe_un_indice y appartenant_à E1 tel_que classe_de y appartient_a (C2) ;
Precede (x,y, 5) ;
Actions :
Attribuer(Etiquette)

Figure 9. Un exemple de règle écrite dans le langage Ljava [MIN 01]

Un autre exemple de règle (cf. fig. 10), extrait de [LEP 00], illustre la possibilité de créer différents espaces de recherche pour différents indices, les espaces *E1* et *E2* et d'exprimer des conditions de précédences entre les différents indices. De plus, F. Le Priol a étendu les possibilités du langage Ljava en construisant des règles génériques de repérage d'arguments d'une relation sémantique (statique) identifiée.

Nom de la règle : Rapp01 **;**
Tâche déclenchante : RelationStatique
Commentaire : capte une occurrence du type *: A représente les éléments de B*
Classe de l'Indicateur : &Verbeinclusion ;
E1 := Créer_espace(Anterieures)
E2 := Créer_espace(Suivantes)
C1:= &app1
C2:=&LlingPrep2
C3:=&ingcl
Condition : Il_existe_un_indice x appartenant_à E1 tel_que classe_de x appartient_a (C1) ;
Condition : Il_existe_un_indice y appartenant_à E2 tel_que classe_de y appartient_a (C2) ;
Condition : Il_existe_un_indice z appartenant_à E2 tel_que classe_de y appartient_a (C3) ;
Precede (x,y) ;
Precede (y,z) ;
Actions :
Attribuer(Etiquette) ;
RechercherAgument(C1, C2, x1→x2)

Figure 10. Un exemple de règle écrite dans le langage Ljava [LEP 00]

Ces deux exemples démontrent la puissance d'expression du langage Ljava qui a été utilisé pour modéliser différentes tâches d'exploration contextuelle. Il présente néanmoins un inconvénient en terme d'ingénierie linguistique. En effet, ce langage, pour être exploitable par le logiciel ContextO doit être traduit dans le langage de programmation Java. Une règle de Ljava devient ainsi une méthode d'une classe Java. Cette méthode est elle-même compilée par la machine virtuelle et intégrée dans le

code de la plate-forme ContextO. Un compilateur du langage Ljava a été réalisé dans le cadre du mémoire de Amar Khetar [KHE 00] rendant ainsi ce cycle de compilations successives totalement automatique. Néanmoins, comme tout processus de compilation il est sensible à des erreurs lexicales ou syntaxiques, ce qui rend son utilisation délicate pour un non spécialiste. Ce constat a amené S. Ben Hazez, dans le cadre de sa thèse [BEN 00, 02] à proposer le langage Ltext dont l'un des avantages, entre autre, est de supprimer ce cycle de transformation.

2.2.3. Langage Ltext

En élaborant le langage Ltext, S. Ben Hazez poursuit deux objectifs. Tout d'abord, proposer un langage de requête composé d'un ensemble d'opérations. Ce langage se doit d'offrir une puissance d'expression au moins égale à celle du langage Ljava, tout en conservant les notions clefs de l'exploration contextuelle, c'est à dire les notions d'indicateurs, d'indices et de contexte de recherche. De plus, en vue de simplifier le travail du linguiste qui écrit les règles, les éléments structuraux du texte font partie intégrante du vocabulaire de base du langage. En cela, il se distingue des différents travaux de recherche qui proposent des langages de requête sur un texte [BUR 92, CLA 95]. L'autre objectif vise à supprimer le cycle de compilation, en intégrant ce langage de requête dans la déclaration des indicateurs et des indices.

Dans Ltext, un document D est structuré comme suit :

D = {S1 S2 S3….Sn} : vecteur de segments de traitement. Un segment peut désigner : la phrase Ph, le paragraphe P, la section H, le texte T, etc.

Un segment S est représenté comme suit :

S ={U1 U2 U3…Ui Ui+1 ….Un} : vecteur d'unités atomiques (mots).

Les classes

Les classes définies par un utilisateur doivent être préfixées par le caractère « & ». Une classe représente une disjonction d'expressions qu'on peut utiliser dans la description des motifs linguistiques. Par exemple, le motif :

« it would be + &importance »

permet de retrouver des séquences éventuellement discontinues : « *it would be possible* », « *it would be very important* », etc. La classe « &*importance* » utilisée dans le motif permet de localiser les adverbes : « *possible*, *important, évident,…* ».

Les catégories

Le langage Ltext offre la possibilité d'utiliser des catégories prédéfinies. Elles sont reconnues par le symbole « # » placé avant le code de la catégorie. Le tableau 6 présente quelques uns des symboles prédéfinis.

CODE	SIGNIFICATION	EXEMPLES
#VIDE	SYMBOLES NULS	
#TOK	FORMES SIMPLES	
#BD	BALISE OUVRANTE	<s>
#BF	BALISE FERMANTE	</s>
#PONCT	PONCTUATION	: ! , ? - ;
#DATE	DATE	12 Dec 2000, 12/12/00
#NP	ENTITÉS NOMMÉES	Académie des sciences de Bulgarie, J.-P.Desclés
#NB	NOMBRES	1000, quatre, IV
#ABREV	ABREVIATION	Fig., c.a.d,
….	….	…..

Tableau 6 . Les catégories prédéfinies de Ltext [BEN 00].

Dans Ltext, un contexte ou espace de recherche est un bloc textuel déterminé par une borne inférieure « *borneinf* » et une borne supérieure « *bornesup* ». Etant donnée une occurrence d'un indicateur exprimé par le motif E, on peut définir plusieurs types de contextes en spécifiant les valeurs des deux bornes à l'aide des variables suivantes (ces variables sont des éléments d'indexation présentés dans la figure 1 de la section 2) :

E : désigne la proportion du texte, occurrence de l'indicateur (ex. *If* + *would* + *#VB*).

n=1,2,…k désigne le rang d'une forme simple de *E.* Par exemple, dans le motif *(If* + *would*+ #VB), « *if* » possède le rang 1, « *Would* » possède le rang 2 et « *#VB* » possède le rang 3,

D'une manière générale, l'expression *X-n* ou *X+n* désigne respectivement le n ème segment (*X* désigne la phrase *S*, un paragraphe *P*, une section *H*, etc.) qui précède ou qui suit le segment courant qui contient une occurrence du motif E. Un vocabulaire de termes prédéfinis permet de désigner les différents éléments d'un document D :

- *HeadDoc, EndDoc* désignent respectivement le premier et dernier segment (phrase) de *D* ;
- *P* désigne le paragraphe courant qui contient l'occurrence de E.
- *H* désigne la section courante qui contient l'occurrence de E.

- *S* désigne la phrase courante qui contient l'occurrence de E.
- *HeadSeg, EndSeg* désignent respectivement la position du premier et du dernier élément de *S* ;
- *E-n* désigne une position dans *S* à gauche de *E* en comptant *n* formes simples ;
- *E+n* désigne une position dans *S* à droite de *E* en comptant *n* formes simples ;
- *S-n* désigne la phrase de rang *n* qui précède la phrase courante *S* ;
- *S+n* désigne la phrase de rang *n* qui suit la phrase courante *S ;*

Les opérateurs

Les contraintes sur un indicateur sont exprimées à l'aide d'opérations élémentaires de distance (distmax, distmin, dist) et de référence aux contextes (contains, startswith, endswith, etc.) pour chercher des indices complémentaires, des indices de positions, des indices de fréquences, etc. Dans ce qui suit, nous présentons des exemples d'opérateurs pour exprimer des contraintes élémentaires:

- Contrainte à *gauche de l'indicateur* : l'indicateur « *if + would + #VB* » permet de chercher la séquence « *if* » , suivi de « *would* », suivi d'un verbe à l'infinitif. Supposons que nous ne voulons pas avoir un élément de la classe &26 avant le mot « *if* » dans la phrase contenant l'indicateur. Cette contrainte peut être exprimée de la façon suivante :

not(contains(&26, headseg, E-1)).

où la borne inférieure headseg désigne le premier mot de la phrase et la borne supérieure E-1 désigne le mot qui précède immédiatement l'indicateur. Contains est un opérateur de référence à des contextes qui sera expliqué dans la section suivante.

- Contrainte à *droite de l'indicateur* : reprenons l'exemple précédent et supposons que la contrainte est de ne pas avoir un élément de la classe &25 après le verbe à l'infinitif. Cette contrainte peut être exprimée de la façon suivante :

not(contains(&25, E+1, endseg)).

où la borne inférieure E+1 désigne le mot qui suit immédiatement l'indicateur et la borne supérieure endseg désigne le dernier mot de la phrase.

- Contrainte à *l'intérieur de l'indicateur* : reprenons l'exemple précédent et supposons que la contrainte est de ne pas avoir un élément de la classe &24 entre les deux mots « *if*» et « *would ».* Cette contrainte peut être exprimée de la façon suivante :

not(contains(&25,1,2)).

Où la borne inférieure de valeur 1 désigne le mot *if* et la borne supérieure de valeur 2 désigne le mot « *would ».*

Plus généralement, les contextes (S-1,S-1) et (S+1, S+1) désignent respectivement la phrase qui précède et la phrase qui suit la phrase contenant l'indicateur.

Les contextes (P-1,P-1) et (P+1,P+1) désignent respectivement la phrase qui précède et la phrase qui suit la phrase contenant l'indicateur.

Les contextes (H-1,H-1) et (H+1,H+1) désignent respectivement la phrase qui précède et la phrase qui suit la phrase contenant l'indicateur.

Le langage Ltext a été utilisé pour modéliser les connaissances linguistiques spécifiées par S. Baldo dans sa thèse [BAL 00] dont le thème est la traduction en français de phrases contenant le modal *would.* Le travail de formalisation des règles dans le langage Ltext est le fruit du travail de W. Elleuche Yaiche dans le cadre de son mémoire de DEA [ELL 00]. La figure 9 présente d'abord la règle telle qu'était exprimée dans [BAL 00], il s'agit d'une règle qui identifie la valeur « valeur d'hypothétique présent ». La figure 11 montre ensuite comment cette règle a été formalisé avec le langage Ltext .

RÈGLE 25 (translation):
SI on relève une occurrence de *WOULD* (classe 44.2.a)
ET SI elle est suivie d'un *V à l'infinitif* (classe 43.2)
ET SI elle est précédée d'une *conjonction à valeur appréciative* (classe 17)
ET SI un autre verbe est conjugué au *present* (classe 40), au *present perfect* (classe 41) ou au *passé* (classe 37.2, 38.2 ou 39) dans le même paragraphe
ALORS il s'agit de la valeur d'hypothétique présent
DONC traduire *WOULD + V* par le verbe correspondant en français, au conditionnel présent.
À MOINS QU'il n'y ait un *indice qui fractionne temporellement le procès* (classe 9.2).

Figure 11. Un exemple de règle de traduction de *Would* [BAL 00].

La modélisation de cette règle avec le langage Ltext est la suivante :

Classe de l'indicateur : &translation25
Expression de l'indicateur : &17 + would #VB
Contrainte : not(contains(&9.2, S, S)) and contains(#VBP | #VBZ | #VBZ #VBN | #VBP #VBN |#VBN, **P**, **P**)
Action : <txt>would+verbe </txt><vb>verbe_francais</vb><tps2>conditionnel_présent</tps2>.

Figure 12. Un exemple de règle écrite dans le langage Ltext [ELL 00].

Une interface est en cours de développement par S . Ben Hazez afin d'offrir à l'utilisateur du langage Ltext des outils inter-actifs qui le guident dans l'utilisation de ce langage[30]. Comme on peut le voir sur la figure 13, cette interface présente l'ensemble des informations que l'utilisateur manipule

[30] S. Ben Hazez [BEN 02] travaille maintenant au développement du système SEMANTEXT, système autonome de repérage d'informations dans les textes.

pour construire une règle d'exploration contextuelle. Nous reviendrons sur cet aspect, à savoir l'interactivité avec l'utilisateur dans la partie concernant la plate-forme ContextO.

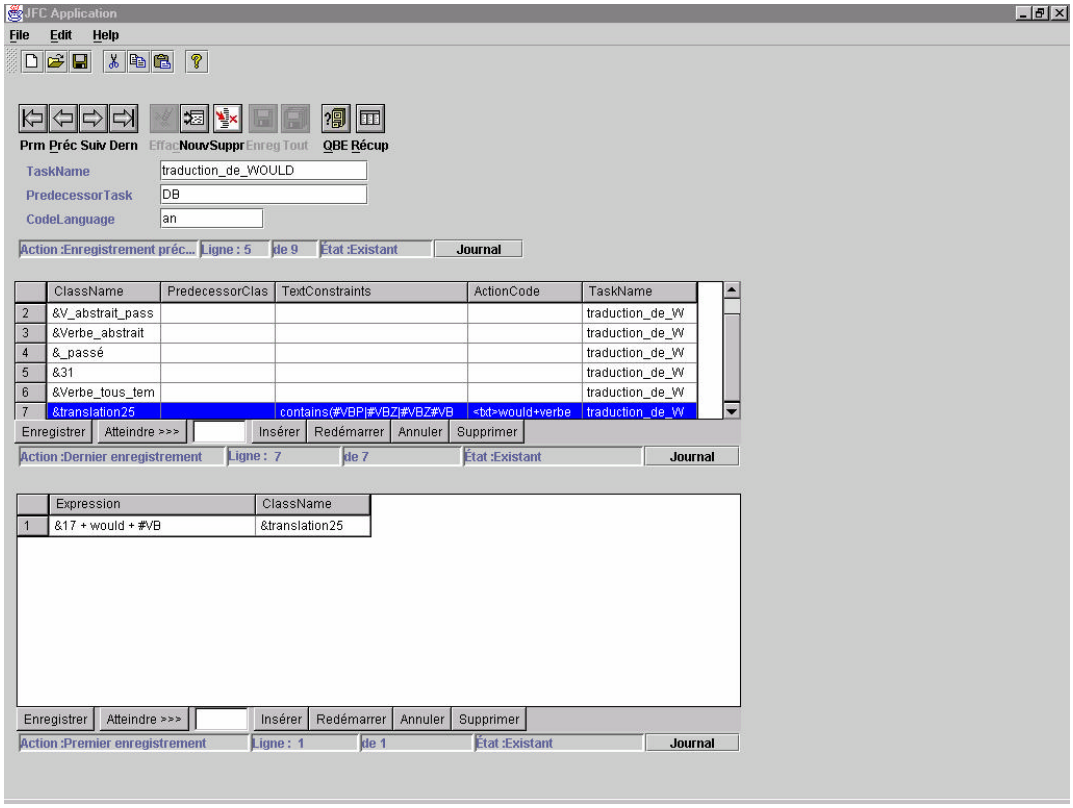


Figure 13 : Interface du langage Ltext [BEN 00]

2.2.4. Un langage plus spécialisé

Les deux langages présentés précédemment sont fondés sur le principe de l'exploration contextuelle qui consiste à identifier des indices dans le contexte d'un indicateur en s'appuyant notamment sur la structure du texte qui reste inchangée.

Dans le cadre de sa thèse, D. Wonsever [WON 01, 02] a développé un système de reconnaissance de propositions dans des textes français. Elle a d'abord cherché à utiliser l'exploration contextuelle puis a constaté que l'identification des propositions ressortait d'une démarche différente. En effet, le processus de reconnaissance d'une proposition doit modifier dynamiquement la structure du texte analysé. De plus, si l'identification de la présence d'une proposition, dans une phrase, reste déclenchée par la présence d'un indicateur (une unité verbale), la recherche du début et de la fin de la proposition s'appuie en partie sur la vérification de l'absence de marques. En conséquence, face à ce nouveau besoin, D. Wonsever a proposé un langage spécialisé, qui reprend certains des concepts de l'exploration contextuelle, tout en proposant de nouvelles fonctionnalités.

Nous présentons brièvement ce langage, en reprenant la présentation qui en est faite dans [WON 01] ; nous comptons, par ailleurs, intégrer ce travail dans la plate-forme ContextO.

Compte-tenu de l'objectif recherché, ce langage n'utilise pas le modèle hiérarchique du texte comme dans les deux langages précédents. En effet, l'identification d'une proposition est indépendant de la place de cette proposition dans le texte, par conséquent D. Wonsever propose de considérer d'abord le modèle T0 d'un texte T :

$T0 : T = w_1 ... w_n$, où n est la longueur du texte T et w_i , $1 \leq i \leq n$, sont les tokens du texte T

Puis, après le balisage et l'étiquetage morpho-syntaxique, ce modèle est transformé pour obtenir le modèle T1, composé d'un ensemble de segments de texte et de l'étiquette de ce segment, s'inspirant ainsi des travaux de [CLA 95]. Chaque segment du texte est représenté par une paire d'indices indiquant respectivement sa première et sa dernière position. Voici un exemple qui illustre cette conception du texte[31] :

T0 :L'homme que j'ai vu hier est ton père.

T1 = {[(0,1),L'], [(1,2), homme], [(2,3), que], [(3,4), j'], [(4,5), ai], [(5,6), vu], [(6,7), hier], [(7,8),est], [(8,9),ton], [(9,10),père], [(10,11),.], [(0,1),det], [(1,2), nom], [(2,3), Pronrel], [(3,4), Pronpers], [(4,5),Aux], [(5,6), participe_passé], [(6,7),adverbe], [(7,8), Verbe], [(8,9),poss], [(9,10),nom], [(10,11),punct], [(4,6), finUV], [(7,8), finUV] }

Le langage des règles contextuelles est alors défini pour pouvoir s'appliquer sur ce modèle.

Description des règles contextuelles

Une règle contextuelle est une expression dont l'objectif est d'identifier et d'étiqueter une portion du modèle du texte défini précédemment. Une portion du texte est étiquetée si elle satisfait une condition qui est une fonction de cette même portion du texte et, éventuellement, d'une portion de texte qui la précède (contexte gauche) ou qui lui succède (contexte droit). Une portion de texte satisfait la condition d'étiquetage si elle contient certaines marques (mots, ponctuation, étiquette) dans un ordre spécifique. Le contenu de la règle détermine quels sont les éléments qui doivent être présents et dans quel ordre. Ces éléments ne sont pas nécessairement contigus, d'autres portions de texte peuvent s'intercaler entre eux ; dans ce cas, il est possible de préciser une liste d'éléments qui ne doivent pas être présents dans ces portions de texte.

Syntaxe des règles contextuelles

Soit V un ensemble fini d'étiquettes, une règle contextuelle est alors définie comme une expression de la forme :

Etiquette → ContexteGauche \ Corps / ContexteDroit ; Spécifications

[31] Les étiquettes morpho-syntaxiques utilisées sont suffisamment explicites pour que nous ne les commentions pas, à l'exception cependant de *finUV* qui signifie une unité verbale finie.

où:

- *Etiquette* ∈ V (ensemble fini d'étiquettes).
- *ContexteGauche*, *Corps* et *ContexteDroit* sont des chaînes composées de deux types d'éléments : des étiquettes appartenant à V et aux zones d'exclusion.
- *ContexteGauche* et *ContexteDroit* peuvent être vides, mais le *Corps* ne peut pas l'être.
- La chaîne *ContexteGauche*. *Corps*.*ContexteDroit* (où le symbole '.' signifie concaténation) sera référée comme la partie *condition* de la règle. Dans cette chaîne les étiquettes peuvent entourer n'importe quelle zone d'exclusion.
- Une zone d'exclusion est une expression de la forme **(EnsemblesExclus, Taille),* où *EnsemblesExclus* est le nom d'un ensemble d'étiquettes et *Taille* un nombre entier qui spécifie l'étendue de la zone d'exclusion.
- *Spécifications* est la définition par énumération des ensembles spécifiés dans la zone d'exclusion de la règle. Un ensemble dans *Spécifications* peut être vide. Néanmoins, les deux étiquettes qui entourent la zone d'exclusion sont considérées comme lui appartenant. Si aucune zone d'exclusion n'est spécifiée, la partie *Spécifications* est absente de la règle.

Enfin pour des raisons de consistance que nous n'expliquerons pas ici[32], une contrainte additionnelle est ajoutée : l'étiquette (*Etiquette*) de la règle ne peut pas appartenir à une zone d'exclusion.

Voici un exemple de règle contextuelle (CR) :

CR : *Proprel → \Pronrel *(S,5) UVfin *(S,10) / UVfin; S={Pronrel, iniProp, UVfin}*

où:

- V = {Proprel, Pronrel, UVfin, iniProp} ;
- Le *ContexteGauche* est vide, le *ContexteDroit* est défini comme la chaîne *UVfin* et le *Corps* comme la chaîne *Pronrel*(S,5) UVfin *(S,10)* ;
- La règle CR a deux zones d'exclusions : **(S,5)* et **(S,10)* avec le même *EnsemblesExclus* (*S*) qui est défini dans la partie terminale de la règle comme *S={Pronrel, iniProp, UVfin}.*

La règle CR peut être paraphrasée ainsi : « rechercher autour d'unité verbale, sur une étendue de cinq tokens, une étiquette de pronom relatif (*Pronrel*) en excluant entre ce pronom et l'unité verbale, toute étiquette du type, pronom relatif (*Pronrel*), ou début de proposition (*iniProp),* ou encore toute autre unité verbale (*Uvfin*) »; ceci permet d'identifier le début de la proposition. La fin de la proposition est identifiée de la même manière mais la recherche est effectuée sur une étendue de dix tokens.

Une description complète des règles en utilisant la notation EBNF est la suivante :

```
ContextualRule ::= label '→' Rhs (';' SetSpecs)?.
Rhs ::= '\' R '/' | R Oi R '/' | '\' R Od R | R Oi R Od R.
R ::= label ( ExZ label | label)*.
Oi ::= '\' (ExZ)? | (ExZ)? '\'.
Od ::= (ExZ)? '/' | '/' (ExZ)?.
SetSpecs ::= (SetSpec)+ .
SetSpec ::= identifier '=' '{' (label (',' label)*)? '}'.
ExZ ::= '*' '(' identifier ',' integer ')'.
```

Figure 14. Grammaire EBNF du langage de règle contextuelle [WON 01]

[32] Le lecteur intéressé trouvera les explications complètes dans [WON 01].

D. Wonsever a d'abord décrit sous forme axiomatique ce système, ce qui lui a permis d'argumenter sur sa consistance et sa complétude, puis un premier prototype a été implémenté. La figure 15 illustre le découpage obtenu sur un texte d'agence de presse.

Antarctique -médecin -USA WELLINGTON -
[prop / Le médecin américain de la station de recherche Amundsen -Scott , au pôle sud [Proprel / , qui se traite elle-même contre un cancer du sein depuis le mois de juillet , /Proprel] va pouvoir être évacué par un avion militaire américain /Proprel/ qui est parvenu à atterrir samedi sur la base [Proprel/ où règne une température de proche de 50 degrés celsius /Proprel] /Proprel] /prop] (Agence France Presse, 16/10/1999)

Figure 15. Exemple de découpage en propositions [WON 01]

2.2.5. En guise de synthèse

La présentation de ces trois langages démontre le travail de recherche auquel a donné lieu la méthode d'exploration contextuelle. Ces langages proposent de formaliser la notion de contexte en l'associant à un modèle du texte, hiérarchique pour Ljava et Ltext, séquentiel pour le dernier présenté. Ce qui nous semble manquer actuellement, c'est la possibilité d'utiliser dans ces langages des structures plus discursives comme des citations, des cadres de discours, des structures argumentatives, etc. Nous pensons en effet que la notion de contexte de recherche ne se limite pas à celle de partie physique du texte.

Or la conceptualisation, en terme informatique, de ces notions ne peut se fonder que sur les résultats de la recherche en linguistique textuelle. Des travaux en ce sens sont ainsi initiés dans l'équipe LaLIC qui visent à unifier différentes recherches effectuées dans le domaine du traitement du temps et de l'aspect. Nous pensons aussi que d'autres travaux de recherche en linguistique textuelle, notamment ceux de M.P. Woodley et de M. Charolles devraient pouvoir nous aider dans ce travail de formalisation.

2.3. Construction d'un système d'exploration contextuelle

L'acquisition des données linguistiques (cf. fig. 16) nécessite une fouille systématique des textes en vue d'accumuler les indicateurs, les indices et d'exprimer les règles qui les combinent ; cette fouille est complétée par un travail de réflexion linguistique afin de dégager les régularités textuelles.

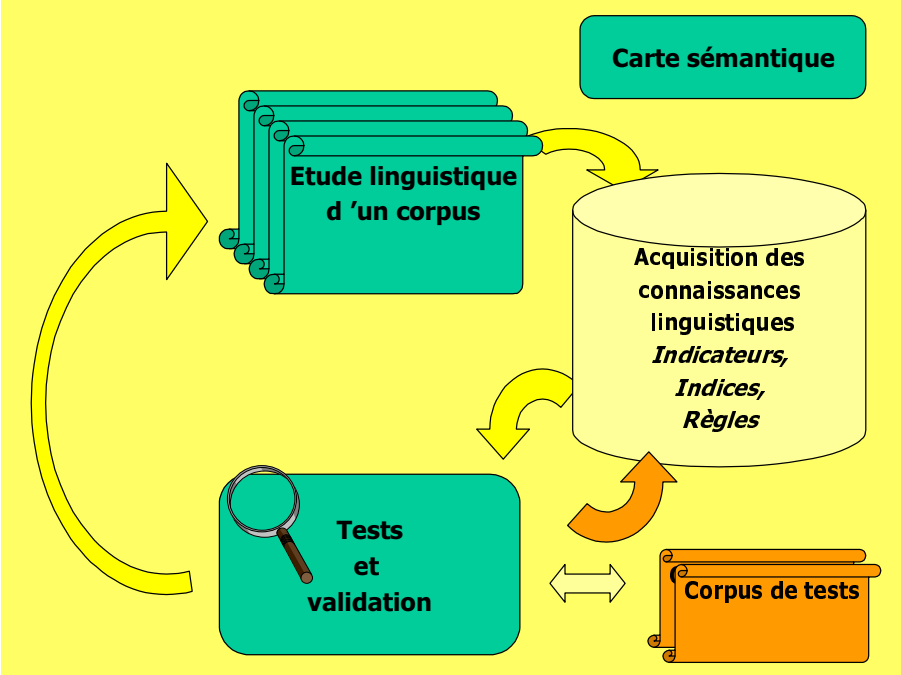


Figure 16. Acquisition des connaissances linguistiques

Le choix du corpus de départ constitue donc une étape importante. De ce point de vue, l'exploration contextuelle s'inscrit dans le courant de la linguistique de corpus ou le travail linguistique se focalise sur les observables plutôt que sur des exemples d'énoncés construits hors contexte. Le choix des textes qui constituent le corpus de référence n'est pas indépendant de la tâche :

- le corpus de travail pour la tâche « Résumé » dans le système SERAPHIN [BER 96a] est constitué essentiellement de rapports techniques et d'articles extraits de revues spécialisées dans le domaine de l'énergie ;
- le corpus de travail pour la tâche « Causalité » de [JAC 98] s'appuie essentiellement sur des articles extraits de la revue « La Recherche » et du « Monde Diplomatique » ; celui de [GAR 98] porte exclusivement sur des comptes rendus de maintenance rédigés par des techniciens de l'EDF ;
- le corpus de travail pour la tâche « Traduction de Would » de [BAL 00] comprend environ une vingtaine d'ouvrages d'auteurs contemporains (Carroll, Dahl, Doyle, Kipling, etc.);
- le corpus de travail pour la tâche « Citation » de [MOU 00] comprend une centaine d'articles de presse spécialisée (Le Monde Diplomatique, 01 Informatique), une trentaine d'articles scientifiques traitant de l'effet de serre et quatre romans policiers ;

- Le corpus de travail pour la tâche « Relations Statiques » dans le système SEEK-JAVA de [LEP 00] comprend des textes descriptifs sur des gravures rupestres. Au sujet de son corpus de travail, F. Le Priol fait remarquer que certains textes peuvent se révéler totalement inadaptés :

> *« ..Le corpus utilisé n'est pas adapté à l'extraction de relations descriptives. En effet, ce corpus ne décrit pas le véhicule mais son utilisation. Les relations sémantiques entre les concepts sont donc essentiellement évolutives (changement, mouvement). De plus, ce manuel n'est pas rédigé, c'est une suite d'alinéas et de liste de commandes. Les marqueurs linguistiques classiques ne se retrouvent pas dans ce type de texte. »* [LEP 00] p. 213.

Cette dépendance vis à vis du corpus de travail dans la phase d'acquisition nécessite la mise en place d'une méthodologie d'acquisition et l'utilisations d'outils logiciels qui permettent de contrôler l'ajout, la suppression ou la modification des marqueurs [BEN 99]. Reste une question cruciale : comment le linguiste peut-il déterminer la complétude des connaissances décrites dans le système d'exploration contextuelle ? Ou encore, quelle est la sensibilité du système aux effets de style, aux particularités d'un domaine, au genre textuel ? La définition de la carte sémantique prend ici toute son importance puisqu'elle guide le linguiste dans la tâche de délimitation du périmètre des classes sémantiques qu'il faut construire. Néanmoins, l'exploration contextuelle, comme toute méthode fondée, en partie, sur l'étude des corpus, n'échappe pas aux difficultés inhérentes à la linguistique de corpus. Nous citerons à l'appui de cette interrogation un extrait de [PER 00].

> « *La disponibilité croissante de bases de textes sur support électronique accompagnées d'interfaces de recherche conduit un nombre également croissant de linguistes de toutes disciplines, et ne se réclamant pas nécessairement des linguistiques de corpus, à travailler sur des exemples extraits de textes plutôt que fabriqués. On utilise donc la base Frantext, ou les grands quotidiens disponibles sur CD-Rom, comme source d'exemples pour des études dans tous les domaines de la linguistique. Beaucoup de ces études, malgré ce changement méthodologique, restent toutefois centrées sur la langue, et traitent le recours au corpus avant tout comme une manière commode de se constituer des données pour aborder la langue. Dans l'univers du TAL, la possibilité de recueillir et de traiter des milliards de mots pousse parfois les chercheurs à accepter sans sourciller la devise "more data is better data". Il me semble important de s'interroger sur le statut de ces données, sur leur adéquation à l'objet d'étude, sur leur représentativité. Les questions sont complexes : le fait que des milliers d'articles du Monde deviennent accessibles grâce à la collection de CD-Roms regroupant plusieurs années de ce quotidien*

> *ne change rien quant à leur aptitude à représenter la diversité des potentialités du français. Ils sont tous issus du même journal. En revanche, si l'on désire envisager Le Monde comme représentant un usage non marqué, une "langue générale", il faudra s'interroger non seulement sur sa représentativité mais aussi sur son homogénéité (les rubriques sportives y parlent-elles la même langue que les éditoriaux ?). Il y a lieu de se demander s'il existe des domaines de la linguistique descriptive pour lesquels l'origine des données n'importe pas. La question se pose a fortiori en TAL, puisque sont visés sur la base d'observations à partir de corpus, des traitements informatisés dont la qualité et l'efficacité va dépendre de l'adéquation de la modélisation initiale aux textes à traiter.* » [PER 00], p. 122-123.

En guise de réponse, M. P. Péry-Woodley propose de modéliser la variation dans les textes en traitant de façon séparée les caractéristiques internes (linguistiques) et les caractéristiques externes (fonctionnelles et situationnelles). Elle suggère d'adopter la méthode de Biber [BIB 88, 89] qui induit les types de textes à partir d'analyses de corpus et distingue ainsi clairement les « types (linguistiques) » et les « genres (situationnels) ». Reste, comme l'indique M-P. Woodley que :

> *« Lorsqu'on travaille sur une langue "minoritaire" et relativement démunie sur le plan des ressources en corpus comme le français, on ne peut cependant bénéficier dans l'immédiat de cette avancée. On est loin en effet de pouvoir procéder à une adaptation de la méthodologie biberienne pour le français, non pas qu'on manque de travaux descriptifs qui pourraient servir de base à la constitution d'une liste de traits pour une typologie émergente, mais parce qu'on ne dispose actuellement d'aucun corpus de référence. »* [PER 00], p. 132.

Nous ne sommes pas persuadé qu'une typologie des textes fondée sur la méthode proposée par D. Biber nous permettrait de répondre à la question de la complétude des connaissances acquises. En effet, l'expérience acquise dans ce domaine dans le cadre des travaux de l'équipe LaLIC nous fait privilégier une méthode que l'on peut qualifier, à l'instar de ce qui se pratique en ingénierie logicielle, de développement en spirale qui organise la tension entre l'étude d'un corpus et le recours à l'introspection linguistique.

Chapitre 3

Plate-forme FilText

3.1. Objectifs

Historiquement, on peut considérer que la notion de plate-forme d'ingénierie linguistique trouve son origine dans les Langages Spécialisés pour la Programmation Linguistique [VAU 85] . Le système INTEXT [SIL 93] qui intègre un langage fondé sur les automates à états finis et d'importantes ressources linguistiques illustre parfaitement, de notre point de vue, l'approche des langages spécialisés. La nécessité d'enrichir cette approche avec les notions de « modèle de texte » et de « représentation des données linguistiques » a fait émerger le concept de plate-forme d'ingénierie linguistique dédiée au traitement textuel. Celle-ci est conçue comme « une boîte à outils » qui vise à fournir à l'utilisateur des moyens d'intégrer ses propres outils et ses propres modèles de représentation dans des outils génériques.

L'atelier KES (« Knowledge Extracting and Structuring »), développé dans le cadre du projet GRAAL par l'Aérospatiale, GSI-Erli et la DER (Direction des Etudes et de la Recherche) de l'EDF, est une des premières réalisations qui ait proposé un modèle conceptuel de représentation des connaissances linguistiques. Elle visait à fournir les outils nécessaires en vue de permettre à l'utilisateur de passer de textes bruts à un ensemble structuré des données. Il était ainsi possible de bâtir des applications KES de constitution de terminologies, ou encore d'aide à l'interprétation sociologique Nous reprenons ci-après la présentation faite par [HER 96].

> *« Une phase initiale de "préparation" des textes [...] en fournit une représentation sous forme de liste de mots et de groupes syntagmatiques. Cette représentation est ensuite gérée sous forme de base de données objet, puis utilisée dans une phase interactive sous forme d'"**ecks**" (élément de*

> *connaissance **K**ES). Des attributs (définis par l'utilisateur) sont attachés aux ecks : catégorie, langue, fréquence ... Si l'usage privilégié de l'eck est de modéliser un terme, on n'y est pas du tout contraint. Un eck peut par exemple représenter un texte ou une classe de termes, il peut aussi être une notion, un domaine, un représentant linguistique ...*
>
> *Une interface permet la visualisation et la mise à jour des données. Elle se compose de matrices appelées "**vues de travail**" (cf. fig. 17), paramétrables par l'utilisateur, qui présentent les ecks dans des colonnes. Les ecks sont éditables, on peut naviguer d'un eck à un autre, ou bien d'un eck vers ses contextes dans les textes analysés par des liens hypertextes.*
>
> *En outre, une bibliothèque de "règles" (écrites dans un langage proche de SQL3) permet d'effectuer des opérations automatiques sur les données: filtrage, comptage, mise à jour, création de liens ou encore création de nouvelles vues de travail, pour présenter autrement les données. Nous exploitons en particulier, dans l'application présentée, la possibilité d'utiliser les colonnes comme des conteneurs de termes. Des opérations courantes de validation ou élimination des termes peuvent ainsi s'exécuter sur des listes entières, définies par l'appartenance à une colonne. Une règle pouvant en appeler une autre, on peut bâtir des scénarios d'utilisation, qui définissent une méthode de travail pour un utilisateur particulier. »* [HER 96].

L'atelier KES a été implémenté en C++ sur station Unix puis testé par des utilisateurs de l'EDF. Des problèmes de performance, notamment lors du traitement de textes de grande taille, et des modifications dans la stratégie de la politique de recherche et développement de la DER de l'EDF ont entraîné l'abandon de cet atelier.

De son côté, Jean Guy Meunier dans [MEU 98] a proposé de définir un modèle conceptuel puis de développer des tâches spécialisées qui coopèrent entre elles :

> *« Un logiciel qui, dans l'accès à l'information textuelle, ne réalise qu'un seul type de tâche devient vite insatisfaisant parce qu'il ne correspond pas à la nature cognitive de ce que font les lecteurs et les analystes de textes. Ceux-ci ont des lectures multiples des textes et ils veulent parcourir un texte dans diverses perspectives. »* [MEU 98] p. 6.

La plate-forme ALADIN, projet de l'équipe de J-G. Meunier, devait implémenter ces concepts, mais diverses difficultés tant techniques que financières n'ont pas permis sa mise en œuvre.

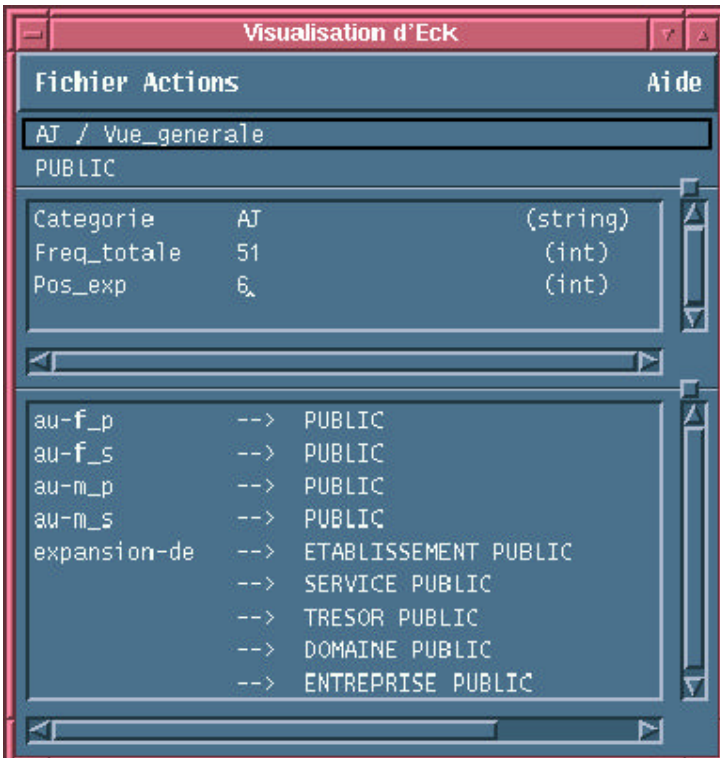
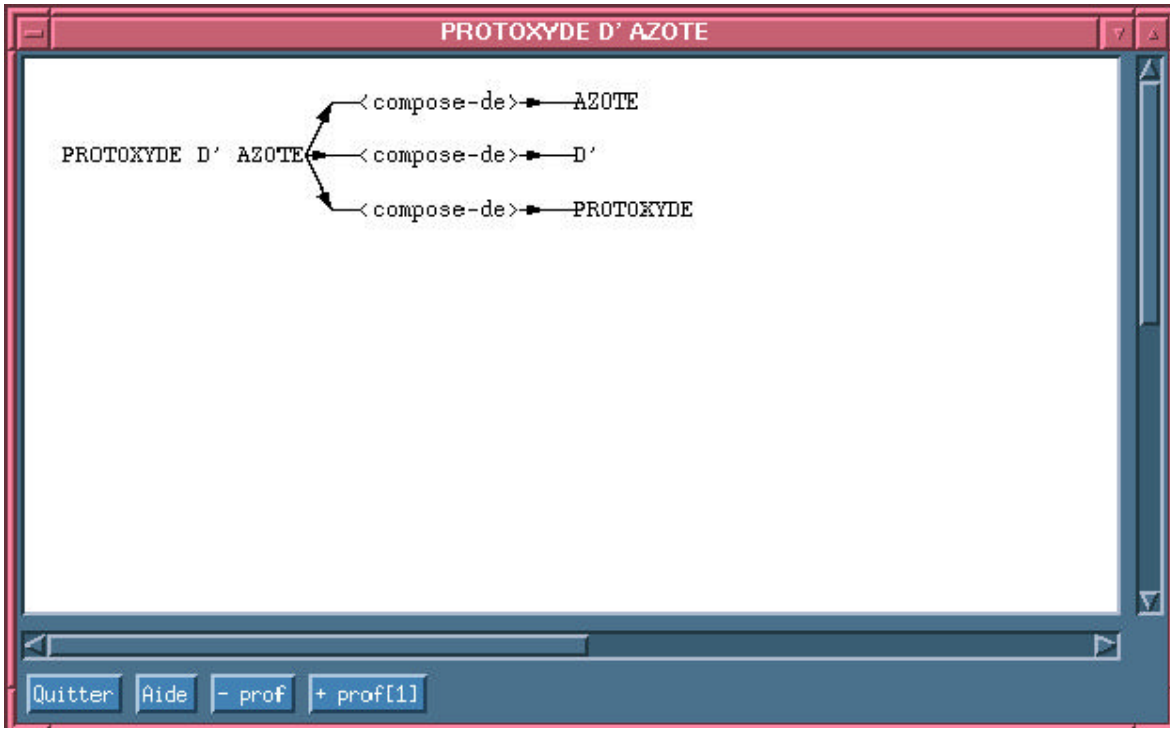


Figure 17. Exemples de vue de travail de la plate-forme KES [HER 96]



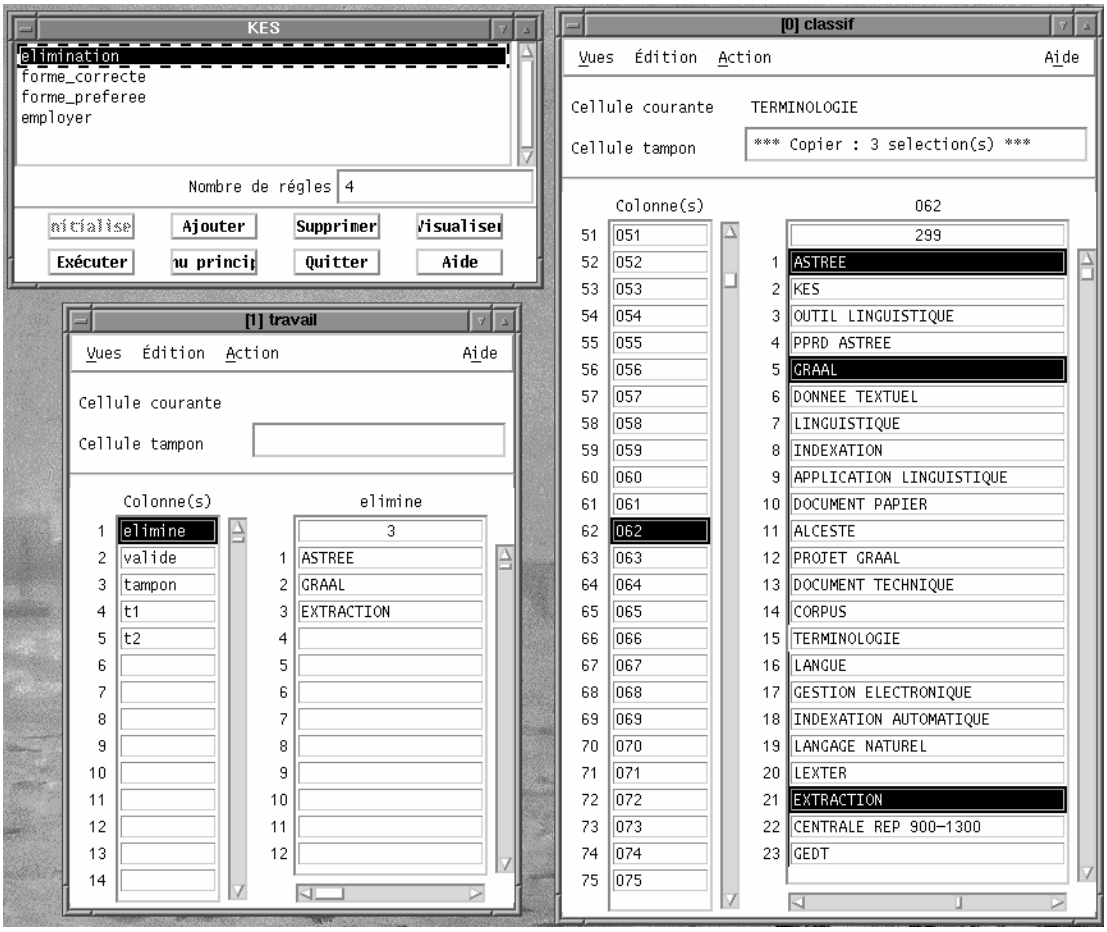


Figure 18. Une interface utilisateur de la plate-forme KES [HER 96]

La plate-forme FilText [CRI 99a, BEN 00, BEN01, BEN 02, CRI 02, MIN 01] reprend en partie ce paradigme, mais elle se veut à la fois plus réaliste dans ses ambitions et plus ouverte techniquement. Plus réaliste, au sens ou le modèle de représentation des connaissances linguistiques est relativement figé puisqu'il s'appuie sur la méthode d'exploration contextuelle [DES 91, 97a, 97b] présentée précédemment ; plus ouverte techniquement, puisqu'elle privilégie la notion de composants dotés d'interfaces logicielles (API).

L'expérience acquise lors du développement des systèmes dédiés précédents [JOU 93, BER 96a, BER 96b, GAR 98, JAC 98] nous a orienté vers une plate-forme qui offre d'une part, des fonctionnalités propres au processus d'acquisition des connaissances linguistiques et des fonctionnalités propres au processus de fouille de textes. La figure 19 illustre les différents cas d'utilisation de la plate-forme Filtext et identifie trois types d'acteurs [33]: le linguiste, l'architecte et l'utilisateur. Comme nous l'avons indiqué précédemment un modèle conceptuel sous-tend un langage

de description de ces données linguistiques. Ce modèle est enrichi par les différentes recherches en cours et il revient au linguiste et à l'architecte informaticien d'implémenter ces modèles dans la plate-forme.

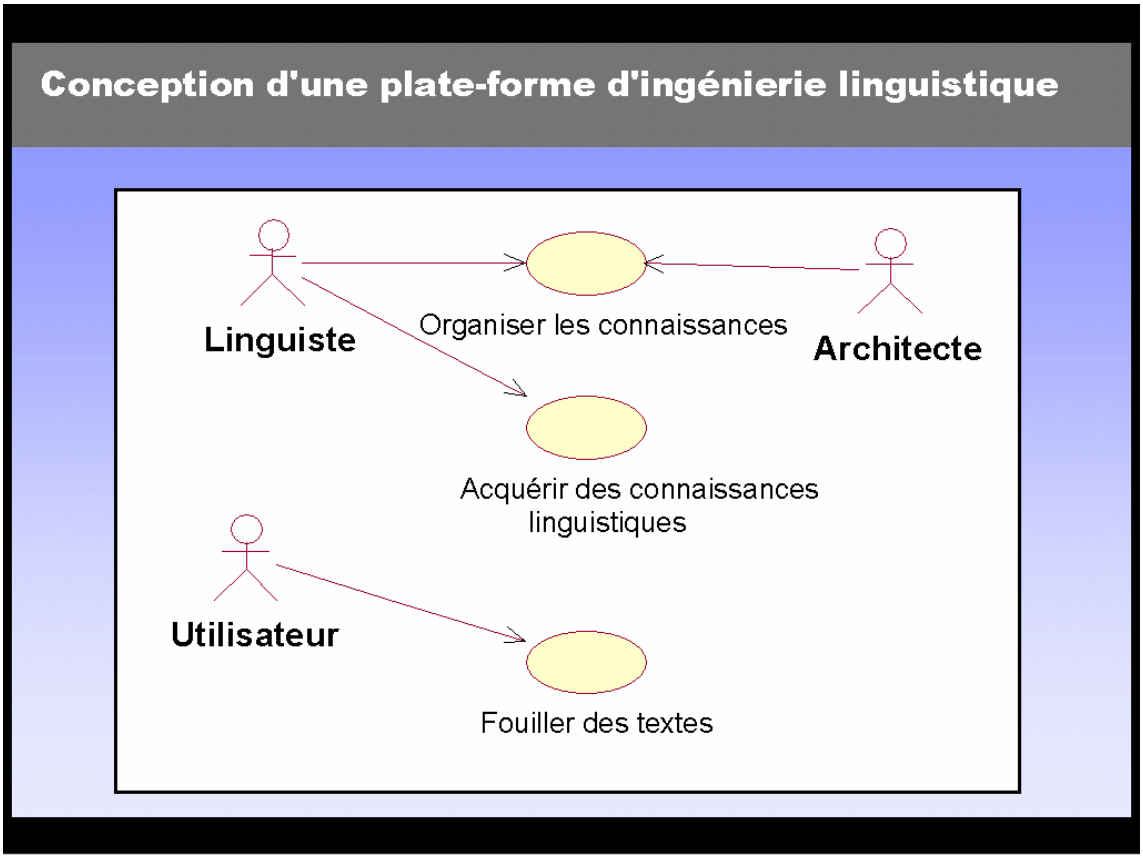


Figure 19. Cas d'utilisation de la plate-forme FilText

L'architecte spécifie le modèle conceptuel, langage de description des données linguistiques et construit les vues de travail nécessaires au linguiste et à l'utilisateur final.

Le linguiste, dans la phase d'acquisition des connaissances, doit disposer d'outils flexibles de gestion des marqueurs, indicateurs et indices. En effet, pendant cette phase les connaissances accumulées sont volatiles et les classes qui sont construites doivent pouvoir être facilement modifiées. De même, les règles sont dans un premier temps simplement ébauchées, puis enrichies au fil du processus. Remarquons que des outils de fouilles de corpus, extérieurs à la plate-forme, peuvent être utilisés et qu'il convient alors de proposer des formats d'échange standard.

L'utilisateur final doit disposer de fonctionnalités qui lui permettent d'interpréter les résultats de la fouille des textes ; ce qui signifie que des connaissances spécifiques à la tâche et à la présentation de

[33] Nous utilisons le langage UML, norme de modélisation objet adoptée par *l'Object Management Group.*

ces résultats doivent pouvoir venir enrichir plate-forme FilText. Ce point est très important en TAL, car il n'est pas possible d'envisager de développer une plate-forme qui réponde à tous les besoins. Il est nécessaire de concevoir celle-ci de telle manière qu'elle puisse accueillir des développements spécifiques ou que réciproquement, les résultats produits par la plate-forme puissent être exploités par ceux-ci. C'est donc l'inverse du concept de boîte noire.

3.2. Architecture de FilText

Conceptuellement, la plate-forme FilText s'organise en plusieurs systèmes (cf. fig. 20) qui coopèrent :

- un système de gestion des connaissances linguistiques qui a pour charge d'accueillir les connaissances linguistiques dans un modèle conforme à la méthode d'exploration contextuelle. Il est donc doté d'un interpréteur du langage de description de ces connaissances linguistiques ;
- un moteur d'exploration contextuelle qui a pour charge d'appliquer sur un ou plusieurs textes les connaissances linguistiques d'une ou plusieurs tâches de fouille textuelle ;
- un ensemble d'agents spécialisés, munis de connaissances et de modèles de présentation des résultats produits par le moteur d'exploration contextuelle.

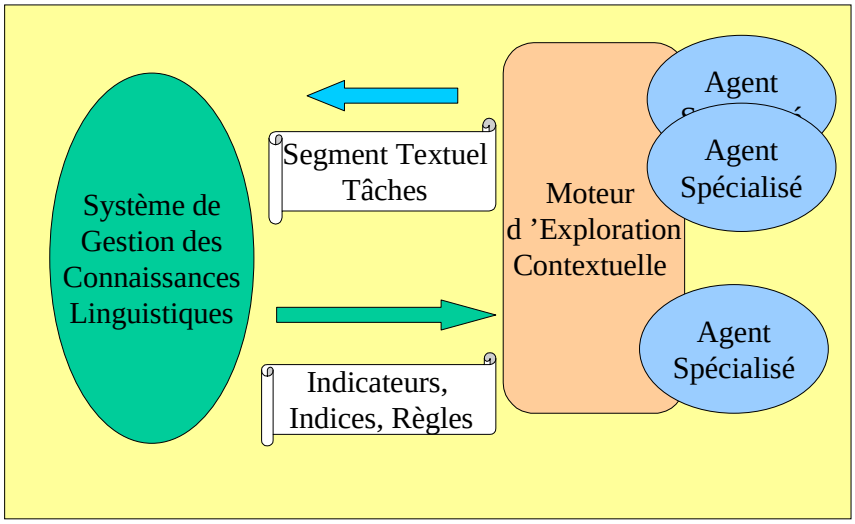


Figure 20. Architecture de la plate-forme FilText.

La coopération de ces systèmes s'articule autour d'un modèle hiérarchisé (cf.. fig. 21) du texte. C'est ce modèle du texte qui va être décoré avec les résultats produits par le moteur d'exploration

contextuelle en appliquant les règles d'exploration contextuelle. Ce modèle est construit à partir d'informations de balisage qui sont présentes dans le texte ou qui seront dynamiquement construites en analysant les marques de surface du texte. Il ne peut plus ensuite être altéré dans sa structure par l'application des règles d'exploration conceptuelle.

Ce choix conceptuel est cohérent avec les principes de la méthode d'exploration contextuelle exposés précédemment, et d'autre part motivé par le souci d'offrir un noyau commun et stable aux différents systèmes, éventuellement externes à Filtext, qui coopèrent. Jusqu'à présent ce choix n'a pas soulevé de difficultés, exception faite du système d'identification des propositions. Soulignons que ce choix ne signifie pas qu'il ne soit pas possible de créer d'autres structures qui viendraient se greffer sur ce modèle, mais simplement que ces nouvelles structures ne peuvent être exploitées que par d'autres tâches ; nous en verrons un exemple lorsque nous présenterons l'agent spécialisé qui construit des résumés par filtrage sémantique.

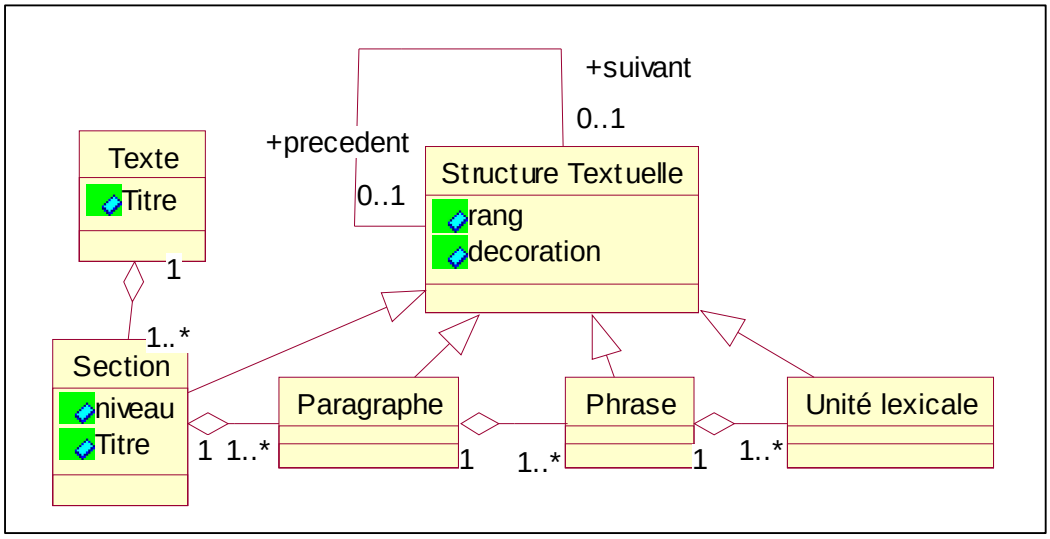


Figure 21. Modèle partiel de représentation d'un texte.

3.3. Plate-forme logicielle ContextO

La plate-forme logicielle ContextO [CRI 99a, CRI99b, CRI 99c, MIN 01, BEN 02, CRI 02] qui implémente FilText se devait d'être un système le plus ouvert possible ce qui impliquait :

- le choix d'un langage de développement favorisant une architecture ouverte ;
- le choix de standards pour les formats d'échange.

Nous avons ainsi opté, en 1998, pour l'utilisation de la technologie objet et plus particulièrement le langage orienté objet JAVA[34], notre préoccupation principale étant une portabilité la plus large possible et la volonté de développer des composants logiciels réutilisables à moindre coût par la communauté du TAL. En effet, comme le souligne P. Laublet dans [LAU 98] :

> *« Ces offres technologiques multiples proposent des solutions pour une informatique sans cesse plus hétérogène et plus distribuée dans des réseaux locaux ou globaux. Elles s'appuient sur la métaphore des objets actifs communiquanst entre eux et ouvrent des perspectives pour une large (ré)utilisation de composants logiciels. Les bénéfices attendus sont des logiciels de meilleure qualité : correction, robustesse, fiabilité, efficacité, portabilité et extensibilité comme notées par B. Meyer [MEY 88]. Ces logiciels sont plus modulaires, présentent un taux plus élevé d'utilisation de modules déjà codés, et disposent avant tout de meilleures capacités d'évolutivité. Celle-ci peut être définie, dans ce contexte, comme la capacité de modifier, de raffiner et d'étendre un ensemble existant de classes en réponse aux changements des besoins utilisateurs ou des contraintes opérationnelles. »*

Il faut de plus signaler que ce choix s'est révélé particulièrement adapté au développement distribué et incrémental qui est en général caractéristique du mode de travail d'une équipe de recherche[35]. Les développements réalisés ou en cours mettent ainsi à profit la notion d'interface logicielle de programmations (API).

Le choix d'un format d'échange standard des ressources linguistiques utilisées ou produites par la plate-forme ContextO est aussi un élément important qui a guidé notre démarche et nous partageons pleinement le diagnostique que porte à ce sujet L. Romary :

> *« L'absence d'un cadre unifié permettant de diffuser largement les ressources linguistiques disponibles dans les laboratoires explique en grande partie la faible diffusion d'outils au sein de la communauté du traitement automatique des langues. Ainsi, si l'on examine un certain nombre de plates-formes d'analyse syntaxique, on observe que, même aux niveaux les plus élémentaires, les lexiques sur lesquels elles reposent ne sont pas du tout interopérables. La conséquence immédiate est que tout chercheur qui s'implique dans l'utilisation d'un outil particulier est contraint, sur le long terme, de se tenir au format qui lui est imposé et que se forme ainsi une*

[34] Suite aux conseils avisés et judicieux de Philippe Laublet.

[35] La réalisation de la plate-forme logicielle ContextO est le fruit d'une collaboration entre l'équipe LaLIC du CAMS et l'Université de la République (Uruguay) ; elle a reçu le soutien du programme ECOS-Sud (Action n° U97E01).

communauté ayant du mal à se remettre en cause et extrêmement fragmentée. »
[ROM 00] p. 195

Nous avons ainsi opté pour le langage XML [XML 98] qui fournit le vocabulaire et les principes pour la description et l'échange de tout type d'informations numériques. En tant que méta-language, XML permet de définir un modèle de codage propre à chaque application. Mais le risque de voir se multiplier des descriptions hétérogènes est limité par les principes mêmes de XML qui distingue les documents « bien formés » et les documents « valides ». La définition d'une DTD (Definition de Type de Document) offre ainsi à l'utilisateur rigoureux la possibilité d'éditer un document « valide » qui respecte le modèle de la DTD et ensuite de diffuser un document « bien formé » dans la communauté. Un certain nombre d'associations internationales du domaine des sciences humaines ont élaboré une représentation générique connue sous le terme de TEI (Text Encoding Initiative) qui propose un format commun de représentation de ressources textuelles informatisées. Le tableau 6 contient la DTD que nous avons définie pour les textes exploitables ou produits par ContextO.

```
<?xml version="1.0" encoding="ISO-8859-1" standalone="yes"?>
<!DOCTYPE texte-contexto [
<!ELEMENT texte-contexto (h1+)>      // une section au moins de type H1
<!ELEMENT h1 (t1?, (p+ | (p*, h2+) ))>
<!ELEMENT t1          (#PCDATA)>   // titre
<!ELEMENT h2 (t2?, (p+ | (p*, h3+) ))> // sous sections H2 à H5
<!ELEMENT t2          (#PCDATA)>
<!ELEMENT h3 (t3?, (p+ | (p*, h4+) ))>
<!ELEMENT t3          (#PCDATA)>
<!ELEMENT h4 (t4?, (p+ | (p*, h5+) ))>
<!ELEMENT t4          (#PCDATA)>
<!ELEMENT h5          (t5?, p+)>
<!ELEMENT p (a+)>                    // paragraphe
<!ELEMENT a (tk, lm?, tx?)+ >        // phrase
<!ELEMENT tk          (#PCDATA)> // le token
<!ELEMENT lm          (#PCDATA)>  // la forme
<!ELEMENT tx          (#PCDATA)>  // l'etiquette morphosyntaxique
]>
```

Tableau 6. DTD d'un texte analysable par ContextO.

Il faut aussi souligner que le choix de ces deux langages, Java et Xml, nous ont offert l'accès à de nombreuses ressources logicielles, nous permettant ainsi de développer beaucoup plus rapidement nos outils spécifiques et de nous concentrer sur notre problématique de recherche.

Pour être analysable par Contexto, un texte doit nécessairement être balisé en sections, paragraphes, phrases ; le balisage en mots étiquetés morpho-syntaxiquement est par contre optionnel. Si le balisage en sections et paragraphes peut être automatisé en exploitant les marques spécifiques de fin de paragraphe, il n'en est pas de même pour le balisage des phrases. En effet, la ponctuation est souvent ambiguë comme le montre la figure 22 extraite de [MOU 99].

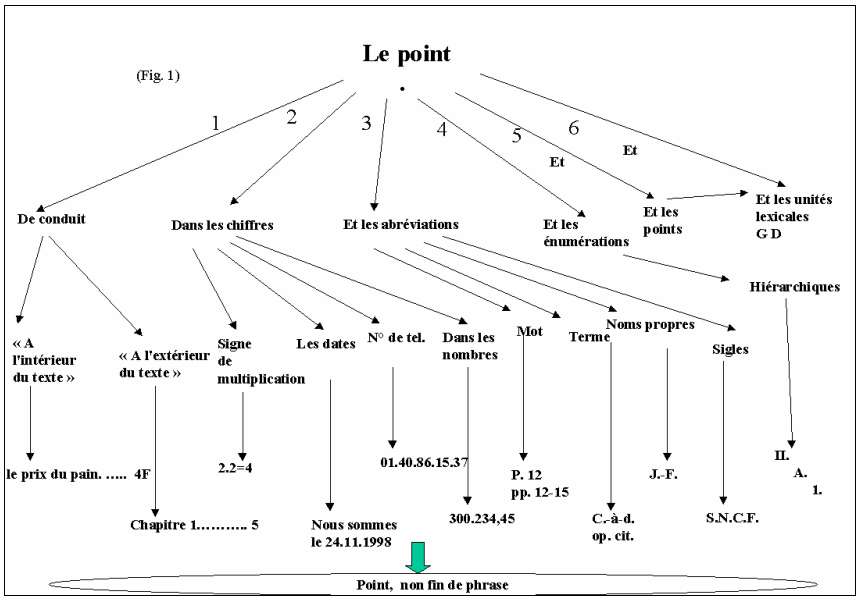


Figure 22. L'ambiguïté du signe point [MOU 99].

<L 'espace BL> <Le point PT> <Le deux-points DPT> <Le point virgule PTV> <Le point d'interrogation PINT > <Le point d'exclamation PTEX> <Parenthèse ouvrante PO> <Parenthèse fermante PF> <Crochet ouvrant CO> <Crochet fermant CF> <Tiret TIRET> <Guillemets GI> <Retour à la ligne \r> <Début de ligne \n> <Tabulation \t> <Chiffres Arabes CA> <Chiffres Romains CR> <Lettres majuscules LM> <Lettres minuscules LM> <Un mot qui commence par une majuscule MMaj> <Un mot qui commence par une minuscule MMin> <Liste des particules d 'interjection (Ah, oh, Ouf…) LPI> <Liste des abréviations toujours suivies d'un chiffre ou d'une majuscule (p., pp., MM., …) LA> <Liste des indices pour la segmentation dans le cas de deux-points (http, « , …) LIDPT> <Liste des indices qui sont suivis par une lettre majuscule (vitamine C, Hépatite A,…) LILM> <Texte TXT>.

Tableau 7. Unités graphiques permettant de segmenter les textes. [MOU 99].

Un segmenteur [MOU 99] de textes en phrases a été développé par G. Mourad dans le cadre de sa thèse et intégré dans la plate-forme ContextO. Ce segmenteur se fonde sur la méthode d'exploration contextuelle pour décider, en fonction du contexte, si un signe de ponctuation peut être interprété comme une fin de phrase. Le tableau 5 dresse l'inventaire de toutes les unités graphiques qui sont prises en compte pour prendre cette décision. On remarque que ce tableau contient des listes d'unités lexicales, notamment des abréviations, des termes de langages spécialisés comme par exemple *Vitamine A* ,ce qui rend illusoire toute tentative d'obtenir un résultat totalement fiable, puisque dépendant de domaines. Ce constat fait d'autant plus problème que la plupart des analyseurs morpo-syntaxiques actuels réalisent leur propre découpage en phrases et ce que découpage est en général de qualité médiocre notamment sur les textes scientifiques ou techniques ; la cause principale étant l'usage fréquent des abréviations et des références bibliographiques. Nous avons enrichi ce

segmenteur afin qu'il intègre la reconnaissance des titres et de certaines structures textuelles comme les énumérations.

3.3.1. Architecture logicielle de ContextO

L'architecture de ContextO (cf. fig. 23) est conçue comme un ensemble de sous-systèmes, chacun de ceux-ci offrant une façade de communication (au sens UML du terme). Ce principe d'architecture garantissant la pérennité des développements[36]. Nous décrivons ci-après les fonctionnalités de chacun de ces sous-systèmes.

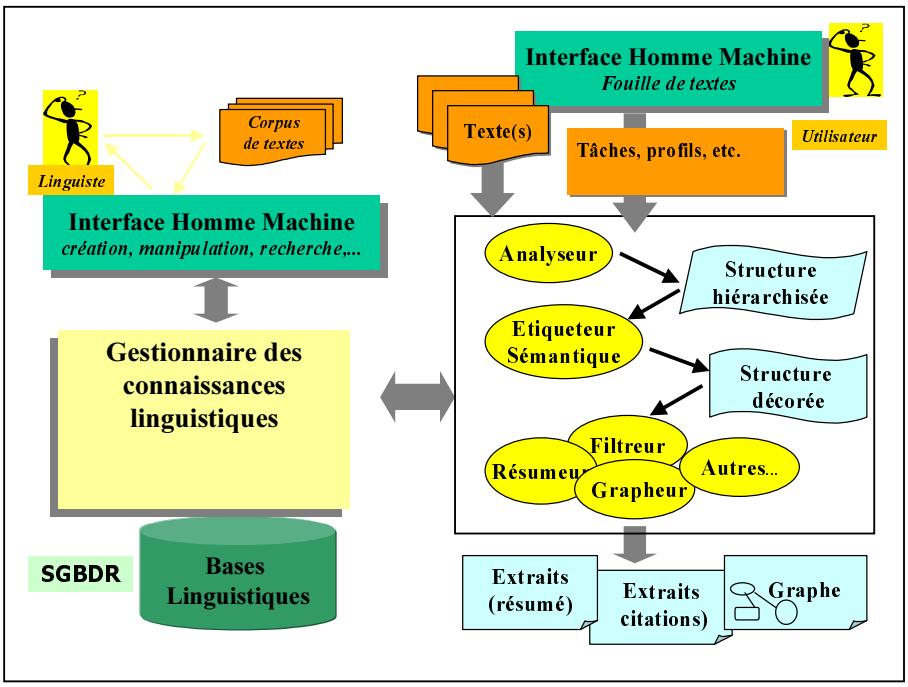


Figure 23. Architecture de la plate-forme logicielle ContextO.

Gestionnaire des connaissance linguistiques

Le gestionnaire des connaissances linguistiques [BEN 99] implémente les fonctionnalités décrites précédemment au paragraphe 2.2. Il permet au linguiste d'organiser en classes les indices et les indicateurs. La persistance de ces données linguistiques repose sur un serveur implanté techniquement avec un système de gestion de bases de données relationnelles (cf. fig. 23). Des pilotes d'accès programmés en Java dialoguent avec des IHM (Interfaces Homme Machine) qui produisent des vues conformes aux besoins de l'utilisateur.

[36] La plate-forme ContextO en est à sa troisième version.

Le choix de s'appuyer sur une base de données relationnelles nous a aussi permis de concevoir une architecture distribuée et personnalisée des connaissances linguistiques.

- La distribution est organisée à partir d'un référentiel qui contient toutes les données linguistiques de toutes les tâches décrites par les linguistes. Le schéma relationnel de cette base est stable et correspond au modèle décrit précédemment ;

- La personnalisation permet à un linguiste de définir sa propre vue des données linguistiques à partir des vues extraites du référentiel (cf. fig. 25). Cette fonctionnalité est très importante pendant la phase d'acquisition des connaissances pendant laquelle le linguiste à besoin d'annoter temporairement ses données comme le souligne S. Porhiel dans [POR 01].

La synchronisation des données est effectuée par des outils conçus notamment par F. Lepriol dans le cadre de sa thèse [LEP 00].

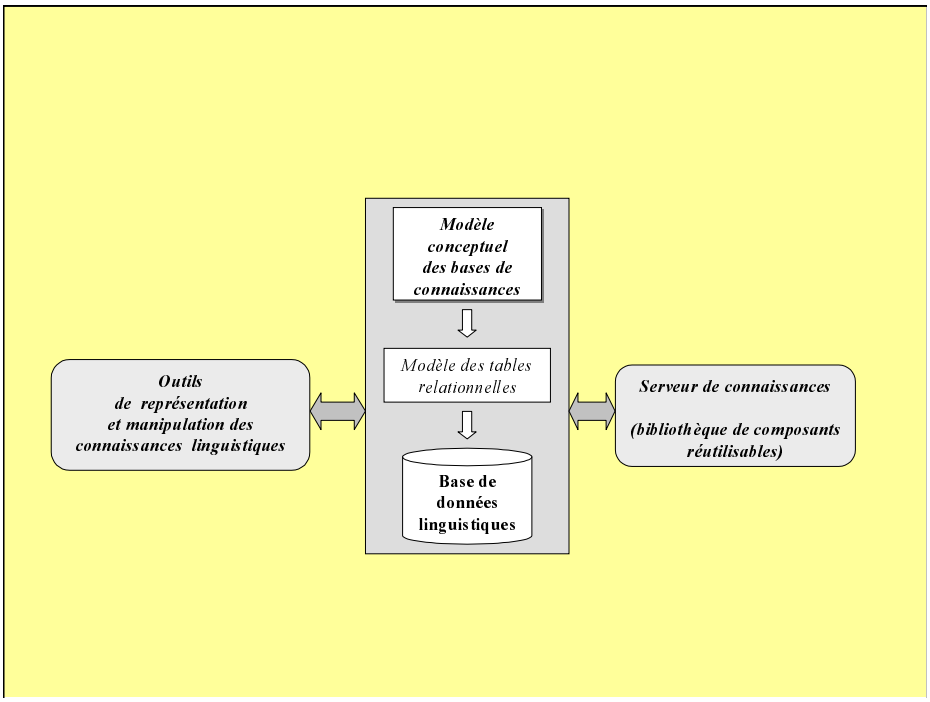


Figure 24. Architecture du gestionnaire de connaissances linguistiques [BEN 99]

Ce gestionnaire a aussi pour charge le repérage des marqueurs, indicateurs et indices, dans un segment textuel que lui fournit le moteur d'exploration contextuelle. Un segment textuel est considéré comme une suite w_i d'unités lexicales, éventuellement composées, ou l'indice i représente le rang de l'unité lexicale dans le segment. La position d'un marqueur est alors représenté par un couple (k,l) représentant respectivement le rang de début et le rang final de ce marqueur dans le segment textuel.

Cette représentation permet de traiter des configurations ou des indices qui sont intercalés dans un indicateur composé.

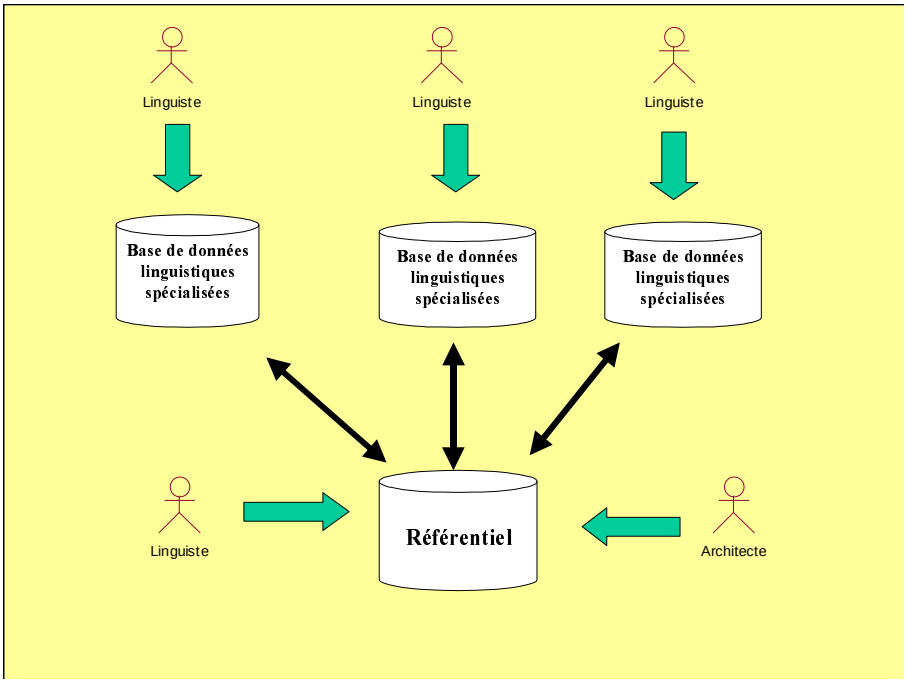


Figure 25. Architecture distribuée et personnalisée

Dans la version à venir de ContextO, ce système intégrera un interpréteur du langage Ltext [BEN 00] évitant ainsi l'écriture des règles d'exploration contextuelle en Java.

Moteur d'exploration contextuelle

Le moteur d'exploration contextuelle a pour charge d'exploiter les données linguistiques pour une ou plusieurs tâches choisies par l'utilisateur. Il est composé de deux systèmes qui coopèrent, l'analyseur de textes et l'étiqueteur sémantique.

Analyseur de texte

L'analyseur de textes construit une représentation qui reflète l'organisation hiérarchique du texte, le traitement textuel nécessitant en effet qu'une tâche spécialisée puisse se focaliser sur les *n* unités lexicales de la *i*ème phrase du *j*ième paragraphe de la *k*ième section. La construction de cette structure hiérarchisée s'appuie sur le texte segmenté produit par le segmenteur [MOU 99] et le « tokeniseur » [BEN 99] déjà décrits précédemment. L'analyseur ne vérifie pas la cohérence de cette structure puisque celle-ci a été produite par le segmenteur ou par un analyseur XML conformément à la DTD ContextO.

Etiqueteur sémantique

Pour chacune des phrases qui composent le texte analysé, l'étiqueteur sémantique adresse une requête au gestionnaire des connaissances linguistiques afin que celui-ci identifie les positions des indicateurs et des indices et le nom des règles d'exploration contextuelle qui peuvent être déclenchées. L'étiqueteur déclenche l'exécution des règles qui sont écrites en Java, chaque règle correspondant à une méthode d'une classe. Ce choix permettant de mieux identifier un problème éventuel dans une règle.

```
public static void RCenthe131(Texte texte,Phrase ph,RegleEC regle,Integer pi, Integer pf)
{/*
* capte un schema du type : apporter, mener, repondre
* avec comme argument : question, argumentation
*                    pages articles qui suit, ci-dessous
* Classe de l'indicateur : &Caction-enonc1.1, &Caction-enonc1.2 ,
*                      &Caction-enonc1, &Caction-enonc1.4
* Etiquette : Thematique*/
IL1 = "&Coperat-enonc1";
 L1 = "&Cobjectif3";
 L2 = "&document2.1";
 L3 = "&Caction-enonc3.1";
TacheEC ta = texte.getmemo().gettache();
EspaceRecherche econd1 = new EspaceRecherche(ta,texte, ph, pi, pf);
 ensemblesousespacesc1 = econd1.voisinage(5);
if  (!explorerContexte(ta,ensemblesousespacesc1,IL1))
return;
EspaceRecherche econd2 = new EspaceRecherche(ta,texte, ph, pi, pf);
 ensemblesousespacesc2 = econd2.voisinage(10);
if ( (!explorerContexte(ta,ensemblesousespacesc2,L1))
   &&(!explorerContexte(ta,ensemblesousespacesc2,L2))
   && (!explorerContexte(ta,ensemblesousespacesc2,L3)) )
return;

ajoutetiquette(ph,regle.getEtiquette(), texte) ;}
```

Figure 26. Une règle écrite en Java

Le code d'une règle se présente comme dans l'exemple de la figure 26 ; on peut constater que ce code reste très proche du langage Ljava avec lequel le linguiste décrit ses règles. Néanmoins, comme nous l'avons déjà souligné, le codage des règles dans un langage de programmation est un maillon faible dans la conception de ContextO, puisqu'il introduit un cycle de compilation où de simples erreurs typographiques ne sont identifiées qu'en fin de cycle.

L'étiqueteur sémantique déclenche, pour toutes les tâches choisies par l'utilisateur, toutes les règles associées à celles-ci. Les règles sont en effet considérées comme indépendantes ; l'ordre de leur déclenchement, pour une tâche donnée, est indifférent. Ce mode de fonctionnement correspond à l'hypothèse que, pour une tâche donnée, certains marqueurs sémantiques ne sont pas exclusifs entre

eux. Par exemple, la présence d'une négation dans une phrase conclusive n'implique pas que cette phrase ne soit pas par ailleurs une « conclusion ». Par ailleurs, les étiquettes attribuées par différentes tâches ne sont pas incompatibles entre elles. Ainsi une phrase étiquetée comme « définitoire » peut aussi être étiquetée comme « conclusion ». Toutes les déductions effectuées par les règles sont attribuées aux éléments qui composent la hiérarchie du texte et produisent ainsi une structure hiérarchisée « décorée » par des informations sémantiques.

Agents Spécialisés

Un agent spécialisé exploite les « décorations sémantiques» du texte en fonction de ses connaissances, qui correspondent à des objectifs définis par l'utilisateur. Un agent spécialisé possède ses propres IHM de présentation des résultats obtenus après l'étiqueteur sémantique. Cette notion d'agent spécialisé rend ainsi la plate-forme ContextO ouverte à tout système extérieur puisque des API ont été définies pour accéder à la structure hiérarchisée. Les agents spécialisés permettent ainsi de développer des traitements spécifiques pour un utilisateur tout en exploitant le modèle générique de traitement des connaissances linguistiques ; ils se rapprochent en cela de la notion « d'intelliciel » développée au LANCI [MEU 98].

Pour illustrer ce concept Nous allons présenter deux exemples d'agents spécialisés, le résumeur filtreur [MIN 01] et SEEK-JAVA [LEP 00].

3.3.2. Exemples d'agents spécialisés

3.3.2.1. Résumeur Filtreur

Cet agent exploite les connaissances du système SERAPHIN fruit d'une collaboration avec la Direction des Études et des Recherches (DER) de la société EDF et de SAFIR, que nous avons décrit dans le chapitre précédent. Rappelons que les connaissances linguistiques de ces systèmes permettent le repérage des énoncés structurants, de connaissances causales et des définitions. Nous allons montrer comment un agent spécialisé dans la plate-forme ContextO peut exploiter les décorations sémantiques pour réaliser un ou plusieurs résumés.

Certaines phrases étant étiquetées, il devient possible de sélectionner un sous ensemble de ces phrases pour construire un extrait ciblé pour un utilisateur. L'algorithme de sélection qui constitue les connaissances de l'agent spécialisé, se compose d'une stratégie d'exploration (SE) du texte, de profils de filtrage (Pf_i), et d'un seuil de filtrage (S).

La stratégie d’exploration SE précise l'ordre et la profondeur d’exploration (et donc de sélection) des sections du texte. Par exemple, une stratégie standard explore le texte linéairement en sélectionnant les phrases qui correspondent au profil de filtrage. Par contre, une stratégie entrelacée privilégie l’exploration de l'introduction et de la conclusion, en tenant compte aussi de la structure en paragraphes de ces deux segments textuels, puis poursuit une exploration linéaire du texte. D’autres stratégies peuvent être définies au gré des besoins spécifiques d’un utilisateur. La profondeur d’exploration permet ainsi de ne pas prendre en compte les sections les plus profondes qui correspondent généralement à des explications détaillées de l’auteur sur un point précis et de privilégier d’autres informations.

Les profils de filtrage Pf_i sont soit prédéfinis, soit construit interactivement par l'utilisateur (cf. fig. 27). Ils précisent l'importance de chaque étiquette sémantique. Ainsi pour un certain type de recherche, des énoncés « conclusifs » sont considérés comme plus importants que les énoncés « d'annonce thématique » lorsqu’ils se trouvent dans la dernière section du texte. Un profil de filtrage se présente sous la forme d'une liste hiérarchisée d'étiquettes, ce qui permet aussi d'ignorer un certain type d'informations.

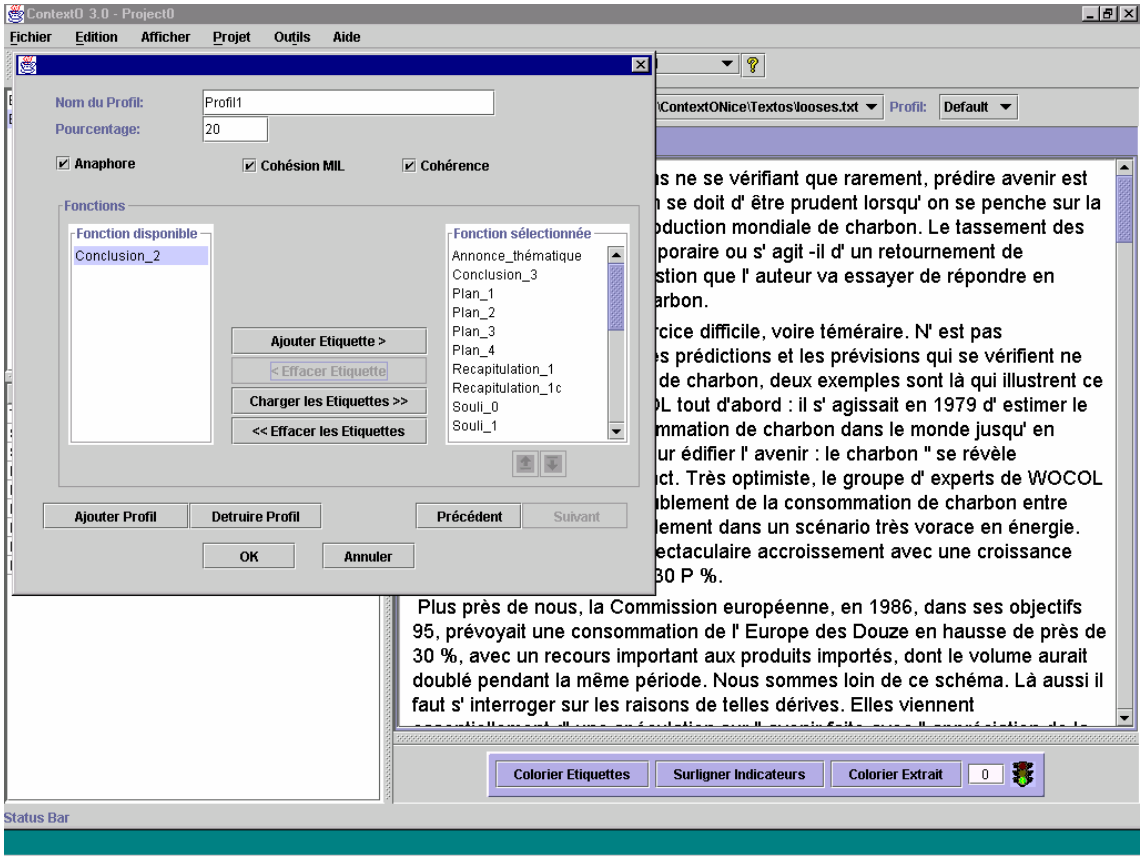


Figure 27. Construction d’un profil de filtrage [COU 01]

Le seuil de filtrage S permet de produire un extrait composé d'un nombre de phrases relativement à la taille du texte source ; par exemple, le seuil de sélection peut être de 10 % ou 20 % du texte source.

Chaque agent spécialisé possède ses propres IHM de présentation. La figure 28 présente deux des interfaces de l'agent résumeur. Ces interfaces ont été développées par J. Couto[37] [COU 01]dans le cadre d'un master de l'Université de la République de Montevideo. Dans l'IHM supérieure, la fenêtre supérieure droite affiche le texte source dans lequel les phrases de l'extrait sont surlignées. L'extrait est affiché dans la fenêtre inférieure droite. Le simple fait de modifier le profil de filtrage permet d'obtenir immédiatement un autre extrait.

De manière générale on peut considérer que cette extraction brise la cohérence du texte source et peut même introduire des contresens. Le développement de ce type d'IHM qui permet à l'utilisateur de naviguer entre l'extrait et le texte source constitue une des réponses à ce problème offertes par ContextO. Par ailleurs, plutôt que de vouloir simuler le travail d'un résumeur professionnel en produisant un résumé indépendant du texte source, nous avons cherché à construire un nouvel « objet textuel » qui articule des données textuelles « décorées » et des procédures de fouille de ces données. Le résumé n'est plus considéré comme indépendant du texte dont il est issu. L'informatique, et plus généralement les outils du multimédia fournissent en effet des fonctionnalités qui permettent d'offrir au lecteur les moyens de naviguer entre le résumé et le texte. Plutôt que de chercher à produire un résumé autonome en résolvant des problèmes comme la résolution des anaphores, le repérage des liens de cohésion et de cohérence, l'objectif se déplace vers la production d'un texte réduit aux informations jugées saillantes pour le lecteur, et la construction de liens qui permettent au lecteur, au vu des informations partielles qui lui sont présentées, de fouiller, à la demande, le texte source.

La figure 28 montre deux des interfaces du système. La première, celle du haut sur la figure 28, visualise le texte source dans la fenêtre supérieure et le résumé obtenu est affiché dans la fenêtre inférieure. Le fait de positionner le curseur sur une phrase du résumé permet de voir, dans le texte source le contexte complet de cette phrase. Enfin, il est possible de visualiser, dans le texte source, les phrases qui ont été sélectionnées. Ces phrases sont « colorisées » afin d'attirer l'attention de l'utilisateur.

L'utilisateur peut aussi choisir un deuxième type d'interface, celle du bas sur la figure 28, où le texte est présenté sous la forme d'une structure arborée ; celle-ci permet à l'utilisateur de visualiser rapidement dans quelles parties du texte se trouvent les phrases qui constituent l'extrait.

Enfin, des caractéristiques du texte, nombre de phrases, nombre de phrases étiquetées, etc., sont systématiquement affichées.

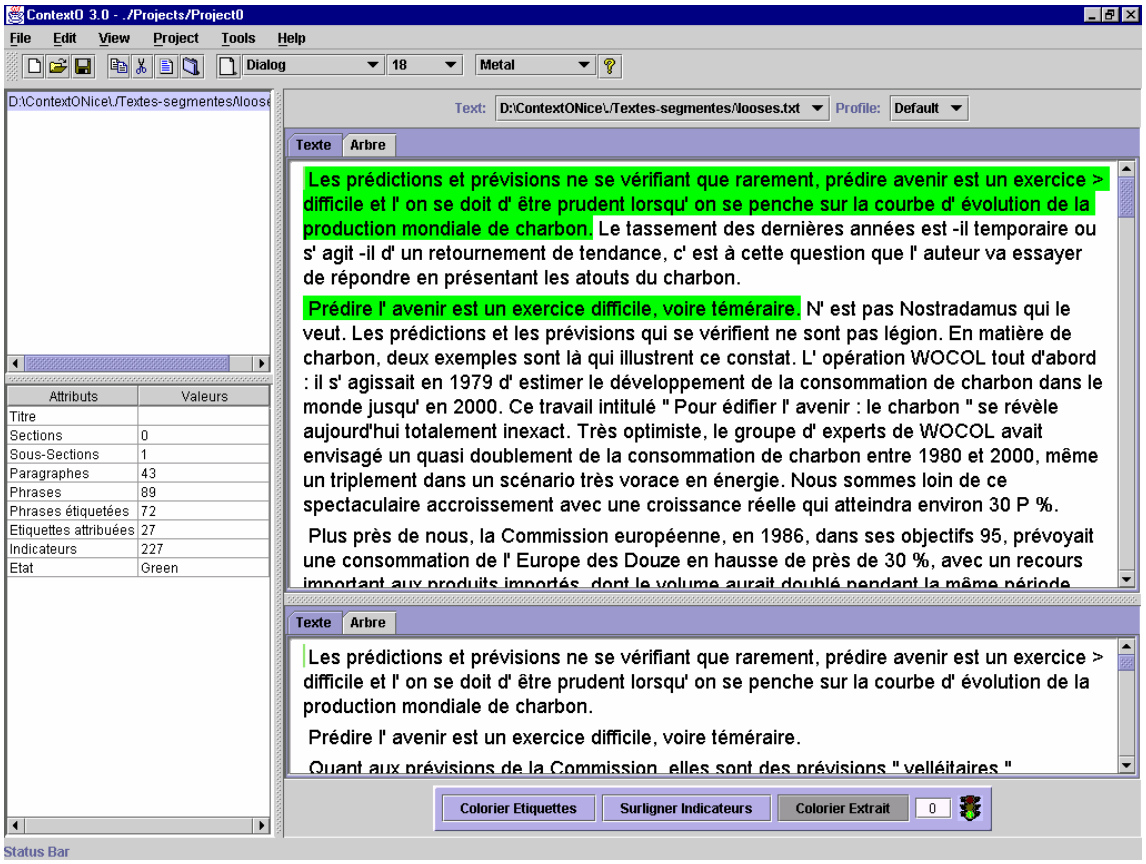
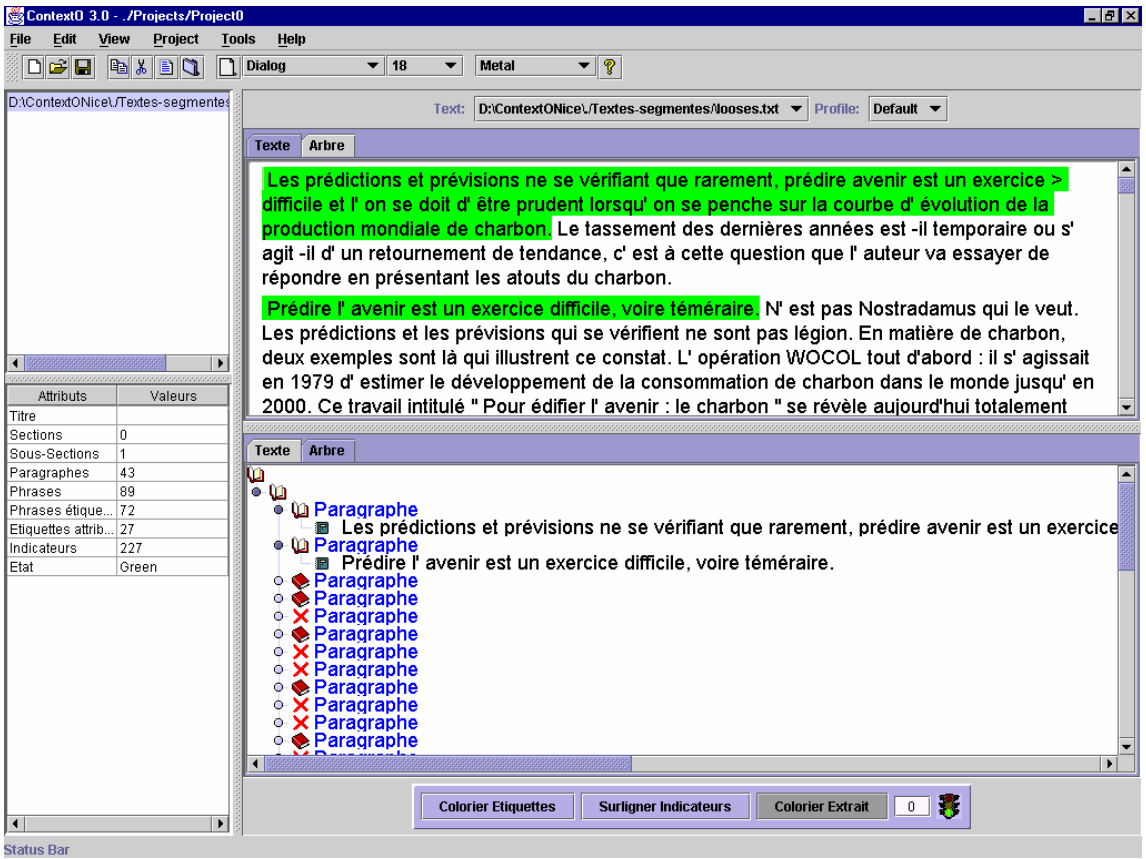


Figure 28. Deux des interfaces de l’agent résumeur filtreur [COU 01]



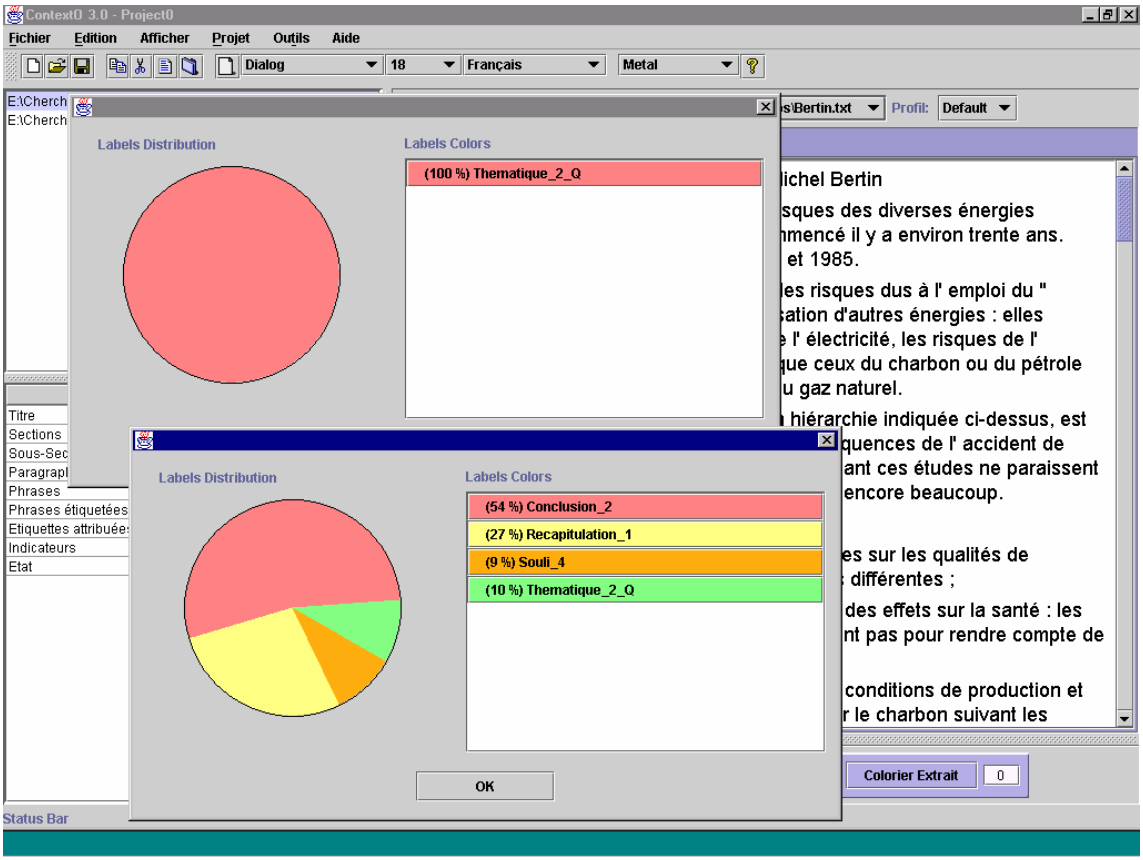


Figure 29. Interface de l'agent résumeur filtreur [COU 01]

La figure 29 illustre le type comparaison qu'il est possible d'effectuer entre plusieurs textes en exploitant les étiquettes sémantiques qui ont été attribuées aux phrases des textes. On voit sur cet exemple que sur un texte le système n'a repéré que des annonces thématiques, alors que sur l'autre texte, les étiquettes sont plus distribuées. Il nous semble que ce type d'outils ouvre la voie à des travaux sur les texte que l'on n'avait pas envisagé jusqu'à présent.

Enfin la figure 30 montre deux possbilités de présenter au lecteur, non plus les organisations structurelles, mais les organisions discursives du texte.

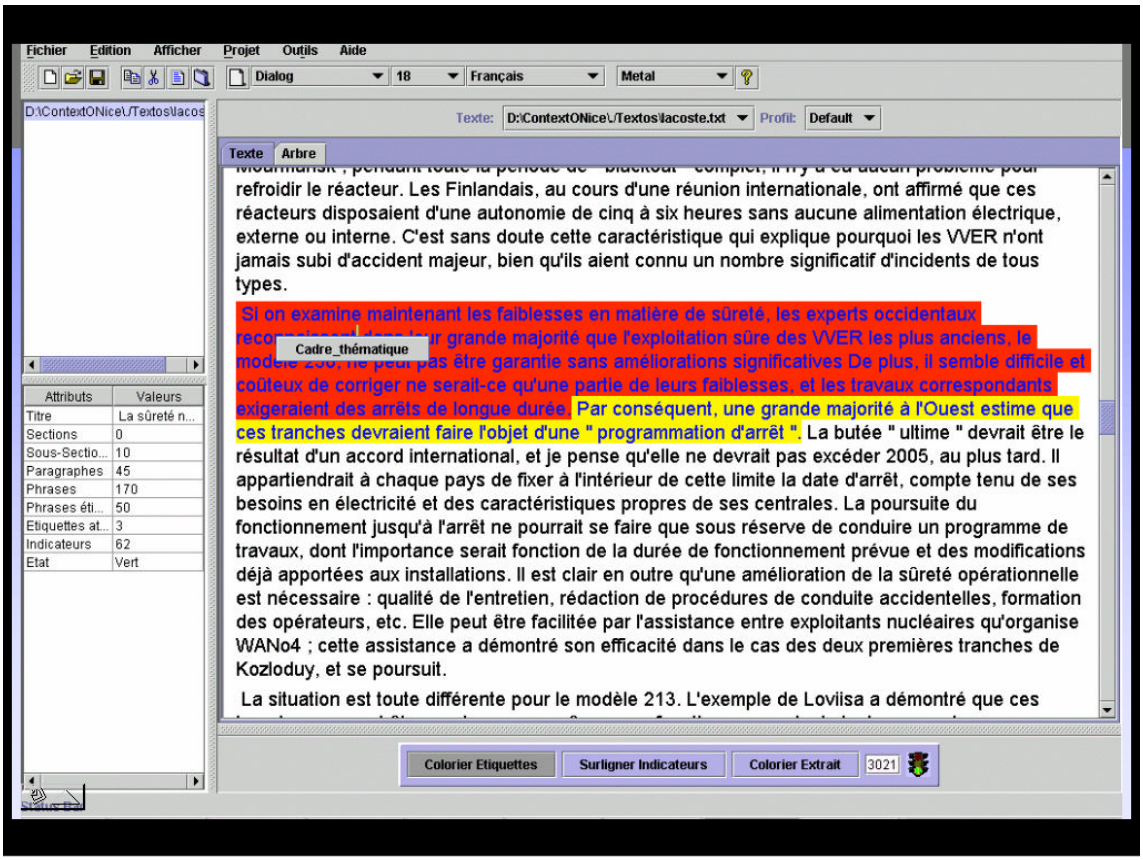
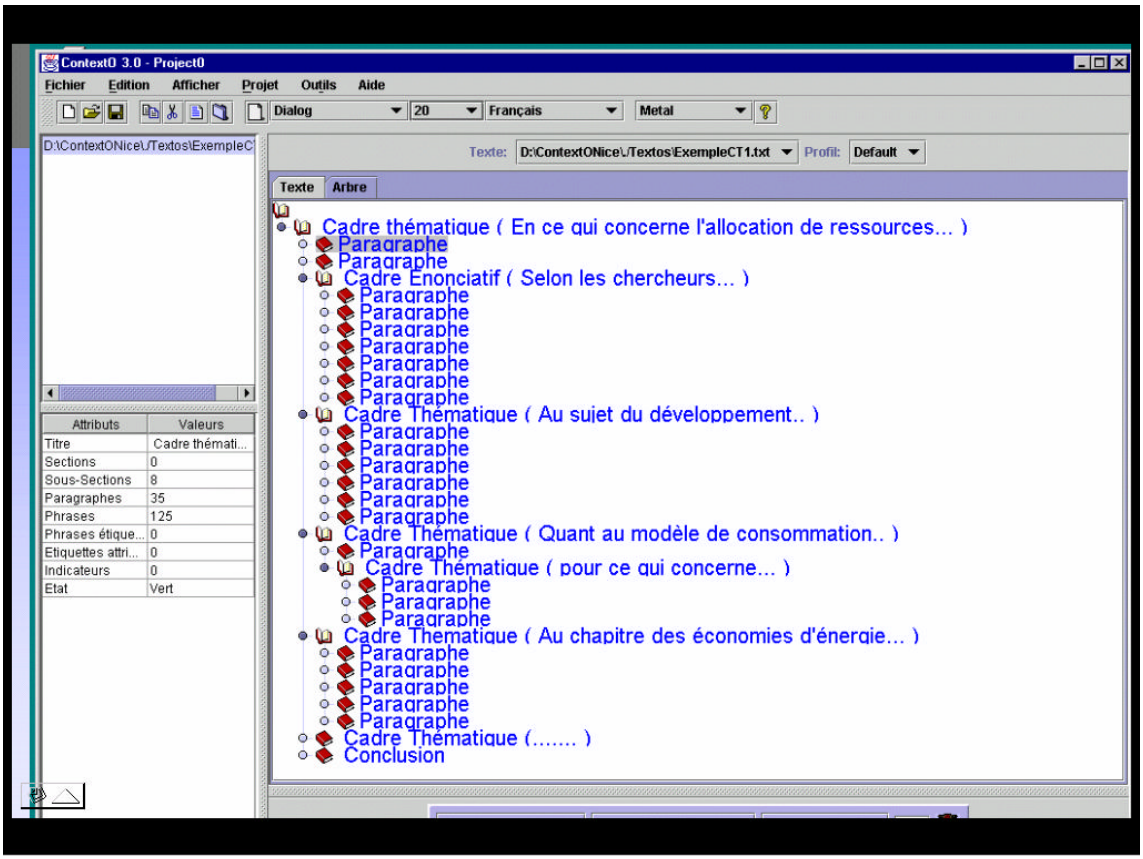


Figure 30. Présentations des organisations discursives



3.3.2.2. SEEK-Java

Cet agent spécialisé [LEP 99, 00] dont nous avons brièvement parlé dans notre introduction permet d'extraire des relations sémantiques entre concepts. Les relations identifiées (hiérarchies classes / sous-classe, attributs / valeur, instances de classes, relations entre un objet et ses parties, etc.) permettent d'enrichir un réseau terminologique en donnant des étiquettes sémantiques aux relations entre termes. Ces relations sémantiques sont représentées sous forme de graphes ou de tables et capitalisées dans une base de données permettant leur exploitation. L'utilisateur intervient uniquement au début du processus afin d'indiquer le thesaurus (ou liste de concepts du domaine) à charger. En fin de traitement, les résultats sont visualisés sous forme de graphes ou de tables comme l'illustre les figures ci-après.

Extraction des concepts

L'extraction des concepts, dans SEEK-JAVA est réalisée de manière automatique, en s'appuyant sur une liste des termes de référence du corpus ou un thesaurus du domaine traité par le corpus. L'extraction est considérée comme une deuxième phase d'exploration contextuelle. En effet, elle est basée sur deux listes de termes assimilées à des indices contextuels et sur quatre règles.

Règles de recherche des arguments

Tête : Rinc04
Tâche : Relation_Statique
Déclenchements
&Ietre B *est* une *spécification de* A
&Iverbeinc1 A *constitue* une *généralisation de* B
Corps
E1 := créer_espace(DonnerAterieures(I))
E2 := créer_espace(DonnerSuivantes(I))
L1 := &inc4
L2 := &inc5
L3 := &inc1
L4 := &LIngPrep2
Condition
Il_existe_un_indice x1 appartenant_à E2 tel_que classe_de x1 appartient_à (L1 ou L2 ou L3)
Il_existe_un_indice x2 appartenant_à E2 tel_que classe_de x2 appartient_à (L4)
Les_indices x1 x2 se_suivent.
Action
1. attribuerEtSem(PhraseParent_de I, "est_inclus")
2. rechercheArgument2SEa(PhraseParent_de I, E1, E2 , x1←x2)

Figure 31. Une règle du système SEEK-Java [LEP 00]

Les relations sémantiques extraites sont orientées. Aussi, les règles de recherche des arguments doivent non seulement trouver les arguments mais aussi ajouter, dans la base de données des

connaissances extraites, le triplet argument-étiquette-argument en respectant l'orientation de la relation. La recherche des arguments est réalisée en s'appuyant sur l'indicateur qui a déclenché la règle en cours d'exécution et des espaces de recherche associés.

Capitalisation et représentation des connaissances

Dans SEEK Java, les relations repérées dans un texte sont d'une part représentées graphiquement et capitalisées dans une base de données. Quela que soit la représentation choisie, il s'agit d'une image des connaissances extraites du corpus en cours de traitement, capitalisées dans la base de données .En effet, l'affichage des graphes ou de la table des relations est réalisé grâce à une requête vers la base de données permettant ainsi d'avoir à tout moment accès aux données les plus récentes. L'interface sous forme de tables donnant la possibilité de réaliser les opérations d'ajout, de suppression et de modification. Ces changements affectent directement la base de données et pas seulement l'image qui en est donnée.

Prenons un exemple afin de présenter cette base de données et les interfaces permettant de représenter les connaissances[38]. Cette base de données est constituée de deux vues, au sens relationnel du terme. Ces deux vues permettent de stocker les relations sémantiques et leurs arguments ainsi que la phrase de laquelle la relation a été extraite.

Le corpus iconographique.
Cinq grandes catégories de gravures rupestres peuvent être distinguées : les corniformes, les armes et les outils, les figures anthropomorphes, les figures géométriques et les figures non représentatives.
Les corniformes : corniformes simples, signes en T, attelages.
Les corniformes représentent 46 % des gravures.
Les armes et les outils : poignards et lames, hallebardes, faucilles, haches, armes diverses, claviformes.
Les armes et les outils représentent 4 % des gravures.
Les figures anthropomorphes : simples, complexes, corniformes anthropomorphisés, réticulés anthropomorphisés.
Les figures anthropomorphes représentent 0.2 % des gravures.
Les figures géométriques : cercles, croix, étoiles, réticulés, spirales ...
Les figures géométriques représentent 7 % des gravures.
Les figures non représentatives : cupules isolées, groupes de cupules isolées, plages, barres, figures inclassables.
Les figures non représentatives représentent 42.8 % des gravures.

Figure 32. Extrait du texte 3 du corpus « gravures rupestres » [LEP 00]

[38] Le texte de la figure 25 est extrait du texte 3 du corpus « gravures rupestres »

Les résultats présentés dans la figure 33, sont ceux obtenus à partir du traitement de l'extrait de texte. En réalité, Le graphe est en mouvement afin de se positionner dans la fenêtre de manière à ce que les termes affichés ne se chevauchent pas. Le mouvement peut être stopper et repris. L'utilisateur peut à l'aide de la souris déplacer les termes afin d'obtenir la disposition qu'il lui convient le mieux.

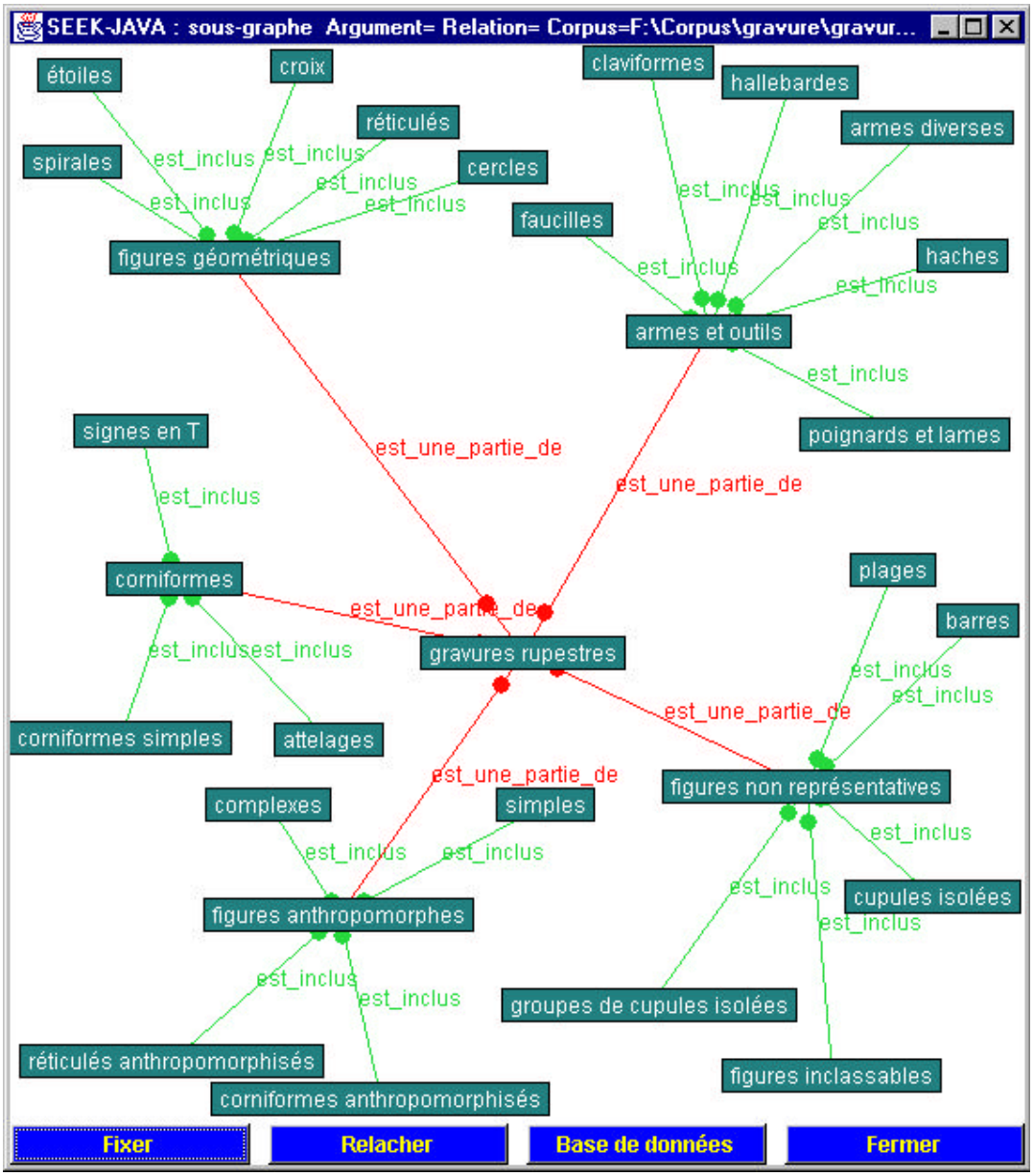
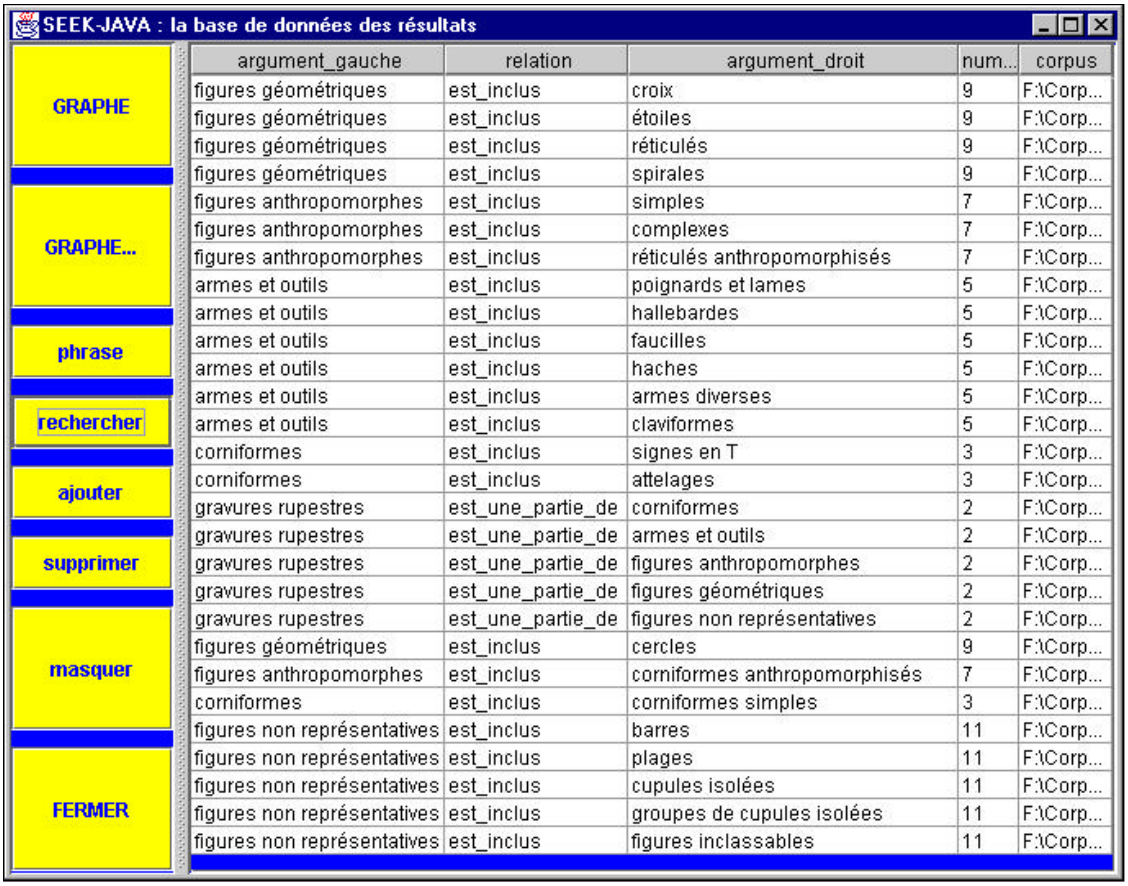


Figure 33 : IHM de l'agent spécialisé SEEK Java [LEP 00]

L'utilisateur peut accéder à une représentation tabulaire des mêmes données comme le montre la 34. La couleur des relations est fonction de leur étiquette : une couleur par étiquette sémantique et l'orientation des relations est symbolisée par un petit rond. A partir de l'interface sous forme de tables,

en plus des opérations classiques (bouton 'rechercher', bouton ajouter', bouton 'supprimer'), il est aussi possible d'accéder à la phrase dont est issue la relation et d'afficher des sous-graphes.



argument_gauche	relation	argument_droit	num...	corpus
figures géométriques	est_inclus	croix	9	F:\Corp...
figures géométriques	est_inclus	étoiles	9	F:\Corp...
figures géométriques	est_inclus	réticulés	9	F:\Corp...
figures géométriques	est_inclus	spirales	9	F:\Corp...
figures anthropomorphes	est_inclus	simples	7	F:\Corp...
figures anthropomorphes	est_inclus	complexes	7	F:\Corp...
figures anthropomorphes	est_inclus	réticulés anthropomorphisés	7	F:\Corp...
armes et outils	est_inclus	poignards et lames	5	F:\Corp...
armes et outils	est_inclus	hallebardes	5	F:\Corp...
armes et outils	est_inclus	faucilles	5	F:\Corp...
armes et outils	est_inclus	haches	5	F:\Corp...
armes et outils	est_inclus	armes diverses	5	F:\Corp...
armes et outils	est_inclus	claviformes	5	F:\Corp...
corniformes	est_inclus	signes en T	3	F:\Corp...
corniformes	est_inclus	attelages	3	F:\Corp...
gravures rupestres	est_une_partie_de	corniformes	2	F:\Corp...
gravures rupestres	est_une_partie_de	armes et outils	2	F:\Corp...
gravures rupestres	est_une_partie_de	figures anthropomorphes	2	F:\Corp...
gravures rupestres	est_une_partie_de	figures géométriques	2	F:\Corp...
gravures rupestres	est_une_partie_de	figures non représentatives	2	F:\Corp...
figures géométriques	est_inclus	cercles	9	F:\Corp...
figures anthropomorphes	est_inclus	corniformes anthropomorphisés	7	F:\Corp...
corniformes	est_inclus	corniformes simples	3	F:\Corp...
figures non représentatives	est_inclus	barres	11	F:\Corp...
figures non représentatives	est_inclus	plages	11	F:\Corp...
figures non représentatives	est_inclus	cupules isolées	11	F:\Corp...
figures non représentatives	est_inclus	groupes de cupules isolées	11	F:\Corp...
figures non représentatives	est_inclus	figures inclassables	11	F:\Corp...

Figure 34. IHM de l'agent spécialisé SEEK Java [LEP 00]

3.4. En guise de synthèse

La plate-forme Filtext a été conçue pour accueillir le modèle conceptuel de tout système fondé sur la méthode d'exploration contextuelle indépendamment de tout choix informatique. C'est donc l'aboutissement de travaux de recherches engagés bien avant mon arrivée dans l'équipe LaLIC. Il faut insister sur l'important travail d'études linguistiques qui a permis de dégager les concepts d'indicateurs, d'indices et de règles d'exploration contextuelle. Dans une deuxième phase dans laquelle mes responsabilités sont plus importantes, les recherches se sont focalisées sur la réalisation de la plate-forme ContextO, réalisation qui par son exigence d'effectivité nous a amenés à questionner le modèle conceptuel défini dans Filtext. C'est ce dialogue constant, fécond, mais aussi quelquefois conflictuel, dans lequel interviennent linguistes et informaticiens qui stimule tous les participants à ce projet.

Chapitre 4

Filtrage sémantique de textes en Sciences Humaines

4.1. Introduction

Comme l'illustrent les pages précédentes, le filtrage sémantique d'informations ne doit pas être conçu comme l'utilisation passive de logiciels figés. L'analyse des premiers résultats de programmes de résumé automatique a montré qu'il fallait de plus en plus tenir compte des besoins des utilisateurs des résumés. Autrement dit, on résume toujours un texte pour répondre à des attentes spécifiques : identifier des informations innovantes, mettre en évidence des résultats ou des hypothèses dans un article scientifique ou dans un rapport technique, retrouver les déclarations d'un auteur sur une question controversée, rechercher les causes d'un phénomène, extraire les approches définitoires d'un concept, etc. Il est donc nécessaire d'introduire la notion de « point de vue » de l'utilisateur et effectuer le résumé en tenant compte explicitement de ses attentes, ce qui suppose qu'on lui donne, dans un système interactif Homme-Machine, la possibilité d'exprimer celles-ci. C'est ainsi que certaines recherches sur l'automatisation du résumé se sont orientées vers un filtrage sémantique qui synthétise des informations extraites des textes, ce filtrage étant guidé par un ou plusieurs points de vue.

Comment introduire et traiter les points de vue de l'utilisateur sans recourir à un hypothétique « modèle (cognitif) de l'utilisateur », avec le risque de ne pas pouvoir entreprendre, du moins à moyen terme, de développements finalisés ? Comment tenir compte des indications sémantiques fournies par l'utilisateur pour la recherche des informations susceptibles de l'intéresser ? Le recours à des requêtes sous forme de simples combinaisons booléennes de mots clés, comme dans les systèmes documentaires ou les systèmes de recherche d'information, apparaît comme peu adéquat car, exprimées de cette manière, les requêtes conduisent à des réponses trop bruyantes et, finalement, peu pertinentes. Les travaux dans ce domaine font actuellement apparaître que l'on peut fonder l'activité

résumante sur l'identification dans les textes de marqueurs linguistiques spécifiques qui correspondent à des points de vue et attentes d'un utilisateur potentiel sur l'organisation même du texte, et sur la position de certaines informations dans le texte.

Ainsi envisagée, la problématique du filtrage d'informations s'étend à d'autres disciplines des sciences humaines dans lesquelles la recherche d'informations dans des textes constitue une activité essentielle des chercheurs. Une coopération[39] scientifique menée avec Sophie Duchesne, chercheur au Centre d'Etude de la VIe POlitique Française (CEVIPOF) va nous permettre d'illustrer notre propos.

Le groupe de chercheurs du CEVIPOF développe des analyses sur la politisation de la parole en tant qu'enjeu pour une société démocratique. Comme S. Duchesne et F.Haegel [DUC 01] le montrent dans leur article, le rapport des citoyens au politique peut être envisagé selon au moins deux approches :

- Dans la première approche, qui s'inscrit dans la tradition des politologues américains, le rapport au politique a été interprété sous l'angle de la capacité à manipuler des connaissances spécifiques, celles du champ politique institutionnel ;

- Dans la deuxième approche, née d'une interrogation des chercheurs sur une autre forme de rapport au politique face à une tendance marquée à la dépolitisation des jeunes par rapport à la politique institutionnelle, il est d'usage de distinguer au sein des productions symboliques - à la suite des travaux de J-C.Passeron et C.Grignon - celles qui mettent en avant les effets de domination légitime (capacité et légitimité à manier les concepts de politique institutionnelle par certains acteurs sociaux) des effets de résistance et d'autonomie populiste (capacité et volonté de produire du discours politique en dehors de la sphère de la politique institutionnelle par ces mêmes acteurs sociaux).

Alors que le premier type de production symbolique a fait l'objet d'une abondante littérature, les chercheurs du CEVIPOF se sont penchés sur le second type de rapport au politique. Il s'agit de savoir comment est définie la politisation de la parole dite « populiste ». Pour S. Duchesne et F. Haegel, il faut « concevoir la politisation de la parole [populiste] comme le passage vers une situation de communication caractérisée par le fait qu'un au moins des interlocuteurs prend position sur un clivage, c'est-à-dire sur une opposition qui les dépasse au sens où elle ne se limite pas à ce qu'ils sont personnellement (l'un pour l'autre). ». A partir de cette conception, les deux auteurs émettent l'hypothèse qu'il est possible de distinguer dans des interviews les passages qui relèvent :

- du récit anecdotique ;
- de la montée en généralité ;

[39] Dans le cadre du mémoire de DEA *Mathématiques et Informatique Appliqués aux Sciences Humaines* de O. Guiraudie co-encadré par S. Duchesne et moi-même.

- de l'expression d'un clivage.

L'étape suivante vise à repérer la prise de position de l'interviewé dans le clivage, c'est-à-dire le passage à une situation politisée, en identifiant les sous-catégories suivantes :

- « Naturalisation » du clivage ;
- désignation d'un responsable ;
- identification à un groupe impliqués ;
- demande d'arbitrage ;
- foi en la capacité d'agir du groupe de référence ;
- foi en la capacité d'agir individuellement ;
- lien avec des acteurs politiques intervenus sur ce clivage.

Voici un exemple d'interview qui illustre certaines de ces catégories[40] (celles-ci ont été ajoutées manuellement) :

Mais il faut aussi voir ce qui se passe à l' intérieur ! Comment on peut essayer - parce que pour moi, moi tout ça au niveau international c'est un peu du cinéma. On dit " oui, il faut aider la Yougoslavie, il faut ... " On fait rien hein ! de toute façon. Ca ça prouve aussi le ... voyez je vous dis, l' éclatement des valeurs. Alors au niveau, au niveau international, on essaie de montrer qu'on est un pays social ; socialisé. qu' on a le, qu' on a le sens de la solidarité, et qu'on veut aider la Yougoslavie ou la Somalie ou etc. Mais en fait on fait rien. C'est du cinéma. Et, on le voit à l' intérieur. On fait ... on fait même pas un effort pour nos, nos propres frères qui sont en train de, de mourir dans la rue et de faire la manche. **Donc, il y a un problème c' est, c'est un peu n'importe quoi**. (*montée en généralité)* Si vraiment on, on avait ce sens de la solidarité, pourquoi aujourd'hui on a rien fait pour la Yougoslavie ? Pourquoi aujourd'hui on, on aide pas encore les gens du Rwanda ou encore plus ? Je sais pas ! On veut essayer de faire quelque chose mais c' est-à mon avis c'est juste du, c'est de la façade tout ça. C'est de la façade. **C'est une ... c'est une société qui a un masque**. (*montée en généralité)* Qui se croit encore une société. Mais qui perd beaucoup de ses valeurs. Ca perd beaucoup beaucoup. On le voit il y a ... il y a beaucoup d' exemples hein. **Il suffit d'ouvrir les yeux et de, et de voir la société telle qu' elle est hein. (***montée en généralité)*. Vous prenez le métro, vous allez balader dans les rues, tout, vous voyez. **Combien que les gens ont changé. (***montée en généralité***) Il y a plus cette idée de solidarité.** (*clivage)*. Parce que, je vous l'ai dit, moi pour moi, la cause principale, c'est le manque de communication. Entre tous les membres de la société**. Les gens ont perdu leurs valeurs, ils ont plus de quoi communiquer.** (*montée en généralité)* Ils ont plus (inaud.) Chacun travaille pour soi et puis voilà. Terminé [Hca38]. Mais moi, bon, je me pose la question hein, je sais pas, vous êtes spécialisé en société (rit), en sociologie. Est-ce que c' est vraiment une société ? C'est vraiment les conditions d' une société ça ? Le fait que chacun vit, vit de son côté ? **La définition de la société c'est groupe.** (*montée en généralité)* C'est pas individuel. C'est ça le problème. Mais bon**. Moi je vous dis en tant que jeune il faut pas, il faut pas se laisser aller et puis, espérons que plus tard, je vous ai dit, quand, quand nous on sera au pouvoir, il y aura plus de SDF (rit).** (*identification)* Ca c'est une grande préoccupation. *Entretiens 4000, François. Stéphane C. Le 1 / 2 / 95.*

On conçoit que le dépouillement de ces entretiens, qui représentent plusieurs dizaine de pages, est une tâche très lourde. L'hypothèse que nous avons posée avec les chercheurs du CEVIPOF est que le repérage de ces catégories (*montée en généralité, identification, clivage*, etc.) pouvait s'appuyer en

partie sur des marques linguistiques et qu'il conviendrait ensuite de réfléchir sur les outils de navigation textuelle nécessaires aux besoins spécifiques de cette recherche, comme par exemple vérifier qu'un *clivage* est toujours encadré par une *montée en généralité*. Il faut préciser, comme on peut le constater en lisant le texte complet de l'entretien [GUI 01], que la syntaxe est souvent malmenée et que les réponses des interviewés sont fréquemment constituées de phrases tronquées. On peut remarquer également que contrairement aux textes écrits, l'oralité des propos entraîne des variations importantes du contenu sémantique d'une phrase à l'autre. Par exemple tel enquêté parlera d'abord de ses difficultés quotidiennes à vivre son chômage, puis fera « *monter en généralité* » son propos en parlant de la condition des chômeurs en France, pour reprendre ensuite une anecdote personnelle explicitant ses propos. Et on observe ce phénomène de retour à un récit anecdotique pour toutes les catégories de la grille.

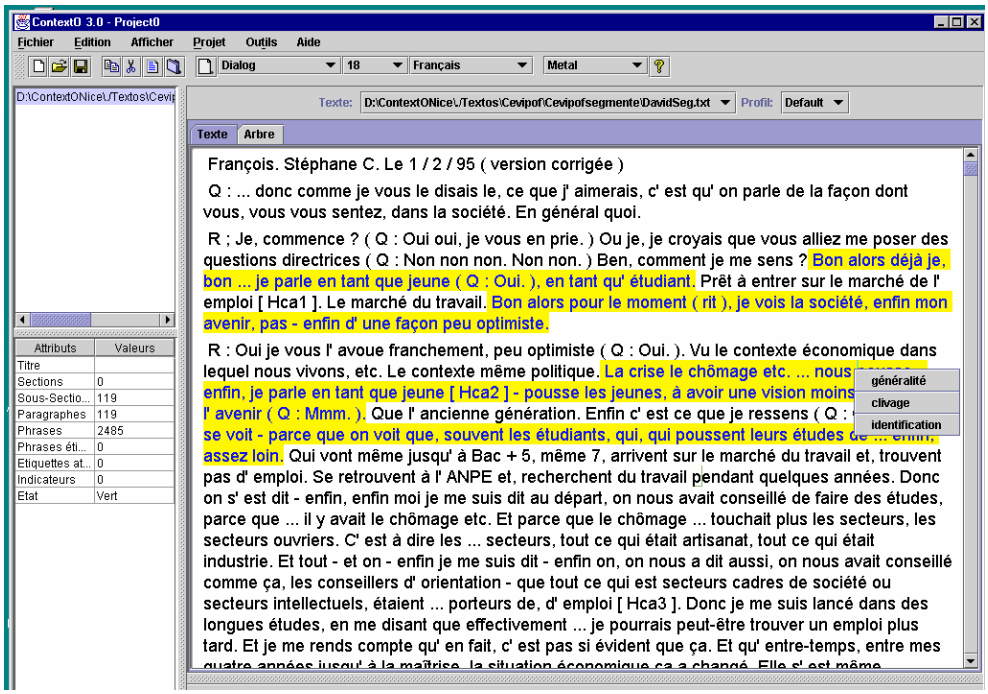


Figure 35. Repérage de la politisation de la parole [GUI 01]

Un outil de filtrage de texte comme ContextO fournit précisément aux chercheurs du CEVIPOF, et plus généralement aux sociologues qui dépouillent des textes, un moyen de repérer rapidement des séquences textuelles qui caractérisent les catégories recherchées. Ce repérage se fonde sur l'hypothèse qu'il existe des marques linguistiques qui reflètent ces catégories. La figure 35 illustre ce type d'identification effectuée automatiquement par ContextO. Les phrases repérées sont surlignées (avec des couleurs différentes sur l'écran) et l'utilisateur peut demander l'affichage des catégories identifiées

[40] Les phrases identifiées sont en **gras** suivies du nom de la catégorie placée entre parenthèses et en *italique souligné*.

(*généralité, clivage et identification*). En fait, comme le note O. Guiraudie dans son mémoire [GUI 01], l'intérêt de ce type d'outils est multiple : Il permet d'abord de valider des hypothèses de travail, il constitue ensuite un outil d'aide au dépouillement et enfin il ouvre de nouvelles perspectives de traitement comme le repérage des interactions entre l'enquêteur et l'enquêté.

Cet exemple souligne un fois de plus que l'activité consistant à sélectionner des informations à l'intérieur d'un texte donné est une démarche hautement intelligente qui varie en fonction des utilisateurs, de leur expertise du domaine traité dans le texte ainsi que de leurs besoins spécifiques. En conséquence, cette activité sollicitant l'intelligence générale des utilisateurs et leurs capacités cognitives n'est pas modélisable dans sa globalité. C'est pourquoi nos recherches dans ce domaine se sont orientées selon trois axes :

- construire des modèles et des outils de représentation d'un texte ;
- élaborer une méthode et les outils nécessaires à l'acquisition des connaissances linguistiques qu'il convient de mobiliser pour identifier les séquences textuelles pertinentes pour la tâche visée ;
- développer les inter-actions entre l'utilisateur et le système en construisant des outils de navigation textuelle.

Nous allons expliciter la problématique et esquisser nos propositions pour chacun de ces axes.

4.2. Pour une représentation de l'organisation textuelle

La construction d'un modèle de représentation d'un texte constitue l'étape préalable à tout traitement textuel dans les sciences humaines. Jusqu'à ces dernières années, les linguistes, essentiellement pour des raisons théoriques, s'étaient tenus à distance de cette problématique et les propositions sont plutôt issues des disciplines de l'analyse du discours [MAN 88] ou de la psycholinguistique [RUM 75]. Ce constat a été maintes fois souligné dans les années précédentes par différents linguistes comme en témoignent les extraits suivants :

> *« Aussi vaut-il mieux, ici comme ailleurs, se détacher de ces phrases isolées de tout contexte, que l'on va jusqu'à fabriquer soi-même. Réfléchissons plutôt sur ce qu'est véritablement un texte. C'est manifestement une totalité où chaque élément entretient avec les autres des relations d'interdépendance. Ces éléments et groupes d'éléments se suivent en ordre cohérent et consistant, chaque segment textuel compris contribuant à l'intelligibilité de celui qui suit. Ce dernier, à son tour, une fois décodé, vient éclairer rétrospectivement le précédent : on constate, si l'on s'y reporte, que la*

compréhension s'en est encore enrichie. Ainsi procédons-nous pour comprendre un texte ; toute phrase (et peu importe ce que l'on entend exactement par là) est subordonnée à chacune des autres dans la mesure où elle n'est pas simplement déchiffrée en elle-même mais participe à la compréhension de l'ensemble des autres. Cela prouve seulement la solidarité de tous les éléments dans ce réseau de déterminations qu'est un texte. » Weinrich H*, La grammaire du texte 1974*

« Jusqu'à présent, les linguistes n'ont guère progressé dans cette étude et sont, pour l'essentiel, restés dans les limites de la phrase. Car l'Analyse du Discours, sans être en soi un domaine vierge, l'est au moins du point de vue technique, en ce sens qu'aucune de ses parties fondamentales n' a encore été sérieusement pénétrée. Certes, il y a l'ouvrage bien connu de Harris, Discourse Analysis Reprints (1963), mais son objet réel, les réarrangements structurels au niveau de la phrase, le rend tout à fait étranger aux problèmes qui nous intéressent ici. » Labov W.*, Le parler ordinaire, Vol1, Editions de Minuit, 1978.*

« Le texte écrit nous force, de façon exemplaire, à comprendre que l'on ne peut pas passer de la phrase (hors prosodie, hors contexte, hors situation) à l'énoncé, par une procédure d'extension. Il s'agit en fait d'une rupture théorique, aux conséquences incontournables. » Culioli A.*, La langue au ras du texte, Préface aux Editions Universitaires de Lille, 1984 .*

« On sait que le texte n'a guère été étudié pour lui-même par les linguistes. C'est la phrase que les théories linguistiques ont, pour la plupart, adoptée comme unité d'analyse, et les études menées sur l'ambiguïté et la paraphrase reflètent très largement cette tendance : on s'y intéresse aux ambiguïtés de phrases isolées et aux relations de synonymie entre phrases prises deux à deux, sans considération de contexte plus vaste. » Fuchs C.*, Aspects de l'ambiguïté et de la paraphrase dans les langues naturelles Edition Peter Lang, 1985.*

De son côté, l'informatique, sous la pression des besoins du marché, notamment de celui de l'édition et dernièrement de celui du Web, a proposé des modèles de description qui se limitent à la prise en compte d'informations structurelles, comme la section, le paragraphe, le titre, la phrase, etc. Ces travaux ont ainsi donné lieu à la production de langages de description comme SGML puis XML et de leurs analyseurs associés (SAX, DOM) [XML 98] ; un important travail de standardisation est actuellement en cours, en vue de proposer des standards de description [BON 00]. Remarquons, néanmoins, que ces langages présentent une limitation importante dans leur pouvoir d'expression puisqu'ils ne permettent de décrire que des structures emboîtées et qu'ils restent très marqués par leur origine, à savoir les pratiques éditoriales. Les questions essentielles, qui concernent autant les

linguistes que les informaticiens sont : que décrire dans un texte ? Quelles organisations textuelles faut-il décrire ? Existe-t-il des unités textuelles élémentaires, et comment celles-ci sont elles organisées ? Comme le fait remarquer M.P. Péry-Woodley [PER 01] à la suite de J. Virbel [VIR 85], « les actes textuels peuvent soit être réalisés par la présence dans le texte de performatifs, par exemple, *j'organise le chapitre 1 en trois parties*, soit être inférables à partir de traces de l'effacement du performatif, traces qui constituent la mise en forme matérielle et recouvrent des marques lexico-syntaxiques, typographiques, dispositionnnelles et ponctuationnelles (par exemple blancs verticaux, titres et numération pour l'organisation d'un chapitre en trois parties). »

Nous pensons, en nous fondant sur l'expérience accumulée dans le filtrage d'informations et au regard des divers modèles proposés par la communauté, que les représentations textuelles à construire doivent l'être dynamiquement en fonction des finalités de la tâche. En effet, il existe potentiellement une multiplicité d'organisations textuelles (cf. 4.3).

Conceptuellement, la représentation d'un texte sur laquelle nous travaillons est, au départ, analogue à celle utilisée dans le modèle d'« hypertexte ». Rappelons que dans ce modèle, un texte est considéré comme un graphe ; chaque nœud du graphe représente une unité textuelle U et deux nœuds U_1 et U_2 peuvent être reliés par un lien hyper-texte. Les navigateurs (« browsers ») exploitent ces liens en offrant à l'utilisateur la possibilité de ce déplacer directement de U_1 vers U_2 (ou inversement) ainsi que deux opérations, « *avancer* » et « *reculer* ».

Nous proposons de remplacer cette notion de lien hyper-texte par celle de « composant ». Un composant encapsule[41] la description des propriétés du lien avec les opérations qu'il est possible d'effectuer sur ce type de lien.

Prenons l'exemple des énumérations qui sont introduites par une amorce puis marquées par des marqueurs d'intégration linéaire comme dans l'extrait suivant :

*Ce rebondissement s'explique par la convergence d'au moins trois facteurs****. En premier lieu****, ... **En second lieu**. ... **Enfin**, ...*

Les propriétés du composant décrivent sa structure, c'est-à-dire la phrase amorce marquée par «*Ce rebondissement.. »* et ses trois phrases introductrices marquées par *« En premier lieu », « En second lieu », « Enfin ».* Les opérations du composant sont « *Initial* », pour se positionner sur la phrase amorce, et « *avancer* » et « *reculer* », pour parcourir les éléments de l'énumération.

Notre deuxième exemple prend appui sur les travaux relatifs à la citation de G. Mourad [MOU 01]. Dans son travail, G. Mourad propose une classification des verbes de citation avec des classes comme « aveu », « dénonciation », « déclaration », etc. Dans ce cas, les unités textuelles sont les « citations »

[41] Dans la modélisation objet, l'encapsulation implique que les propriétés qui décrivent l'objet sont uniquement modifiables par les opérations définies pour cet objet.

et les opérations exploiteront cette typologie, de telle manière qu'il soit possible de se déplacer d'une « déclaration » à un « aveu », d'une « anticipation » à une « proposition ».

En résumé, dans notre modèle de représentation, une unité textuelle U_1 d'un texte est liée à une autre unité textuelle U_2 par un lien étiqueté $l_{U1U2}{}^{C}$ où C est un ensemble de composants logiciels muni de propriétés et d'opérations. Ces opérations peuvent viser au repérage et à la gestion de structures physiques (paragraphes, titres, etc.) ou discursives (énumérations, cadres temporels, cadres thématiques, etc.). Un texte est ainsi représenté comme un graphe dont les arcs sont « décorés » par des composants. Pour éviter une explosion combinatoire du nombre de liens à représenter, nous proposons de ne décorer le graphe qu'avec un certain type de composants. En effet, s'il est pertinent d'utiliser les composants qui gèrent les introducteurs de cadres thématiques (*cf.* 4.3) pour la tâche de résumé automatique, ceux-ci ne présentent pas d'intérêt dans le cas du repérage de la politisation de la parole, présenté précédemment.

On peut alors en s'inspirant des technologies de la modélisation objet, construire des composants abstraits d'analyse textuelle[42] qui seront ensuite instanciés et enrichis par des applications particulières. Les travaux de S. Ben Hazez [BEN 02] constituent à ce titre une expérimentation intéressante.

4.3. Acquisition et capitalisation des ressources linguistiques

La méthode d'exploration contextuelle constitue le socle sur lequel s'appuient nos travaux sur l'acquisition et la représentation des connaissances linguistiques.. Nous nous distinguons ainsi d'un courant de recherche, représenté notamment par [JAC 01], qui cherche à acquérir automatiquement des connaissances linguistiques par le repérage de collocations ou de patrons morpho-syntaxiques[43]. Nous nous appuyons tout au contraire sur le savoir des linguistes et sur le résultat de leurs analyses. Remarquons que de nombreux travaux, indépendamment de toute référence à l'exploration contextuelle, vont aussi dans ce sens. C'est le cas par exemple de C. Rossari qui déclare :

> *« Il ne s'agit pas de construire une théorie sémantico-pragmatique permettant de rendre compte du sens des connecteurs, ni d'en rendre compte dans le cadre d'une théorie existante, mais de construire des outils qui permettent de saisir les répercussions que les connecteurs ont sur les relations de discours en se basant exclusivement sur les contraintes qu'ils exercent sur les suites linguistiques qui leur sont adjacentes ». [ROS 00] p. 35.*

[42] Dans le sens que lui donne le génie logiciel dans la modélisation objet.
[43] Ces outils peuvent par contre fournir au linguiste des moyens de systématiser le repérage des configurations de marqueurs recherchés. Ils constitueraient ainsi un maillon dans la chaîne de traitement linguistique que nous cherchons à construire

La principale difficulté réside alors dans la définition d'un modèle de représentation des connaissances linguistiques et dans la construction d'un langage de manipulation de ces connaissances. Les travaux menés en syntaxe depuis plusieurs dizaine d'années [ABE 00] et les différents formalismes proposés dans ce domaine montrent l'ampleur et la difficulté de la tâche, d'autant plus que le passage de l'analyse de la phrase à l'analyse du texte introduit plus qu'une dimension supplémentaire dans le modèle de représentation à construire. Avec l'exploration contextuelle, nous pensons disposer d'un atout, au sens où nous ne cherchons pas à construire un modèle théorique mais à modéliser des observables et à rendre cette modélisation accessible. Ce travail d'ingénierie linguistique se situe à la frontière de trois disciplines : la linguistique, la cognition et l'informatique. Au plan cognitif, il nous faut mettre en place des protocoles expérimentaux qui valident, ou non, les stratégies d'identification qui sont mises en œuvre par des linguistes, tant dans la phase d'acquisition que dans la phase d'organisation des données textuelles. Sur le plan linguistique, le principe même de la contextualité de la méthode d'exploration contextuelle tend à restreindre les potentialités de réutilisabilité des ressources linguistiques construites pour une tâche spécifique. En clair, il nous faut découvrir les opérations linguistiques ou discursives qui se cachent derrière les notions de classes et règles d'exploration contextuelle, tout en sachant que ces opérations ne sont pas indépendantes des organisations textuelles. Enfin, sur le plan informatique, il faut pouvoir inscrire ces concepts dans des langages et des composants logiciels.

4.4. Pour des outils de navigation textuelle

Comme nous l'avons dit précédemment, nous ne disposons pas actuellement des concepts et des outils nécessaires à la construction d'un modèle cognitif de l'utilisateur, et rien ne laisse présager que cette situation puisse changer à court terme. Or les deux exemples présentés ci-dessus montrent que la gestion des inter-actions entre l'utilisateur et le système, qui se limitent aux opérations de navigation dans un texte, est une réponse possible. Moins ambitieuse conceptuellement, cette approche se heurte tout de même à de sérieuses difficultés qui sont liées d'ailleurs au problème de la représentation d'un texte et à celui de la modélisation des connaissances linguistiques que nous venons d'évoquer .

Quand un système de résumé automatique affiche un extrait de textes construit à partir de phrases extraites du texte source, la perte d'information dans les domaines de l'organisation thématique du texte, de l'argumentation, de la prise en charge, *etc.*, est considérable[44]. Il faut donc compenser cette perte d'information textuelle par des présentations iconiques (comme celles décrites au § 3.3.2), mais également en offrant au lecteur des stratégies de parcours du texte qui se fondent sur le repérage de certaines relations organisations textuelles et sur l'identification de notions sémantiques. Il faut pallier l'absence de compréhension par l'exploitation des données structurelles et discursives.

[44] On peut néanmoins remarquer qu'il en est de même pour les résumés rédigés par des professionnels.

Nous allons illustrer notre propos en nous appuyant sur deux exemples issus des expérimentations effectuées avec le système ContextO. Le premier exemple nous est fourni par un article scientifique de la revue « Pour la Science » (texte n° 3 de l'annexe) intitulé « Pivée d'émotions, la mémoire flanche ». Le repérage des marqueurs d'intégration linéaire[45] permet dans un premier temps de construire un extrait de texte plus cohérent en plaçant dans celui-ci toutes les phrases qui sont liées, comme le montre l'extrait de texte produit par ContextO.

[...] ***En premier lieu****, l'essor des neurosciences cognitives, tout au long du XXe siècle, a considérablement accru notre savoir sur le cerveau, fournissant ainsi les bases indispensables pour aborder la complexité des phénomènes affectifs.* ***En second lieu****, des perspectives entièrement nouvelles ont émergé grâce à de récents progrès techniques.* [...]_***Enfin****, plusieurs chercheurs contemporains, ouvrant la voie des neurosciences affectives, ont su réactualiser l'idée ancienne selon laquelle les émotions sont en réalité la cheville ouvrière du bon fonctionnement de nombre de nos facultés, adaptation sociale, raisonnement, prise de décision, ou mémoire.*

Dans un deuxième temps, un outil de navigation textuelle permet à l'utilisateur de visualiser rapidement la phrase amorce de cette énumération et de proposer l'extrait suivant [46]:

*Ce rebondissement s'explique par la convergence d'au moins trois facteurs****. En premier lieu****, l'essor des neurosciences cognitives, tout au long du XXe siècle, a considérablement accru notre savoir sur le cerveau, fournissant ainsi les bases indispensables pour aborder la complexité des phénomènes affectifs.* ***En second lieu****, des perspectives entièrement nouvelles ont émergé grâce à de récents progrès techniques.* [...] ***Enfin****, plusieurs chercheurs contemporains, ouvrant la voie des neurosciences affectives, ont su réactualiser l'idée ancienne selon laquelle les émotions sont en réalité la cheville ouvrière du bon fonctionnement de nombre de nos facultés, adaptation sociale, raisonnement, prise de décision, ou mémoire.*

Le repérage dans l'amorce du marqueur anaphorique « ce » apporte une nouvelle possibilité de navigation qui se traduit par la visualisation de la phrase d'un nouvel extrait :

*Les émotions sont aujourd'hui l'objet d'un intérêt grandissant en neurosciences, comme en témoigne la croissance exponentielle des publications dans ce domaine depuis la fin des années 1990. Ce rebondissement s'explique par la convergence d'au moins trois facteurs****. En premier lieu****, l'essor des neurosciences cognitives, tout au long du XXe siècle, a considérablement accru notre savoir sur le cerveau, fournissant ainsi les bases indispensables pour aborder la complexité des phénomènes affectifs.* ***En second lieu****, des perspectives entièrement nouvelles ont émergé grâce à de récents progrès techniques.* [...] ***Enfin****, plusieurs chercheurs contemporains, ouvrant la voie des neurosciences affectives, ont su réactualiser l'idée ancienne selon laquelle les émotions sont en réalité la cheville ouvrière du bon fonctionnement de nombre de nos facultés, adaptation sociale, raisonnement, prise de décision, ou mémoire.*.

Comme on le voit, l'exploitation de ces types d'organisations textuelles, très fréquentes dans les articles scientifiques, ouvre la voie à des possibilités nouvelles dans l'étude des textes.

Le deuxième exemple montre que les possibilités de parcours dynamique sont en fait multiples. Dans le texte présenté au paragraphe *5.2.6.3,* on peut faire l'hypothèse que la phrase suivante :

[45] Ecrit en caravtère gras pour illustrer notre propos.

[46] En se fondant soit sur le repérage de marqueurs linguistiques comme « trois facteurs », qui peut être rapproché du nombre de marqueurs d'intégration linéaire employé par le rédacteur, soit sur une heuristique qui exploite la position de cette phrase, car la phrase amorce est généralement la phrase qui précède le premier élément de l'énumération.

(1) Cette situation a des conséquences décisives, particulièrement sur trois variables stratégiques du processus de développement : l'allocation des ressources, le modèle de consommation et l'intégration en amont de l'activité industrielle.

est placée dans l'extrait présenté au lecteur[47]. L'identification des introducteurs de cadre thématique [CHA 97] associés à un repérage des groupes nominaux permettrait d'offrir à l'utilisateur les moyens d'afficher l'extrait suivant :

Cette situation a des conséquences décisives, particulièrement sur trois variables stratégiques du processus de développement : l'allocation des ressources, le modèle de consommation et l'intégration en amont de l'activité industrielle.

***En ce qui concerne** l'allocation des ressources, en Côte-d'Ivoire, l'excédent prélevé par l'Etat et consacré à l'expansion du marché intérieur transite nécessairement par les firmes multinationales, finançant en grande partie leur implantation ou l'élargissement de leur capacité productive.* [...]

***QUANT au** modèle de consommation, en Côte-d'Ivoire, la production de biens relève de la stratégie propre à la firme multinationale, sans rapport avec le niveau moyen des revenus et les habitudes traditionnelles de consommation.* [...]

*Enfin, **pour ce qui concerne** l'intégration en amont de l'activité industrielle, dans un pays comme la Côte-d'Ivoire, où le secteur industriel est contrôlé par les firmes étrangères, la taille du marché a constitué l'obstacle insurmontable à la diversification de la structure productive.* [...]

Mais la lecture du texte source montre bien la part d'arbitraire qu'il y a, pour un lecteur donné, dans cette sélection ; un autre lecteur pourrait vouloir plutôt rechercher les présentations contrastées que fait le rédacteur en utilisant les introducteurs de cadre spatiaux que sont *en Corée du Sud* et *en Côte-d'Ivoire*, pour obtenir ce type d'extrait[48] :

***En Côte-d'Ivoire,** les firmes contrôlent pratiquement l'ensemble de l'industrie produisant pour le marché interne. Au contraire, l'accès à ce dernier leur est interdit dans la plupart des branches **en Corée du Sud**.* [...]

*En ce qui concerne l'allocation des ressources, en **Côte-d'Ivoire**, l'excédent prélevé par l'Etat et consacré à l'expansion du marché intérieur transite nécessairement par les firmes multinationales, finançant en grande partie leur implantation ou l'élargissement de leur capacité productive. **En Corée du Sud**, les firmes multinationales sont exclusivement concentrées dans les branches exportatrices, ce qui permet à l'Etat de prélever des ressources externes additionnelles que les entreprises publiques ou privées coréennes utilisent selon les orientations précises du plan dans le cadre d'une stratégie d'intégration industrielle orientée vers le marché intérieur.*

*QUANT au modèle de consommation, **en Côte-d'Ivoire**, la production de biens relève de la stratégie propre à la firme multinationale, sans rapport avec le niveau moyen des revenus et les habitudes traditionnelles de consommation.*

[...] ***En Corée du Sud**, la diversification des biens offerts aux consommateurs est un processus progressif et contrôlé en relation étroite avec la capacité d'achat de la population.*

[...] ***En Corée du Sud**, la maîtrise absolue de l'Etat sur la décision économique au niveau du marché interne a permis ce que l'on appelle la "remontée des filières" vers les industries lourdes - sidérurgie, chimie et industries de biens d'équipement - et assuré une autonomie notable du processus d'industrialisation, même si, dans certains secteurs, la dimension du marché était manifestement insuffisante.*

---

[47] Compte tenu de la présence, dans une énumération, des marqueurs « conséquences décisives » et « variables stratégiques ».

[48] On remarquera que ce repérage échouera pour la phrase « Enfin, pour ce qui concerne l'intégration en amont de l'activité industrielle, *dans un pays comme la Côte-d'Ivoire*, où le secteur industriel ... ».

Il faut que ces fonctionnalités de parcours du texte soit dynamiquement proposées. Par exemple, après que le système ait extrait la phrase (1), c'est l'utilisateur qui doit pouvoir déclencher l'identification des groupes nominaux «*allocation des ressources », «modèle de consommation »* et *« intégration en amont de l'activité industrielle »* et la recherche, dans la suite du texte, des introducteurs de cadre qui introduisent ces groupes nominaux.

Les stratégies de parcours s'appuient sur les représentations textuelles présentées précédemment et sont dépendantes de la tâche ; dans l'exemple du repérage de la politisation de la parole, les stratégies à mettre en œuvre se focaliseront sur d'autres organisations discursives comme les enchaînements entre « clivage » et « montée en généralité ». Par ailleurs, les structures textuelles comme les paragraphes, les sections ou même les phrases n'ont aucune pertinence pour le sociologue qui cherche à explorer un entretien ; dans ce cas, il faut plutôt repérer les structures textuelles comme les questions et les réponses.

Il existe donc une profonde corrélation (cf. fig. 36) entre les stratégies de parcours et le modèle hypertexte construit. La modélisation des connaissances qui sont mises en jeu dans ces stratégies et leur inscription dans des outils d'interaction relèvent de la construction de ce que nous avons appelé (cf. 3.3.1) des « agents spécialisés ».

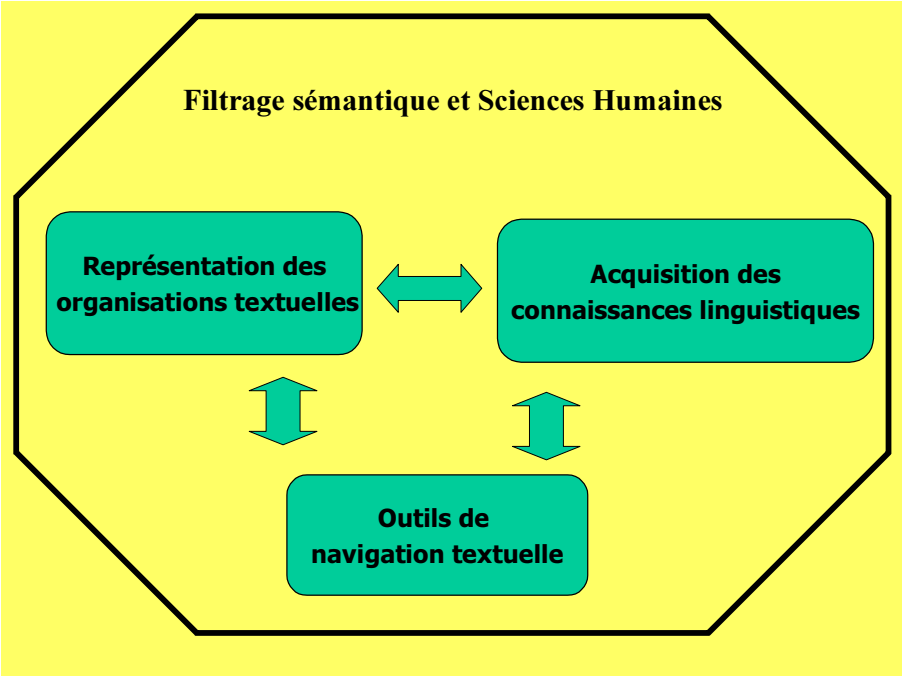


Figure 36. Projet de recherche pour le filtrage sémantique en Sciences Humaines

Jusqu'à présent les travaux sur la représentation textuelle [BAE 99]se sont surtout intéressés aux problèmes de la représentation de données agrégées et à l'élaboration de langages de requêtes spécialisés. Notre approche vise plutôt à étudier finement comment le repérage des organisations textuelles se combinent avec des stratégies de parcours spécifiques à une tâche, en vue de nous permettre de produire de nouveaux instruments d'études des textes. Bien que nous nous soyons appuyé sur des exemples de tâches dans le domaine de la linguistique et de la sociologie, nous pensons que d'autres disciplines des Sciences Humaines sont concernées.

Par cette démarche, à la fois exploratoire et effective, nous cherchons à illustrer le fait qu'un problème très pratique, résumer des documents textuels, peut conduire à entreprendre de nouvelles descriptions linguistiques, à initier des collaborations que nous espérons fécondes et à orienter la linguistique textuelle vers de nouvelles perspectives théoriques.

Chapitre 5

Utilisations de la plate-forme ContextO

5.1.Filtrage d'informations sur la Toile

5.1.1. Projet RAP

Les systèmes de résumé automatique ont été confrontés , avec la montée en puissance de la Toile, à un nouveau type de problème : la « surcharge informationnelle » ; une requête sur un thème donné peut se traduire par un accès potentiel à des milliers de textes. Dans l'optique de répondre à ces nouveaux défis, l'équipe LaLIC[49], en collaboration avec la société Pacte Novation et l'URFIST, à répondu à un appel d'offre du Ministère de l'Education nationale pour proposer un prototype, le système RAP [NAI 98, 99]. Ce système s'appuie pour une part sur l'agent spécialisé « Résumeur Filtreur » de la plate-forme ContextO pour présenter des résumés à un utilisateur.

[49] La réponse à cet appel d'offre a été rédigée en collaboration avec Jean-Pierre Desclés, Philippe Laublet et moi-même.

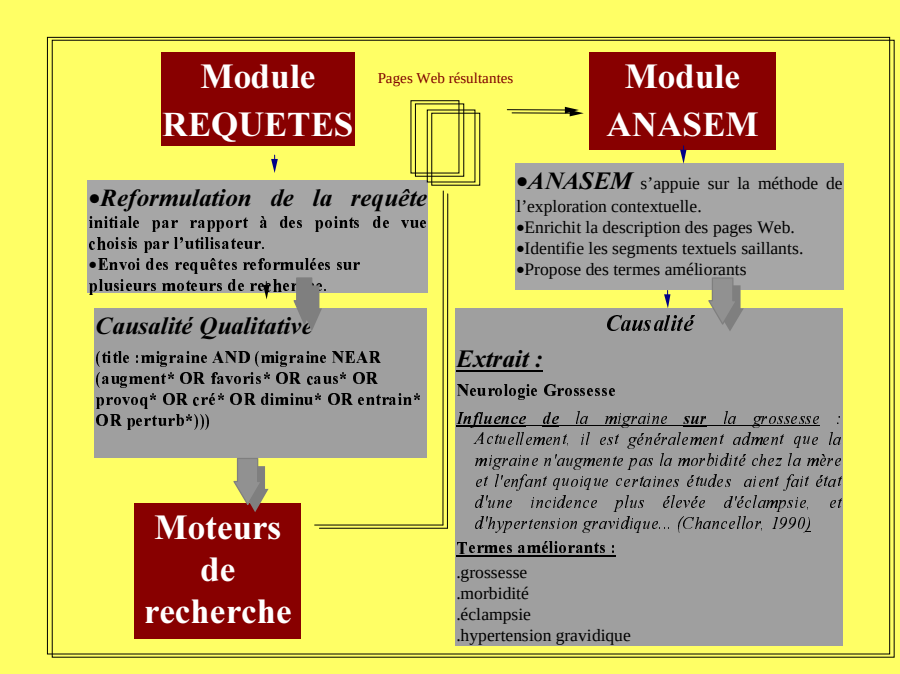


Figure 37 : Architecture du système RAP [NAI 98]

Le système RAP poursuivait deux objectifs :

- Réduire de manière drastique le nombre de documents fournis par les moteurs de recherche en réponse à une requête utilisateur ;
- Proposer pour chaque URL considérée comme pertinente, un extrait significatif du texte.

Pour répondre à ces objectifs tout en bénéficiant des capacités des outils existants, l'architecture du système RAP faisait collaborer quatre modules : un module de reformulation, un module de requête et un module de filtrage et un module de présentation des extraits de textes.

Module de reformulation

Ce module a pour finalité de reformuler la requête de l'utilisateur en s'appuyant sur la notion de « point de vue » [NAI 01]. Cette notion de point de vue permet à un utilisateur de combiner des critères de recherche sur le type de document avec des notions discursives comme « la causalité », « la définition », etc. La figure 38 illustre cette notion de point de vue concrétisée par des interfaces réalisées par B. Djioua.

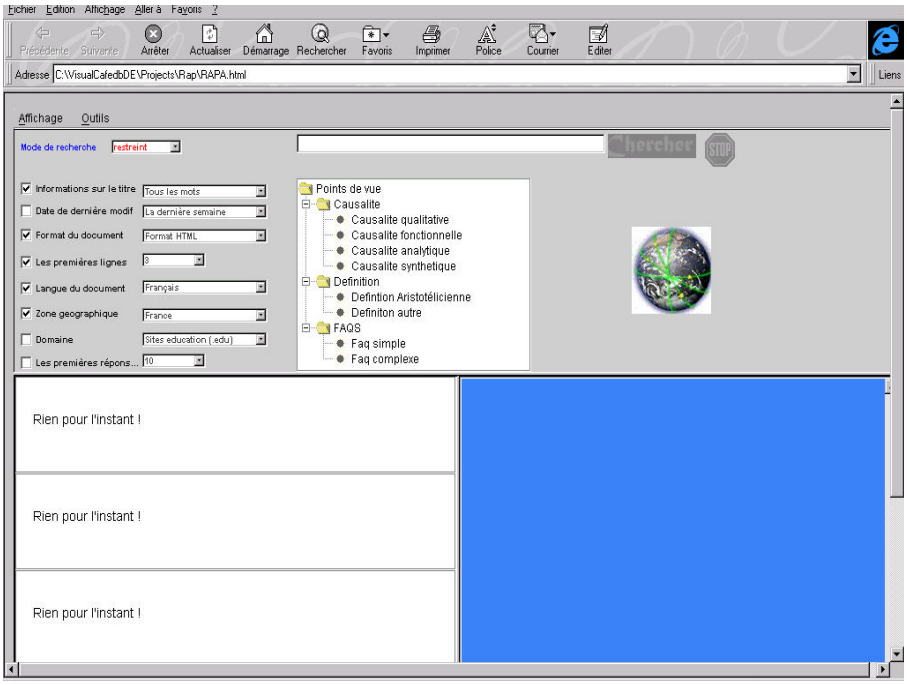


Figure 38 : Interface du module de reformulation [NAI 99]

Module de requête

Le module de requête transforme la requête utilisateur en une requête conforme aux exigences des langages utilisés dans les moteurs de recherche classique. La requête produite est donc une combinaison d'opérateurs booléens (and, or), d'opérateurs de distance (near, adj, etc.) et de termes linguistiques représentatifs du point de vue.

Module de filtrage

Le module de filtrage est s'appuie sur la méthode d'exploration contextuelle pour extraire les phrases considérées comme pertinentes. Dans un premier temps un module spécifique a été réalisé, puis dans un deuxième temps le moteur d'exploration contextuelle développé dans ContextO a été utilisé. Pour ce faire M. Parienti [PAR 99], dans le cadre de son mémoire de DEA, a réalisé un module afin de pouvoir traiter des textes au format HTML.

Module de présentation

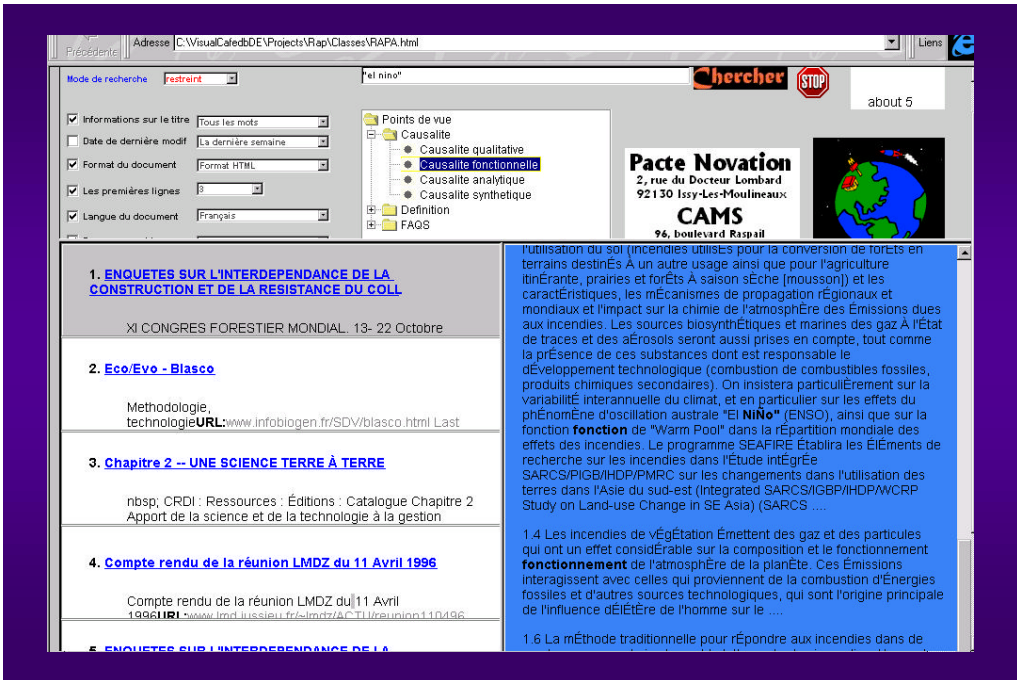


Figure 39 : Interface du module de présentation [NAI 99]

Ce module présente dans une première fenêtre (à gauche sur la figure 39) une liste des documents jugés pertinents et dans une deuxième fenêtre (à droite sur la figure 39) l'extrait significatif du texte. L'utilisateur dispose de diverses fonctionnalités pour naviguer entre les différents extraits.

5.1.2. Acquis du projet RAP

Les évaluations réalisées par l'URFIST sur le système RAP ont démontrées le bien fondée de l'approche tout en confirmant ce que nous avions constaté lors de l'évaluation du système SERAPHIN : l'absence de repérage des structures thématiques des textes est souvent un frein à la production d'extraits de textes significatifs pour l'utilisateur

5.2. Projet de « Modèle d'exploration sémantique de textes guidé par les points de vue du lecteur »

Mes propres recherches s'orientent actuellement vers l'intégration d'autre approches du traitement linguistique, l'objectif étant de capter des informations textuelles plus globales en vue de capter les éléments thématiques du texte. C'est dans ce cadre que je participe au projet intitulé « Modèle d'exploration sémantique de textes guidé par les points de vue du lecteur ». Ce projet, financé par le Ministère de la Recherche dans le cadre des ACI Cognitique 2000, réunit quatre partenaires : le LIMSI (UPR du CNRS), le LATTICE (UMR du CNRS), le CEA et l'équipe LaLIC du CAMS. Je reproduis partiellement ci-après le texte du projet rédigé par ces quatre partenaires ainsi que les premiers résultats obtenus [FER 01].

5.2.1. Situation du sujet et objectifs généraux

Les moteurs de recherche actuels, que ce soit dans des grandes bases de documents ou sur le Web, sélectionnent souvent un très grand nombre de documents en rapport avec une requête. Dans ce contexte, il est intéressant d'offrir des outils de visualisation rapide des textes sélectionnés afin que l'utilisateur puisse évaluer leur pertinence par rapport à sa requête.

Le projet soumis vise au développement d'un modèle d'analyse et de représentation de textes permettant de s'adapter aux points de vue d'un lecteur lorsqu'il consulte un document afin de lui fournir l'information qu'il juge pertinente. Cette problématique repose sur les techniques d'analyse nécessaires à la production de résumé, la dimension originale de notre projet consistant à produire un résumé « dynamique », fonction des désirs et des besoins d'un utilisateur, donnant ainsi la possibilité d'une exploration personnalisée d'un texte.

Le modèle envisagé doit pouvoir s'appliquer à tout type de texte, quel que soit le sujet traité, et fera collaborer une analyse thématique fondée sur la récurrence et la distribution des mots avec une analyse fine des marqueurs linguistiques, qu'ils soient de nature argumentative ou significatifs de liaisons et de

ruptures thématiques. Ce modèle sera implémenté dans la plate-forme Filtext du CAMS qui a déjà développé, avec ContextO, un système de résumé par extraction.

5.2.2. Cadre scientifique

Le modèle met en jeu deux idées essentielles qui expliquent les développements prévus dans le projet soumis. Ces deux idées sont les suivantes :

- le système de fouille et de filtrage doit être adapté aux besoins des utilisateurs et conçu de telle sorte que ceux-ci puissent l'enrichir en fonction des informations qu'il recherchent ;

- ce système, pour répondre aux attentes du maximum d'utilisateurs, doit reposer sur des indicateurs linguistiques indépendants des sujets abordés dans les textes traités, l'intégration de connaissances du domaine demeurant cependant possible, pour améliorer ses performances.

La conception et le développement d'un tel système supposent une collaboration étroite et de longue haleine entre linguistes et informaticiens. Les améliorations prévues dans le projet soumis vont dans le sens d'un approfondissement et d'un élargissement de cette collaboration. Ils portent sur deux points qui sont étroitement liés, à savoir : le repérage des unités thématiques et la segmentation des données textuelles.

En ce qui concerne le repérage des unités thématiques, il s'agira de faire collaborer des procédures de calcul prenant en compte des indicateurs lexicaux, à même de fournir très rapidement des indications sur le thème d'un segment de texte et les changements thématiques d'un segment à un autre, avec des marqueurs linguistiques porteurs d’informations quant au rôle des segments du point de vue argumentatif ou discursif. Ces procédures, déjà explorées dans les travaux de l'équipe L&C du LIMSI, seront affinées et intégrées dans ContextO.

En ce qui concerne la segmentation, la plupart des systèmes s'en tiennent au découpage en paragraphes qui sont des unités assez grossières et parfois peu motivée. Pour étoffer cette dimension, il est prévu d'intégrer dans le système les expressions introductrices de cadres de discours (groupes adverbiaux détachés en tête de phrases) qui signalent la façon dont un rédacteur répartit les informations dont il fait état dans des rubriques homogènes. L'analyse linguistique de ces marqueurs et les principes gouvernant la mise en place des cadres de discours seront prises en charge par la composante Paris III de l'équipe LATTICE.

La fouille et le filtrage sémantique d'informations textuelles sont des activités auxquelles les lecteurs se livrent en permanence. On ne dispose pas de donnés psychologiques précises sur ces activités. Il est probable que les connaissances du domaine jouent un rôle important dans ces activités ;

toutefois, on peut penser que les lecteurs s'appuient aussi pour les mener à bien sur des indicateurs linguistiques du genre de ceux pris en compte dans le système ContextO. Les contraintes liées à l'implémentation, en ce qu'elles obligent à expliciter le maximum de facteurs à même de peser sur une décision d'extraction, constituent une excellente base pour la formulation d'hypothèses neuro-psychologiques sur les démarches naturelles de fouille et de filtrage. Une des dimensions du projet soumis consistera à rechercher des partenaires en vue de l'opérationnalisation à court terme de ces hypothèses.

5.2.3 Objectifs scientifiques

Ce projet de recherche vise à identifier les activités cognitives mises en œuvre par un lecteur qui cherche à dégager les points importants d'un texte et les liens qu'ils entretiennent. Selon le point de vue adopté lors de la lecture, les intentions du lecteur et ses connaissances, ce qui est pertinent dans un texte varie. C'est ainsi que notre objectif consiste à proposer un modèle d'analyse de texte rendant compte de cette variabilité, reposant sur la détection de la structure thématique des textes. Une autre caractéristique importante dont il nous faut alors tenir compte concerne la fait que la structure thématique d'un texte varie selon le type de texte (article de journal, article scientifique, etc.). Il faut donc aussi tenir compte de cette variabilité dans l'élaboration d'un modèle général ou de modèles différents selon les types de textes.

Afin de répondre de manière fine aux demandes des utilisateurs du système, nous chercherons à articuler ce ou ces modèle(s) avec des identifications sémantiques plus locales. Notre projet visant à traiter des textes portant sur des domaines quelconques, cela introduit la contrainte de ne pas proposer de modèles reposant sur des connaissances difficiles à modéliser ou à acquérir, en particulier des modèles reposant sur des ontologies spécifiques des domaines.

Dans le domaine de la recherche d'information, la sélection d'un document et l'extraction d'informations pertinentes d'un texte adaptées aux besoins d'un utilisateur sont classiquement fondées sur la notion de profil d'utilisateur. Les profils d'utilisateurs, à l'image de la représentation des textes, prennent généralement la forme d'ensembles de termes, éventuellement pondérés. Le filtrage proprement dit est alors réalisé par l'application d'une mesure de similarité entre les profils et les textes considérés. La mise en évidence de la structure thématique des textes permet ainsi de complexifier la notion de profil afin d'atteindre une plus grande précision vis-à-vis des attentes de l'utilisateur. Il s'agit non seulement de rassembler un ensemble de termes relatifs à un thème général intéressant l'utilisateur mais de les structurer en sous-ensembles thématiquement cohérents et de granularité compatible avec le niveau d'analyse des textes visé. Une telle structuration devient en effet nécessaire dès lors qu'un profil a pour objectif de sélectionner non seulement des textes mais au-delà, des parties de texte articulées entre elles, comme c'est le cas dans l'optique du résumé dynamique.

Le matériau textuel sera constitué d'articles et de rapports techniques issus de domaines divers (brevets, énergie, sociologie, financier, etc.) ainsi que de textes obtenus à l'aide de moteur de recherche sur le WEB. Nous travaillerons aussi sur des articles de journaux provenant de la presse française (« Le Monde » par exemple).

En ce qui concerne la mise en œuvre, nous développerons un système informatique qui s'appuiera sur l'architecture du système FilText développé par le CAMS, et sur les composants d'analyse thématique développés au LIMSI. Nous construirons une base de modèles et les ressources linguistiques nécessaires au filtrage et au typage. Ceux-ci seront spécifiés et stockés à l'aide de formalismes et de composants logiciels qui les rendront facilement réutilisables par la communauté. En effet, nous pensons que l'architecture de la plate-forme Filtext, en privilégiant le concept de composants logiciels et d'agents spécialisés, la rend apte à accueillir différents types de traitement linguistique car il devient possible de construire de nouvelles bases de marqueurs linguistiques adaptés à de nouvelles tâches d'étiquetage sémantique. Cette plate-forme vise ainsi à faciliter les étapes d'acquisition et de modélisation des connaissances linguistiques en proposant des formats, et des langages de représentation des données, des outils de consultation, de manipulation, de recherche et d'analyse, etc.

5.2.4. Méthodologie

Le projet repose sur les résultats obtenus par les différents partenaires en ce qui concerne, d'une part, l'identification de la valeur sémantique de certains segments textuels et, d'autre part, la segmentation thématique des textes. Après un bref exposé de ces travaux montrant en quoi ils sont complémentaires, nous présentons comment nous les ferons collaborer afin d'élaborer un modèle général de la structuration thématique d'un texte.

5.2.4.1. Filtrage sémantique

Le système ContextO permet d'accéder au contenu sémantique des textes pour en extraire des séquences particulièrement pertinentes. A cet effet, la plate forme Filtext exploite des connaissances purement linguistiques. Les indicateurs sollicités sont constitués de marqueurs discursifs explicites (morphèmes, mots, expressions et locutions...) interprétés en contexte et considérés comme révélateurs d'une intention informationnelle de l'auteur du texte, intention que le système doit être capable d'identifier sémantiquement et pragmatiquement, à l'instar du lecteur peu averti du domaine traité.

La méthode d'exploration contextuelle développée dans l'équipe LALIC met en jeu des processus décisionnels [DES 97b] qui sont déclenchés, dans un premier temps, par l'identification d'indicateurs linguistiques relatifs à une tâche. Ces marqueurs étant généralement polysémiques, l'exploration contextuelle prévoit un ensemble de règles qui, pour un marqueur donné et une décision à prendre,

tiennent compte d'autres indices explicites dans un espace de recherche (proposition, phrase, paragraphe...), afin d'affecter une valeur sémantique à ce marqueur dans le contexte particulier où il apparaît.

D'une façon générale, une *base d'exploration contextuelle* se compose :

- d'un ensemble d'*indicateurs linguistiques* - c'est-à-dire de marqueurs de valeurs sémantiques (grammaticales ou discursives) - jugés pertinents pour la résolution de la tâche ;
- d'un ensemble de *règles d'exploration contextuelle* qui cherchent à reconnaître la présence *d'indices linguistiques* complémentaires et présents dans le contexte d'un marqueur ; ces règles orientent vers une prise de décision immédiate ou vers la recherche d'autres indices complémentaires plus fins.

Les règles d'exploration contextuelle sont des heuristiques, et non des règles impératives, et elles sont, rappelons-le, indépendantes des connaissances des domaines traités : leur application n'oblige pas à construire des *représentations des connaissances préalables* à l'analyse sémantique des textes.

Lors de la production de résumé et plus généralement de la fouille sémantique de textes, le système exploite directement l'organisation du texte source. Le jugement d'importance est fondé essentiellement sur ce que l'auteur a lui-même explicitement mis en valeur dans son texte à l'aide, par exemple, d'expressions comme : *voici ce qui doit être noté…, j'insiste sur…, …*. Dans cette perspective, il s'agit donc de sélectionner, par le biais de marqueurs linguistiques révélateurs, des séquences textuelles exprimant un certain point de vue de l'énonciateur.

La construction d'extraits par extraction de phrases issues du texte source est confrontée à plusieurs problèmes. Parmi les difficultés bien connues auxquelles conduit ce genre d'approche, on peut signaler :

- les risques de mauvaise interprétation ou d'impossibilité d'interprétation d'une anaphore,
- les risques de mauvaise interprétation ou d'impossibilité d'interprétation d'un connecteur qui sera considéré comme signalant une relation avec une phrase filtrée alors que dans le texte source le lien met en cause une autre phrase (par exemple, si une phrase, étiquetée comme *conclusive* est sélectionnée, il est impossible de savoir à quelle phrase, étiquetée comme *hypothèse* elle se réfère, puisque le système ne construit aucune représentation conceptuelle du texte source).

Dans l'état actuel des connaissances et des outils disponibles pour traiter correctement ce genre de problèmes, les concepteurs n'ont d'autres ressources que de recourir à des palliatifs. Ainsi dans ContextO, des outils de navigation permettent à l'utilisateur de revenir au texte source.

5.2.4.2. Analyse thématique

Le groupe L&C du LIMSI, par ailleurs, a développé des méthodes d'analyse thématique de textes, que ce soit dans le cadre du résumé ou de l'apprentissage. Il a plus particulièrement étudié le problème de la segmentation thématique, qui consiste à déterminer les endroits du texte où le sujet traité change. Les méthodes réalisées correspondent aux deux grandes approches du domaine : i) segmentation par calcul de distance entre vecteurs de descripteurs représentatifs de deux zones de textes contiguës ; ii) segmentation reposant sur l'étude des valeurs de cohésion associées aux descripteurs du texte.

La description de zones de texte par un vecteur de descripteurs repose sur la répétition et la répartition de mots significatifs pour le sujet traité. Une première méthode [MAS 95) consiste à mesurer une distance entre deux vecteurs représentant deux paragraphes consécutifs, et à décider d'une rupture thématique lorsque la distance passe un certain seuil. Cette méthode s'applique bien à des textes scientifiques issus de magazine de vulgarisation scientifique, comme « La Recherche » et « Pour la Science », mais donne de très mauvais résultats sur des articles de journaux. Afin d'appliquer la même approche sur des textes de type narratifs, le LIMSI a expérimenté l'utilisation d'une source de connaissance générale permettant d'introduire une mesure de la cohésion lexicale entre deux paragraphes [FER 98a]. Chaque paragraphe est alors décrit par ses propres termes mais aussi par les termes du texte qui leur sont particulièrement liés afin de lier deux paragraphes consécutifs utilisant des termes différents mais très liés sémantiquement. Ces valeurs de liaisons entre termes proviennent d'un réseau de cooccurrences construit automatiquement sur un corpus d'articles de journaux. Par ailleurs, afin de traiter des textes narratifs courts, tels que des dépêches, l'équipe du LIMSI a développé une méthode reposant entièrement sur la notion de cohésion lexicale [FER 98b]. A chaque mot du texte est associé une valeur de cohésion traduisant les liens qu'il entretient avec ses voisins, cette valeur étant calculée à partir des valeurs des liens trouvés dans le réseau de cooccurrences. Cela produit une courbe qui est analysée automatiquement afin de détecter des ruptures.

Ces deux approches ont l'avantage d'être applicables à tous les domaines traités, mais sont peu fiables quand aux décisions de segmentation.

5.2.4.3. Indicateurs circonstanciels et la segmentation en cadres de discours

Dans les textes scientifiques et techniques, on rencontre très souvent, détachées en tête de phrase, des expressions partiellement figées du type de « en x », « dans x », « selon x », « au sujet de x », etc.,

qui sont plus ou moins dépendantes de la prédication verbale principale. Du fait de leur propension intégratrice, ces marqueurs indiquent des cadres de discours qui peuvent intégrer toute une série de phrases, voire un paragraphe entier.

Les cadres de discours [CHA 97] contribuent à la segmentation des discours. Les unités qu'ils configurent interagissent avec d'autres unités induites par d'autres marqueurs linguistiques comme les chaînes de référence ou les séquences de propositions reliées par des connecteurs. Les cadres de discours jouent de ce fait un grand rôle dans le filtrage.

Dans le projet soumis on s'intéressera particulièrement aux introducteurs thématiques dits aussi "indicateurs d'intérêt" [POR 98] comme « concernant x », « à propos de x », « en matière de x », question x, etc. qui signalent que le rédacteur souhaite attirer l'attention sur un sujet ou une dimension des faits rapportés qui lui paraît importante.

Plus généralement, le découpage en cadres représente dans l'analyse des données textuelles une étape essentielle pour le filtrage dans la mesure où un segment extrait peut parfaitement être intégré dans un cadre introduit par une expression qui ne figure pas dans ce segment. D'où l'intérêt de recenser et de classer les expressions potentiellement introductrices de cadres de discours et en même temps de répertorier celles à même de les fermer. Cette tâche vaut pour les introducteurs énonciatifs mais, plus généralement, pour tous les cadres : spatiaux, temporels, épistémiques, gnoséologiques, etc.).

5.2.4.4. Complémentarité des approches

L'exposé des différentes approches montre leur complémentarité : une approche globale de l'analyse d'un texte, l'approche fondée sur des critères statistiques, qui sera complétée et améliorée par une approche locale, l'approche fondée sur le repérage des cadres thématiques, mettant en évidence des indicateurs de segmentation et de structuration argumentative.

Pour mettre en oeuvre cette complémentarité, nous proposons d'identifier les segments d'un texte par une analyse de sa cohésion lexicale puis d'explorer les segments afin de détecter la présence de marqueurs permettant de prendre une décision plus fiable quant à la rupture ou à la liaison thématique. Cependant, segmenter ainsi un texte en ses différents sujets donne une vision plus globale du texte, cela ne fait pas apparaître la structure du texte du point de vue thématique et argumentatif : quels sont les liens entre les différents thèmes, l'un est-il le détail d'un point particulier d'un autre ? ou bien passe-t-on à un sujet différent ? Lorsqu'on a identifié un sujet, comment est-il développé ? quelles sont ses différentes parties du point de vue argumentatif ? La reconnaissance de marqueurs linguistiques significatifs de cette structure doit alors compléter la structure thématique.

5.2.5. Evaluation

L'évaluation du produit de la recherche ne pose pas de problème théorique : les travaux se traduiront par des améliorations de ContextO qui seront testables empiriquement. Un protocole permettant la comparaison des résultats obtenus par le système avec ceux produits par des juges externes sera mis en place. Compte tenu du délai imparti pour l'évaluation, cette comparaison sera effectuée sur des échantillons.

5.2.6 Premiers résultats

Les premiers résultats du projet ont été publiés dans un article [FER 01] que nous reproduisons partiellement ci-après.

L'extrait analysé est tiré d'un texte de quatre pages du *Monde Diplomatique* (voir ci-après) et développe un des points qui ont favorisé la réussite économique de la Côte d'Ivoire et de la Corée du sud. Dans cet extrait, les cadres soulignent la cohérence et la cohésion du passage, et segmentent le texte en unités thématiques.

5.2.6.1. Analyse linguistique

Comme nous l'avons déjà souligné, les cadres sont des guides qui permettent une interaction entre cohérence et cohésion et qui participent à la dynamique du texte. Deux sortes de cadres, qui correspondent à deux stratégies textuelles, sont instanciés dans ce passage : des cadres thématiques *(en ce qui concerne (...), quant au (...), pour ce qui concerne (...)*) et des cadres spatiaux (*en Côte-d'Ivoire, en Corée du Sud*). Les premiers soulignent le passage d'une thématique à une autre, les seconds le passage d'un point particulier à un autre (ici pays). Les introducteurs thématiques assurent la cohérence et la cohésion générale du passage. Les trois thématiques qu'ils ré-introduisent ont été précédemment énumérées en termes identiques et dans le même ordre. Les cadres spatiaux, eux, servent à illustrer localement les thématiques introduites par les introducteurs thématiques. Conformément à ce qui a été annoncé dès le début, il y en a toujours deux, ce qui assure la « cohérence illustrative » de l'ensemble du texte.

En ce qui concerne la segmentation, les expressions introductrices de cadre sont détachées en début de phrase. Les trois cadres thématiques apparaissent en début de paragraphe ce qui facilite la lecture car l'alinéa « signale au lecteur qu'il vient de traiter une unité de sens et, qu'il va passer à une unité ultérieure » [BES 88). Chaque introducteur thématique introduit ici une nouvelle unité thématique (et par là même ferme la précédente) qui correspond au paragraphe, ce principe n'étant cependant pas absolu car un cadre thématique peut en subordonner un ou plusieurs autres. Chaque point introduit par

un introducteur thématique est ensuite développé au regard de deux exemples la Côte-d'Ivoire et la Corée du Sud introduits par des *en* qui instancient des cadres spatiaux. Ces derniers se délimitent l'un l'autre et sont subordonnés aux cadres thématiques du début de chaque paragraphe.

En conséquence, dans cet extrait, on a une double structuration et une double cohérence. Chacune correspond à une stratégie textuelle et est instanciée par des expressions linguistiques différentes exerçant un contrôle interprétatif spécifique : les expressions qui instancient des cadres sont des marques linguistiques de surface qui fournissent des indications sérieuses de marques de cohérence et de segmentation.

5.2.6.2. Analyse statistique

La méthode de segmentation thématique décrite dans [MAS 95][50], révèle que les paragraphes 11 à 14 sont liés les uns aux autres. Si on cherche à faire des distinctions, les résultats ne peuvent s'interpréter qu'en termes de tendance : la méthode appliquée a tendance à lier les paragraphes 11 et 12 ainsi que les paragraphes 13 et 14. Il y a une coupure entre les paragraphes 12 et 13 mais elle est moins nette qu'entre les paragraphes14 et 15.

Les résultats de l'analyse statistique ne sont pas aussi nets que ceux de l'analyse linguistique. En particulier, les liens entre les paragraphes 11 et 12 ainsi qu'entre les paragraphes 13 et 14 (celui entre les paragraphes 12 et 13 est beaucoup plus ténu) sont difficiles à justifier sur le plan thématique. Ces résultats sont en fait la conséquence logique de la méthode utilisée qui repose sur le repérage de répétitions significatives (noms, adjectifs et verbes) pour caractériser le thème du segment. Les mêmes exemples (*Côte d'Ivoire, Corée du Sud*) relatifs à des firmes (*firmes*) étant repris d'un paragraphe à l'autre, la répartition du vocabulaire indique logiquement des liens et non des ruptures entre les paragraphes. Ce résultat n'est pas complètement en contradiction avec l'analyse linguistique et rejoint ce que nous avons appelé la « cohérence illustrative » du texte ; elle lui donne d'ailleurs la préséance, à l'inverse de l'analyse linguistique qui segmente tout d'abord en cadres thématiques.

La courbe (cf. fig. 40) montre que la segmentation thématique découpe le texte en grandes sections. Elle identifie les ruptures majeures comme celle autour du paragraphe 8 et, pour des coupures moins nettes, les informations fournies peuvent être confirmées par la présence de marqueurs (exemple de *quant au* entre §12 et §13). Combiner une telle segmentation avec des marques linguistiques permet donc de délimiter des cadres plus sûrs et de distinguer plusieurs types de frontières textuelles.

5.2.6.3. Analyse de ces résultats

Ces premiers résultats soulignent que la cohérence intervient à plusieurs niveaux et que les moyens linguistiques qui les expriment sont différents. Pour ce qui est de la prise en charge de la cohérence

thématique, trois points nous semblent intéressants à développer ultérieurement. Premièrement, les conditions dans lesquelles les introducteurs thématiques renforcent les résultats de la segmentation thématique. Deuxièmement, les conditions dans lesquelles les introducteurs corrigent les résultats de la segmentation thématique. Des indices annoncent des thématiques et s'il était possible de repérer des phrases introductrices de thèmes (cf. dernière phrase du § 11), la cohérence thématique des résumés en serait grandement améliorée. Enfin, prendre en compte les différents niveaux de segmentation, en utilisant les expressions linguistiques instanciant des cadres à l'intérieur même des paragraphes.

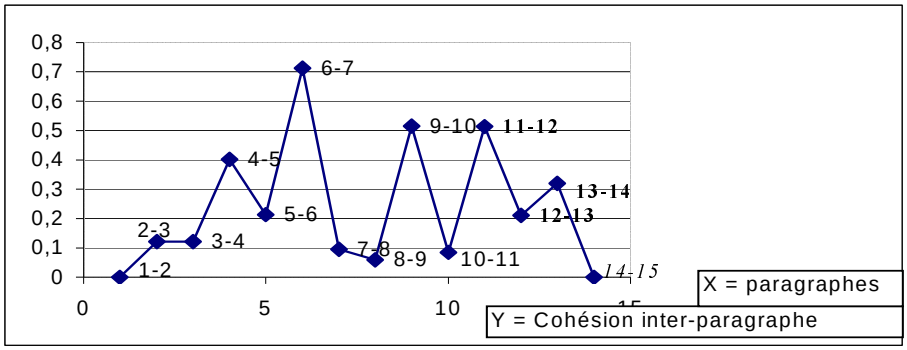


Figure 40. Repérage de la structure thématique par l'analyse statistique

[50] Les méthodes décrites dans [FER 98) et dans [HEA 97] donnent des résultats globalement similaires.

Annexe

Les lexies en gras sont des introducteurs thématiques, celles en italique, des introducteurs spatiaux.

[§11] L'importance quantitative de l'investissement étranger est cependant moins significative de l'impact des firmes multinationales que le type de secteurs où elles se localisent. *En Côte-d'Ivoire*, les firmes contrôlent pratiquement l'ensemble de l'industrie produisant pour le marché interne. Au contraire, l'accès à ce dernier leur est interdit dans la plupart des branches en Corée du Sud. Cette situation a des conséquences décisives, particulièrement sur trois variables stratégiques du processus de développement : l'allocation des ressources, le modèle de consommation et l'intégration en amont de l'activité industrielle.

[§12] **En ce qui concerne** l'allocation des ressources, *en Côte-d'Ivoire*, l'excédent prélevé par l'Etat et consacré à l'expansion du marché intérieur transite nécessairement par les firmes multinationales, finançant en grande partie leur implantation ou l'élargissement de leur capacité productive. *En Corée du Sud,* les firmes multinationales sont exclusivement concentrées dans les branches exportatrices, ce qui permet à l'Etat de prélever des ressources externes additionnelles que les entreprises publiques ou privées coréennes utilisent selon les orientations précises du plan dans le cadre d'une stratégie d'intégration industrielle orientée vers le marché intérieur.

[13] **QUANT au** modèle de consommation, *en Côte-d'Ivoire*, la production de biens relève de la stratégie propre à la firme multinationale, sans rapport avec le niveau moyen des revenus et les habitudes traditionnelles de consommation. Ce phénomène suscite ou accentue à son tour la distribution inégalitaire du revenu. *En Corée du Sud*, la diversification des biens offerts aux consommateurs est un processus progressif et contrôlé en relation étroite avec la capacité d'achat de la population. Cette correspondance entre niveau de revenu et offre de biens contribue fortement à atténuer les tendances à la répartition inégalitaire des revenus. La politique du pouvoir, dans ce domaine, a été très ferme. Les biens de consommation les plus modernes - électroménager, appareils optiques, électronique grand public, - fabriqués en grande partie par les firmes multinationales, ont été longtemps exclusivement destinés à l'exportation. La population coréenne n'a eu accès à ces biens qu'une fois satisfaits les besoins essentiels en matière de nourriture et de vêtement. Mais le développement du marché interne n'a pas profité aux firmes multinationales qui en ont été pratiquement exclues au profit des firmes locales. Dans la branche électronique grand public, par exemple, les ventes sur le marché interne sont réalisés pour 95,4 % par les entreprises coréennes et pour 4,4 % par des entreprises en joint venture.

[§14] Enfin, **pour ce qui concerne** l'intégration en amont de l'activité industrielle, *dans un pays comme la Côte-d'Ivoire*, où le secteur industriel est contrôlé par les firmes étrangères, la taille du marché a constitué l'obstacle insurmontable à la diversification de la structure productive. En conséquence, le processus reste bloqué au niveau des branches légères. *En Corée du Sud*, la maîtrise absolue de l'Etat sur la décision économique au niveau du marché interne a permis ce que l'on appelle la "remontée des filières" vers les industries lourdes - sidérurgie, chimie et industries de biens d'équipement - et assuré une autonomie notable du processus d'industrialisation, même si, dans certains secteurs, la dimension du marché était manifestement insuffisante.

[§15] Cette rapide comparaison montre que le diagnostic établi par les analystes des problèmes du développement n'est pas aussi faux que cela et que la thérapie proposée, qui est une "thérapie douce", loin de conduire à des situations apocalyptiques, peut se révéler efficace.

Bibliographie

[ABE] ABEILLÉ A., P. BLACHE. (2000) *Grammaires et analyseurs syntaxiques*, in *Ingénierie des langues*, (sous la direction de J-M. Pierrel) Paris, Editions Hermès, p. 51-76.

[ABR 97] ABRACOS, J., G. P. LOPES. (1997). Statistical methods for retrieving most significant paragraphs in newspaper articles*, Proceedings of a Workshop : Intelligent Scalable Text Summarization*, EACL 97, Madrid, p. 51-57.

[ADA 90] ADAM, J-M. (1990). *Éléments de linguistique textuelle*, Mardaga, Liège.

[ALT 90] ALTERMAN, R., L. A. BOOKMAN. (1990). Some Computational Experiments in Summarization, *Discourse Processes,* 13, p. 143-174.

[ALT 91] ALTERMAN, R. (1991). Understanding and summarisation, *Artificial Intelligence Review*, 5, p. 239-254.

[APP 93] APPELT, D, J. HOBBS, J. BEAR, D. ISRAEL, M. KAMEYANA, M. TYSON. (1993). FASTUS : a finite-state processor for information extraction from real-world text, *Proceedings of the International Joint Conference on Artificial Intelligence.*

[BAE 99] BAEZA-YATES R., RIBEIRO-NEO B. (1999). *Modern Information Retrieval,* Addison Wesley, NewYork.

[BAL 00] BALDO, S. (2000). *Représentation des valeurs sémantiques de l'auxiliaire modal WOULD en anglais : étude linguistique et système d'exploration contextuelle en vue du traitement informatique de sa traduction automatique en français,* Thèse de doctorat, Université Paris-Sorbonne, Paris.

[BAR 97] BARZILAY, R., M. ELHADAD. (1997). Using Lexical Chain for Text Summarization, *Intelligent Scalable Text Summarization, Workshop* EACL 97, Madrid, p. 11-17.

[BAT 00] BATTISTELLI, D. (2000). *Passer du texte à une séquence d'images : analyse spatio-temporelle de textes, modélisation et réalisation informatique (système SPAT).* Thèse de doctorat, Université Paris-Sorbonne, ParBENis.

[BEN 99] BEN HAZEZ, S. (1999). BDContext : un système de gestion de connaissances linguistiques orientées vers le filtrage sémantique de textes, *Acte de colloque international, CIDE'99*, Damas, Syrie.

[BEN 00] BEN HAZEZ, S., J.-L. MINEL. (2000). Designing Tasks of Identification of Complex Patterns Used for Text Filtering, *RIAO'2000*, *Content-Based Multimedia Information Acess*, Paris, p. 1558 - 1567.

[BEN 01] BEN HAZEZ, S., J.-P. DESCLÉS, J.-L. MINEL. (2001). Modèle d'exploration contextuelle pour l'analyse sémantique des textes, *TALN 2001*, p.73-82, Tours.

[BEN 02] BEN HAZEZ, S. (2002). Thèse de doctorat en cours, Université Paris-Sorbonne, Paris.

[BER 95a] BERRI, J, D. LE ROUX, D. MALRIEU, J.-L. MINEL. (1995). Pour automatiser l'activité résumante, *Actes du colloque JADT 95*, Rome, p. 345-352.

[BER 95b] BERRI, J, D. LE ROUX, D. MALRIEU, J.-L. MINEL. (1995). SERAPHIN un système d'extraction automatique d'énoncés importants, *Actes des Journées de Génie Linguistique,* Montpellier, France, p. 409-419.

[BER 96a] BERRI, J. (1996). *Contribution à la méthode d'exploration contextuelle. Applications au résumé automatique et aux représentations temporelles. Réalisation informatique du système SERAPHIN*. Thèse de doctorat, Université Paris-Sorbonne, Paris.

[BER 96b] BERRI, J, E. CARTIER, J-P. DESCLES, A. JACKIEWICZ, J.-L. MINEL. (1996). SAFIR, système automatique de filtrage de textes, *Actes du colloque TALN'96*, Marseille, p. 140-149.

[BES 88] BESSONAT, D. (1988). Le découpage en paragraphes et ses fonctions, *Pratiques* n°57, p. 85-97.

[BIB 88] BIBER, D. (1988). *Variation across speech and writing*, Cambridge, Cambridge University Press.

[BIB 89] BIBER, D. (1989). A typology of English texts, *Linguistics*, (27), p. 3-43.

[BON 00] BONHOMME P. (2000) *Codage et normalisation de ressources textuelles*, in *Ingénierie des langues*, (sous la direction de J-M. Pierrel) Paris, Editions Hermès, p. 173-91.

[BOU 94] BOURIGAULT, D. (1994). *LEXTER, un logiciel d'Extraction de TERminologie. Applications à l'extraction des connaissances à partir d'un texte*, Thèse de doctorat, EHESS, Paris.

[BUR 92] BURKOWSKI F. J. (1992). An algebra for hierarchically organized text-dominated databases, *Information Processing and Managemen,t* 28 (3), p. 333-348.

[CAR 96] CARTIER, E. (1996). SERAPHIN, marqueurs sémiotiques et linguistiques, règles d'exploration contextuelle et phase d'ajustement, *Rapport interne 96/1 du CAMS-LALIC*.

[CAR 97] CARTIER, E. (1997). La définition dans les textes scientifiques et techniques : présentation d'un outil d'extraction automatique des relations définitoires, *TIA'97*, Toulouse.

[CHA 00] CHAGNOUX,, M. (2000). Perspectives d'implémentation des valeurs aspecto-temporelles inter-propositionnelles*, mémoire de DEA Miash,* Université Paris-Sorbonne, Paris.

[CHA 88] CHAROLLES, M. (1988). Les plans d'organisation textuelle ; période, chaînes, portées et séquences, *Pratiques*, n° 57, p. 3-14.

[CHA 89] CHAROLLES, M. (1989). Marquages linguistiques et résumé de texte, *Recherche Linguistique XVII,* p. 11-27, Klincksieck.

[CHA 91] CHAROLLES, M. (1991). Marquages linguistiques et résumé de texte, *Actes du colloque international de linguistique (aspects linguistiques, sémiotiques, psycholinguistiques et automatiques)*, Charolles M., Petitjean (eds.), Pont-à-Mousson, Paris, Klincksieck, p. 11-27.

[CHA 97] CHAROLLES, M. (1997). L'encadrement du discours : univers, champs, domaines et espaces, Cahier de Recherche Linguistique, LANDISCO, Université Nancy 2, 6, p. 1-73.

[CHA 98] CHAROLLES, M. (1998). L'organisation du texte, le filtrage et le résumé, *RIFRA'98, Rencontre Internationale sur l'Extraction le Filtrage et le Résumé Automatiques*, Sfax, Tunisie.

[CHA 99] CHAROLLES, M. (1999). Phrase, texte, discours, *Langue Française*, n° 121, p. 76-116.

[CHI 86] CHIARAMELLA Y., B. DEFUDE., M.-F. BRUANDET, D. KERKOUBA. (1986). IOTA : a full text information retrieval system, *Proceddings of Conference on Research and Development in Information Retrieval,* p. 207-213, Pisa, Italy.

[CLA 95] CLARKE CH., G. V. CORMACK, F. J. BURKOWSKI. (1995). An algebra for structured text search and a framework for its implementation, *The Computer Journal,* 38(1), p. 43-56.

[COU 01] COUTO J. (2001). ContextO, *Los sistemas de exploracion contextual de cara al usuario*, Mémoire de Master, Université de la République, Uruguay.

[CRI 99a] CRISPINO G., S.BEN HAZEZ., J.-L. MINEL. (1999). ContextO, un outil de la plate-forme d'ingénierie linguistique Filtext, *VEXTAL 99*, Venise, Italie, p. 361-367.

[CRI 99b] CRISPINO G., S. BEN HAZEZ , J.-L. MINEL. (1999). Architecture logicielle de Context, plate-forme d'ingénierie linguistique, *TALN 99,* p. 327-332.

[CRI 99c] CRISPINO G., S. BEN HAZEZ , J.-L. MINEL, G. MOURAD. (1999). ContextO : Una plataforma de ingeniería lingüística orientada al filtrado semántico de textos, *SEPLN, Procesamiento de Lenguaje Natural, Revista nº 25,* p. 215-216

[CRI 02] CRISPINO, G. (2002). Thèse de doctorat,en cours, Université Paris-Sorbonne, Paris.

[DEJ 82] DE JONG G. (1982). An overview of the FRUMP system, *Strategies for Natural Language Parsing,* p. 149-176, W. Lenhert et M. Ringle (eds), Lawrence Erlbaum Associates, Hillsdale, NJ.

[DEN 85] DENHIERE G. (1985). *Il était une fois*, P.U., Lille.

[DES 87] DESCLES, J.-P. (1987). Réseaux sémantiques : la nature logique et linguistique des relateurs, *langages, Sémantiques et Intelligence Artificielle,* n° 87, p. 55-78.

[DES 88] DESCLES, J.-P. (1988). Langage et cognition : avant-propos, *Intellectica*, n°6, p. 1-41.

[DES 91a] DESCLES, J.-P., C. JOUIS, H-G. OH, D. MAIRE REPPERT. (1991). Exploration Contextuelle et sémantique : un système expert qui trouve les valeurs sémantiques des temps de l'indicatif dans un texte, In *Knowledge modeling and expertise transfer*, p. 371-400, D. Herin-Aime, R. Dieng, J-P. Regourd, J.P. Angoujard (éds), Amsterdam.

[DES 91b] DESCLES, J.-P. (1991). Architectures, représentations cognitives et langage naturel, in *Les sciences cognitives en débat,* Edition du CNRS, Paris.

[DES 93] DESCLES, J.-P., C. JOUIS. (1993). L'exploration contextuelle: une méthode linguistique et informatique pour l'analyse informatique de textes, *ILN93*, Nantes.

[DES 97a] DESCLES, J.-P., E CARTIER, A. JACKIEWICZ, J.-L. MINEL. (1997). Textual Processing and Contextual Exploration Method in *CONTEXT 97*, Universidade Federal do Rio de Janeiro, Brésil, p. 189-197.

[DES 97b] DESCLES, J-P. (1997). *Systèmes d'exploration contextuelle. Co-texte et calcul du sens*. (ed Claude Guimier), Presses Universitaires de Caen, p. 215-232.

[DES 01a] DESCLES, J.-P. (2001). *Abduction and non observability*, E. Agazzi and M. Pauri(eds), The reality of the Unobservable, p. 87-112, Kluwer Academic Publishers, Netherlands.

[DES 01b] DESCLES, J.-P. (2001). L'exploration contextuelle, communication au séminaire « Cadres et discours », Université Paris 3, Février 2001.

[DUC 01] DUCHESNE S., F.HAEGEL.(2001). Entretiens dans la cité ou comment la parole se politise, *à paraître*.

[ELL 98] ELLOUZE, M., A. BEN HAMADOU. (1998). Utilisation de schémas de résumés en vue d'améliorer la qualité des extraits et des résumés automatiques, *RIFRA'98, Rencontre Internationale sur l'Extraction le Filtrage et le Résumé Automatiques*, Sfax, Tunisie, p. 95-107.

[END 95] ENDRES-NIGGEMEYER B., E. MAIER , A. SIGEL. (1995). How to implement a naturalistic model of abstracting : four core working steps of an expert abstractor, *Information Processing & Management*, 31(5), p. 631-674.

[ELL 00] ELLEUCHE YAICHE, W. (2000). Aide à la traduction du modal auxiliaire "WOULD" ; représentation des connaissances linguistiques et contribution à l'extension de la plate forme ContextO*, DEA Miash,* Université Paris-Sorbonne, Paris.

[EHR 93] EHRLICH, M-F, H. TARDIEU. (1993). Modèles mentaux, modèles de situation et compréhension de textes in *Les modèles mentaux, approches cognitives des représentations*, Masson.

[FAY 89] FAYOL M. (1989). Le résumé : un bilan provisoire des recherches en psychologie cognitive, *Actes du colloque international de linguistique (aspects linguistiques, sémiotiques, psycholinguistiques et automatiques)*, Charolles M., Petitjean (eds.), Pont-à-Mousson, Paris, Klincksieck, p. 163-182.

[FER 98a] FERRET, O., B. GRAU, N. MASSON. (1998). Thematic segmentation of texts: two methods for two kinds of texts, *ACL/COLING '98*, p. 392-396.

[FER 98b] FERRET, O., B. GRAU. (1998). A Thematic Segmentation Procedure for Extracting Semantic Domains from Texts, *European Conference on Artificial Intelligence (ECAI)*, Brighton, Grande-Bretagne, p. 155-159.

[FER 01] FERRET, O., B. GRAU, J.-L. MINEL, S. PORHIEL. (2001). Repérage de structures thématiques dans des textes, *TALN 2001*, p. 163-172, Tours.

[FIL 68] FILLMORE, C. (1968). The case for case in *Universals in linguistic theory,* B. Harms (éd), Holt, Rinehart and Winston, Chicago, p. 1-90.

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

[GAU 01] GAUVAIN, P. (2001). *La notion de changement dans l'œuvre d'Aristote.* Thèse de doctorat en cours, Université Paris-Sorbonne, Paris.

[GRO 85] GROUPE §. (1985). *La notion de paragraphe*, Textes réunis par Roger LAUFER, Editions du CNRS.

[GUI 01] GUIRAUDIE, O. (2001). *Repérage de la politisation de la parole par exploration contextuelle.* Mémoire de DEA MIASH, Université Paris-Sorbonne, Paris.

[HAN 97] HAN, T.F. (1997). A Proposal for Task-Based Evaluation of Text Summarization Systems, *Workshop Intelligent Scalable Text Summarization,* EACL 97, Madrid, p. 31-38.

[HAR 88] HARRIS, Z.S. (1988). *Language and information.* New York: Columbia University Press.

[HAR 89] HARRIS, Z.S, GOOTFRIED, M., RYCKMAN, T., MATTICK JR, P., DALADIER, A., HARRIS, T.N. (1989). The form of information in science. Analysis of an immunology sublanguage, Dordrecht : Kluwer Academic Publishers.

[HAR 91] HARRIS, Z.S. (1991). A theory of language and information : A mathematical approach. Oxford : Clarendon Press.

[HAR 96] HAMAN, D. , E.M. VOORHES. (1996) *The Fifth Text Retrieval Conference* (TREC 5), Technical Report, SP-500-238, National Institute od Standards and Technology, Gaithersburg, Maryland.

[HEA 97] HEARST, M. A. (1997). TextTiling : Segmenting Text into Multi-paragraph Subtopic Passages, *Computational Linguistics, 23*, 1, p. 33-64.

[HER 96] HERVIOU, M-L, R. QUATRAIN, M-G MONTEIL (1996). Construction de terminologies : une chaîne de traitement supportées par un atelier intégrant outils linguistiques et statistiques, *TALN'96*, Marseille, p. 130-139.

[HOB 92] HOBBS, J, D. APPELT, J. BEAR, M. TYSON, D. MAGERMAN (1992). Robust Processing of Real-World Natural Language Texts*, in Text-based Intelligent Systems : Current Research and Practice in Information Extraction and Retrieval*, Laurence Erlbaum Associates, Hillsdale, N-J.

[KHE 00] KHETTAR A. (2001). Réalisation d'un compilateur de règles d'Exploration Contextuelle, *mémoire de DEA MIASH*, Université Paris-Sorbonne.

[KIN 75] KINTSCH W., T. A. VAN DIJK. (1975). Comment on se rappelle et on résume une histoire, *Langage*, 40, p. 98-116.

[KIN 78] KINTSCH W., T. A. VAN DIJK. (1978). Toward a model of text comprehension and production, *Psycological review*, 85, p. 363-394.

[KIN 88] KINTSCH, W. (1988). The role of knowledge in discourse comprehension : A construction-integration model, *Psycological review*, 95, p. 163-182.

[KUP 95] KUPIEC, J., J. PEDERSEN, F. CHEN. (1995). A trainable document summarizer, *Proceedings of the 18th ACM-SIGIR Conference*, p. 68-73.

[JAC 98] JACKIEWICZ, A. (1998). *L'expression de la causalité dans les textes. Contribution au filtrage sémantique par une méthode informatique d'exploration contextuelle*. Thèse de Doctorat, Université Paris-Sorbonne.

[JAC 01] JACQUEMIN C. (2001). *Spotting and Discovering Terms Through Natural Language Processing.* M.I.T. Press.

[JEA 01] JEANNERET Y. (2001). *Étude et propositions sur la politique scientifique de recherche au GRIPIC,* rapport interne au CELSA, Paris.

[JIN 98] JING, HONGYAN, R. BARZILAY, K. MCKEOWN. (1998). Summarization evaluation methods : Experiments and analysis. In *Symposium on Intelligent Text Summarization ACL,* Stanford, CA.

[JOU 93] JOUIS, C. (1993). *Contribution à la conceptualisation et à la modélisation des connaissances à partir d'une analyse linguistique de textes*. Thèse de doctorat, EHESS, Paris.

[LAU 98] LAUBLET, A. (1998). Méthodes de développements d'applications à objets, H3 228, in *les Techniques de l'Ingénieur, traité informatique*, p. 1-15.

[LEH 95] LEHMAM, P. (1995). *Le résumé de textes techniques et scientifiques, aspects linguistiques et computationnels*, Thèse de doctorat, Université de Nancy 2.

[LEP 99] LE PRIOL, F. (1999). A data processing sequence to extract terms and semantics relations between terms, *10th mini EURO Conference Human Centered Processes, HCP'99*.

[LEP 00] LE PRIOL, F. (2000). *Extraction et capitalisation automatiques de connaissances à partir de documents textuels. SEEK-JAVA : identification et interprétation de relations entre concepts,* Thèse de Doctorat, Université Paris-Sorbonne.

[LER 92] LEROUX, D. (1992). Automatisation de l'activité résumante. Le résumé de texte, Actes publiés par M. Charolles et A . Petitjean, p. 237-253, Klincsieck.

[LER 94] LEROUX, D., J.-L. MINEL, J. BERRI. (1994). SERAPHIN project. *First European Conference of Cognitive Science in Industry*, Luxembourg, p. 275-283.

[LEV 89] LEVINE, P, J.-L.. MINEL. (1989). *A development Tool for Expert Systems in the field of regulations*, in *Expert Systems in Public Administration*, Elsevier.

[LEV 90] LEVINE, P, J-C. POMEROL. (1990). *Systèmes experts dans l'entreprise, Hermés*.

[MAI 92] MAIRE-REPPERT, D. (1992). *Les temps de l'indicatif du français en vue d'un traitement informatique : imparfait*, Thèse de Doctorat, EHESS, Paris.

[MAN 01] MANI, I. (2001). *Automatic Summarization*, John Benjamins Publishing Company, Amsterdam.

[MAN 99] MANI, I., M. T. MAYBURI. (1999). *Advances in Automatic Summarization*, MIT Press, London.

[MAN 88] MANN, W. C., S. A THOMPSON. (1988). Rhetorical Structure Theory : Toward a functional theory of text organization, *Text*, 8(3) p. 243-281.

[MAR 97] MARCU, D. (1997). From discourse structures to text summaries, in *Workshop Intelligent Scalable Text Summarization,* EACL 97, Madrid, p. 82-88.

[MAS 95] MASSON, N. (1995). *An Automatic Method for Document Structuring*, in Proceedings of the 18th Annual International ACM-SIGIR, Conference on Research and Development in Information Retrieval, Seattle, Washington, USA.

[MAS 98] MASSON, N. (1998). *Méthodes pour une génération variable de résumé Automatique : Vers un système de réduction de textes*, Thèse de Doctorat, Université Paris-11.

[MEU 98] MEUNIER, J-G. (1998). La gestion des connaissances et les intelliciels. *RIFRA'98, Rencontre Internationale sur l'Extraction le Filtrage et le Résumé Automatiques*, Sfax, Tunisie.

[MEY 88] MEYER, B. (1988). *Object-Oriented Software Construction*, Prentice Hall.

[MIE 93] MIEVILLE , D., (Editeur). (1993). Relations formelles et non formelles, *Travaux du Centre de Recherches Sémiologiques*, n° 61, Université de Neuchâtel.

[MII 94] MIIKE, S., E. ITOH, K. ONO., K. SUMITA. (1994). A full-text retrieval system with a dynamic abstract generation function, *Proceedings Sigir'94*, ed by W. Bruce Croft and C. J. van Rijsbergen, Springer-Verlag, Dublin, p. 152-161.

[MIL 88] MILLER, G. A., C. FELLBAUM, J. KEGL. (1988). WORDNET : An Electronic Lexical Reference System Based on Theories of Lexical Memory, *CSL Report 11*, Cognitive Science Laboratory, Princeton University.

[MIL 90] MILLER, G. A., R. BECKWITH, C. FELLBAUM, D. GROSS, K. J. MILLER. (1990). Introduction to WordNet : An on-line lexical database, *International Journal of Lexicography (special issue*), 3(4) p. 235-312.

[MIN 97] MINEL , J.-L., S. NUGIER, G. PIAT. (1997). How to appreciate the Quality of Automatic Text Summarization, *Workshop Intelligent Scalable Text Summarization,* EACL 97, Madrid, p. 25-30.

[MIN 00] MINEL J.-L., J.-P. DESCLES. (2000) *Résumé Automatique et Filtrage des textes*, in *Ingénierie des langues*, (sous la direction de J-M. Pierrel) Paris, Editions Hermès, p. 253-270.

[MIN 01] MINEL J.-L., E. CARTIER, , G. CRISPINO, J.-P. DESCLES, S. BEN HAZEZ, A. JACKIEWICZ. (2000). Résumé automatique par filtrage sémantique d'informations dans des textes, Présentation de la plate-forme FilText, *Technique et Science Informatiques,* n° 3, Paris.

[MIN 75] MINSKY M. (1975). A framework for representing knowledge, *The psychology of computer vision*, in P. Winston (ed.), New York, Mc Graw-Hill.

[MIT 97] MITRA M., A. SINGHAL , C. BUCKELY. (1997). Automatic Summarization by Paragraph Extraction*, Proceedings of a Workshop : Intelligent Scalable Text Summarization*, EACL 97, Madrid, p. 39-46.

[MOR 91] MORRIS, J., HIRST G. (1991). Lexical cohesion computed by thesaural relations as an indicator of the structure of the text, *Comptutational Linguistics*, 17(1), p. 21-45.

[MOU 99] MOURAD, G. (1999). La segmentation des textes par l'étude de la ponctuation, *CIDE'99*, Damas, Syrie.

[MOU 00] MOURAD, G., J.-L. MINEL. (2000). Filtrage sémantique du texte, le cas de la citation. 3e Colloque International sur le Document Électronique, *CIDE'2000*, Lyon, p. 41-56.

[MOU 01] MOURAD, G. (2001). . *Analyse informatique des signes typographiques pour la segmentation de textes et l'extraction automatique des citations. Réalisation des applications informatiques SegAtex et CitaRE.*Thèse de doctorat, Université Paris-Sorbonne, Paris.

[NAI 98] NAIT-BAHA, L., A. JACKIEWICZ, P. LAUBLET. (1998). Reformulation de requêtes et extraction de phrases pertinentes pour la collecte d'informations sur le Web, *RIFRA'98, Rencontre Internationale sur l'Extraction le Filtrage et le Résumé Automatiques*, Sfax, Tunisie, p 177-190.

[NAI 99] NAIT-BAHA, L., A. JACKIEWICZ, B. DJIOUA, P. LAUBLET. (1999). Reformulation de requêtes pour la collecte d'informations sur le Web à partir de points de vue : premiers résultats, *ISKO'99*.

[NAI 01] NAIT-BAHA, L. (2001). *Reformulation de requêtes pour la collecte d'informations sur le Web à partir de points de vue,* Thèse de doctorat en cours, Université Paris Sorbonne.

[NAV 01a] NAVA, M. (2001). *Contextualisation automatisée de syntagmes nominaux*, Thèse de doctorat en cours, Université Paris-Sorbonne, Paris.

[NAV 01b] NAVA, M., D. GARCIA. (2001). Contextualisation automatisée de syntagmes nominaux pour la navigation dans une mémoire d'entreprise, *TIA 2001*, Nancy.

[PAI 81] PAICE, C. D. (1981). The automatic generation of literature abstracts : an approach based on the identification of self indicating phrases, *Information retrieval research*, Oddy, R. N. (Ed.), p. 172-191.

[PAI 90] PAICE, C. D. (1990). Constructing literature abstracts by computer : techniques and prospects, *Information processing management*, 26 (1), p. 171-186.

[PAR 99] PARIENTI, M. (1999). *Etudes des liaisons entre le projet RAP et la plate-forme Context0, mémoire de DEA Miash*, Université de Paris-Sorbonne.

[PAZ 97] PAZIENZA, M.T. (1997). (éd.). Information extraction (a multidisciplinary approach to an emerging information technology*), International Summer School, SCIE'97*, Springer Verlag (Lectures Notes in Computer Science).

[PER 00] PÉRY-WOODLEY, M.P. (2000). Une pragmatique à fleur de texte : approche en corpus de l'organisation textuelle, *Carnets de grammaire*, n° 8.

[PER 01] PÉRY-WOODLEY, M.P. (2001). Modes d'organisation et de signalisation dans des textes procéduraux, *Langages*, n° 41, p. 28-46.

[POI 99] POIBEAU, T., A. NAZARENKO. (1999). L'extraction d'information, une nouvelle conception de la compréhension de texte ?, T.A.L., vol 40., n°2, p. 87-115.

[POR 98] PORHIEL, S. (1998). *Les indicateurs d'intérêt,* Thèse de doctorat, Université Paris 13, Lille, Presses du Septentrion.

[POR 01] PORHIEL, S. (2001). Organising Linguistic Data: Thematic Introducers as an Example*, à paraître*, Coyote Press.

[RAT 61] RATH, G. J., A. RESNICK, T. SAVAGE. (1961). The formation of abstracts by the selection of sentences, *American Documentation*, 12(2), p.139-143.

[REB 98] REBEYROLLE, J., M.-P. PERY-WOODLEY. (1998). Repérage d'objets textuels fonctionnels pour le filtrage d'information : le cas de la défintion, *RIFRA'98, Rencontre Internationale sur l'Extraction le Filtrage et le Résumé Automatiques*, Sfax, Tunisie, p. 19-30.

[RIN 94] RINO, L., D. SCOTT. (1994). Automatic generation of draft summaries : heuristics for content selection, *Rapport interne*, Centre IBM, Royaume uni.

[ROM 00] ROMARY L. (2000). *Outils d'accès à des ressources linguistiques,* in *Ingénierie des langues*, (sous la direction de J.-M. Pierrel), Paris, Editions Hermès, p. 194-211.

[ROS 00] ROSSARY C. (2000). *Connecteurs et relations de discours* : *des liens entre cognition et signification*, Nancy, Presses Universitaires de Nancy.

[RUM 75] RUMELHART D. (1975). Notes on a schema for stories, *Representation and understanding : Studies in cognitive science*, Bobrow D. & Collins A. (eds.), New York, Academic Press, p.211-236.

[SAL 83] SALTON, G., M. GILL (1983). Introduction to Modern Information Retrieval, Mac Graw Hill Book Co, New York.

[SAL 89] SALTON, G. (1989). Automatic text processing, the transformation, analysis, and retrieval of information by computer, Addison-Wesley, Reading.

[SAG 20] SAGGION, H., G. LAPALME. (2000). Selective Analysis for Automatic Abstracting : Evaluating Indicativeness and Acceptability, *RIAO 2000, Content-Based Multimedia Information Access*, Paris, p 747-764.

[SAG 98] SAGGION, H., G. LAPALME. (1998). Where does information come from ? Corpus Analysis for Automatic Abstracting, *RIFRA'98, Rencontre Internationale sur l'Extraction le Filtrage et le Résumé Automatiques*. Sfax, Tunisie, p. 84-94.

[SCH 75] SCHANK R. (1975). The structure of episodes in memory, *Representation and understanding : Studies in cognitive science*, Bobrow D. & Collins A. (eds.), New York, Academic press.

[SCH 77] SCHANK R., R. ABELSON. (1977). *Scripts, plans goals, and understanding*, Hillsdale N. J., Erlbaum.

[SIL 93] SILBERZTEIN, M. (1993). Dictionnaires électroniques et analyse automatique de textes : le système INTEX, *Masson*.

[SLA 75] SLATKA, D. (1975). L'ordre du texte, *Cahiers de linguistique appliquée*, 19, p. 30-42.

[SLA 80] SLATKA, D. (1980). *Sémiologie et grammaire du texte*, Thèse d'état, Université Paris 10, Nanterre.

[SPA 93] SPARCK JONES, K. (1993). What might be in a summary ?, in Knorz, Krause and Wormser-Hacke (eds.) *Information Retrieval 93*, p. 9-26, Universitates verlag Konstanz.

[SPA 96] SPARCK JONES, K. (1996). Evaluating, *Nordic journal of linguistics*, 11, p. 89-110.

[SPA 99] SPARCK JONES, K. (1999). Automatic summarizing : factors and directions , in *Advances in Automatic Text Summarization*, (edited by I. Mani & m.T. Maybury), p. 1-14, MIT Press.

[SOL 68] SOLNIK, H. (1968). Historical development of abstracting, *Journal of Chemical Information and Computer science*, 19,4, p. 215-218.

[SPR 80] SPRENGER-CHAROLLES, L. (1980). Le résumé de textes, *Pratiques*, 26, p. 59-90.

[STA 96] STAIRMAN, M. (1996). *A Computational Analysis of Lexical Cohesion with Applications in Information Retrieval,* Ph. D. thesis, Center for Computational Linguistics, UMIST, Manchester.

[TEU 97] TEUFEL S., M. MOENS. (1997). Sentence exraction as a classification task, *Workshop Intelligent Scalable Text Summarization,* EACL 97, Madrid, p. 58-65.

[VAN 83] VAN DIJK, T. A., W. KINTSCH. (1983). *Strategies of discourse comprehension*, New York, Academic press.

[VAU 85] VAUQUOIS, B., C. BOITET. (1985). Automated Translation at Grenoble University, *Computational Linguistics,* 11/1, p. 28-36.

[VIC 98] VICTORRI, B., C. FUCHS. (1998). *La Polysémie*, Hermès.

[VIR 85] VIRBEL, J. (1985). Langage et métalangage dans le texte du point de vue de l'édition en informatique textuelle, *Cahiers de Grammaire*, n° 10, p. 5-72.

[WIT 73] WITTY, F.J. (1973). The beginnings of indexing and abstracting notes toward a history of indexing in antiquity and Middle ages, *Journal of Chemical Information and Computer science*, 8,4, p. 193-198

[WON 01] WONSEVER, D., J.-L. MINEL. (2001). Contextual Rules for Text Analysis, *Proceedings of Cicling 2001*, *Lecture Notes in Artificial Intelligence*, Springer-Verlag, p. 507-521.

[WON 02] WONSEVER, D. (2002). *Identification des propositions*, Thèse de doctorat, en cours, Université Paris-Sorbonne, Paris.

[XML 98] BRAY, T., J. PAOLI, C.M. S P. SPERBERG-MCQUEEN. (1998). Extensible Markup Language (XML) 1.0, W3C Recomendation.

[ZWE 97] ZWEIGENBAUM, P. (1997). *Traitement automatique de la langue médicale*, Habilitation à diriger des recherches, Université Paris-Nord.

Annexes

Exemples de résumés et de textes étiquetés sémantiquement

Exemple 1

Résumé automatique de l'article « Résumé automatique par filtrage sémantique d'informations dans des textes» [7] *publié dans la revue Techniques et science informatique, avec un profil du type « résumé-auteur » et un seuil de filtrage limité à 10 % du texte source. Les titres, les sections, les paragraphes et les phrases ont été automatiquement délimités. Pour ce type de profil, les titres de chaque section du texte original sont systématiquement placés dans le résumé, même si aucune phrase de la section n'a été sélectionnée. Le temps de traitement sur une station de type micro-ordinateur est d'environ 6 secondes. Aucune post-édition manuelle n'a été effectuée à l'exception de la mise aux normes rédactionnelles.*

Résumé

Résumé automatique par filtrage sémantique d'informations dans des textes

1. Introduction

[...] Les techniques traditionnelles fondées sur des techniques purement quantitatives de recherche d'informations ne sont pas toujours très satisfaisantes et ne répondent pas assez bien aux réels besoins des utilisateurs.[...] Un document textuel devra donc être maintenant géré en même temps que son résumé qui sera, par ailleurs, un des moyens d'accès au contenu du document.[...] La première réalisation que nous avons menée avec le projet SERAPHIN, en partenariat avec la Direction des études et des Recherches (DER) d'EDF [LER 94] était fondée sur l'utilisation de ressources linguistiques qualifiées de sémantiques. Elle visait une application précise : la diffusion sélective d'informations par l'interrogation en ligne d'une base de données textuelles, de façon à fournir un produit intermédiaire entre d'un côté, la chaîne d'indexation, trop pauvre, ambiguë et incapable de donner un aperçu complet du texte et d'un autre côté, le texte intégral, trop lourd à manipuler par les utilisateurs car trop riche et pas assez sélectif.

2. Méthodologie : méthode d'exploration contextuelle

L'originalité de notre approche revient à se donner les moyens d'accéder au contenu sémantique des textes, pour mieux les cibler et en extraire des séquences particulièrement pertinentes. A cet effet, nous exploitons un savoir purement linguistique, et plus précisément sémantique. [...] Notre but est de cibler, à l'aide de marqueurs linguistiques et de certaines connaissances grammaticales, des séquences textuelles qui peuvent exprimer un certain savoir sur le monde [...] Il ne s'agit pas d'une utilisation de mots clés ou d'une simple analyse distributionnelle, puisque l'exploration contextuelle met en jeu des processus inférentiels [DES 97 b] qui sont déclenchés, dans un premier temps, par l'identification d'indicateurs linguistiques relatifs à un champ grammatical ou discursif précis.[...] L'exploration contextuelle est donc gouvernée par un ensemble de règles (dites d'exploration) qui, pour un marqueur donné et une décision à prendre, recherchent d'autres indices explicites dans un espace de recherche (proposition, phrase, paragraphe ...) déterminé par la règle.[...] L'acquisition de ces données linguistiques nécessite une fouille systématique des textes en vue d'accumuler les indicateurs, les indices et les règles qui les combinent ; cette fouille est complétée par un travail de réflexion linguistique, afin de dégager les régularités textuelles. [...] Soulignons que l'exploration contextuelle ne nécessite pratiquement pas de connaissances des domaines traités, c'est-à-dire qu'il n'est pas nécessaire de construire des représentations des connaissances préalables à l'analyse sémantique du texte.

3. Plate-forme FilText

[...] Nous avons élaboré un modèle conceptuel et un langage de description de ces données linguistiques

3.1. *Organisation et gestion des données linguistiques*

3.1.1. *Description des données linguistiques*

Nous avons défini un langage de description qui permet au linguiste de constituer sa base de données linguistiques en spécifiant : les tâches, les indicateurs ou les indices pertinents et les règles d'exploration contextuelles associées.

3.1.2. *Règles d'exploration contextuelle*

Les règles d'exploration contextuelle sont exprimées dans un langage formel de type déclaratif. Ce langage est centré sur la notion d'un espace de recherche, c'est-à-dire un segment textuel déterminé à partir de l'indicateur, espace dans lequel les indices complémentaires doivent être recherchés. [...] La partie Déclaration d'un Espace de Recherche E permet de construire un segment textuel, l'espace de recherche, en appliquant différentes opérations sur la structure du texte construite par le moteur d'exploration contextuelle (voir § 3.2.1).

Il est possible de construire plusieurs espaces de recherche dans une même règle. Une dizaine d'opérations ont été définies pour construire un espace de recherche à partir de la structure d'un texte.

3.2. *Moteur d'exploration contextuelle*
3.2.1. *Analyseur de texte*
3.2.2. *Exécuteur*
3.3. *Agents spécialisés*
3.4. *Plate-forme logicielle ContextO*

4. Un exemple d'agent spécialisé : le résumeur - filtreur

[...] - l'extrait fourni par le système, composé de phrases extraites du texte source, doit être compris par un lecteur, ce qui signifie qu'il doit être lisible par un lecteur intéressé par le thème dont traite le texte source, mais que l'extrait n'a pas pour finalité d'être publié en l'état ; [...] Nous présentons, ci-après, les caractéristiques des ressources linguistiques qui ont été développées pour attribuer des étiquettes sémantiques aux phrases du texte à résumer et comment ces étiquettes sont exploitées pour construire un ou plusieurs extraits.

4.1. ***Etiquettes sémantiques***
4.1.1. *Enoncés structurants*
[...] Nous reconnaissons ces énoncés par la co-présence d'un déictique (le présent document, nous ...) ou d'une formulation impersonnelle (il faut ..., il est utile de ...), d'un présentatif (commencer l'étude, présenter, expliquerons, montrerons) et de marqueurs d'intégration linéaire ;

4.1.2. *Définitions*
4.1.3. *Connaissances causales pour le filtrage et le résumé*

Parmi les relations conceptuelles permettant de structurer les connaissances, et donc les informations pertinentes que l'on peut souhaiter extraire d'un texte et insérer dans un résumé, on peut considérer que les relations causales occupent une place relativement privilégiée. En effet, l'intelligibilité et la maîtrise d'un grand nombre de phénomènes (qu'ils soient naturels, sociaux, psychologiques ou économiques) passent par la recherche et l'analyse des liens de causalité entre faits, entre phénomènes, entre événements.

4.2. *Filtrage des phrases étiquetées*

[...] Ainsi pour un certain type de recherche, des énoncés " conclusifs " sont considérés comme plus importants que les énoncés " d'annonce thématique " lorsqu'ils se trouvent dans la dernière section du texte.[...] La stratégie d'exploration SE précise l'ordre d'exploration des sections et la profondeur d'exploration (et donc de sélection) du texte.[...] Pour chaque section du texte, il est possible de préciser un profil de filtrage P.[...] Le deuxième problème concerne la présence des marqueurs d'intégration linéaire comme en premier lieu, en second lieu, etc. dans une des phrases sélectionnées. Dans ce cas également, l'application d'heuristiques de repérage qui exploitent la structure hiérarchique du texte et des listes finies de termes, donne des résultats satisfaisants. Le dernier problème, non résolu, est celui de la rupture argumentative du texte. Par exemple, si une phrase, étiquetée comme conclusive est sélectionnée, il est impossible de savoir à quelle phrase, étiquetée comme hypothèse, elle se réfère, puisque le système ne construit aucune représentation conceptuelle du texte source.

5. Conclusion

[...] Plutôt que de vouloir simuler le travail d'un résumeur professionnel en produisant un résumé indépendant du texte source, nous avons cherché à construire un nouvel "objet textuel" qui articule des données textuelles "décorées" et des procédures de fouille de ces données. Enfin, nous pensons que l'architecture de la plate-forme Filtext, en privilégiant le concept de composants logiciels et d'agents spécialisés, la rend apte à accueillir différents types de traitement linguistique car il devient possible de construire de nouvelles bases de marqueurs linguistiques adaptés à de nouvelles tâches d'étiquetage sémantique. Cette plate-forme vise ainsi à faciliter les étapes d'acquisition et de modélisation des connaissances linguistiques en proposant des formats et des langages de représentation des données, des outils de consultation, de manipulation, de recherche et d'analyse, etc.

Exemple 2

Résumé automatique de l'article « L'etnocentrisme» avec un profil du type « résumé-auteur » et un seuil de filtrage limité à 10 % du texte source. Les titres, les sections, les paragraphes et les phrases ont été automatiquement délimités. Pour ce type de profil, les titres de chaque section du texte original sont systématiquement placés dans le résumé, même si aucune phrase de la section n'a été sélectionnée. Le temps de traitement sur une station de type micro-ordinateur est d'environ 1 secondes. Aucune post-édition manuelle n'a été effectuée à l'exception de la mise aux normes rédactionnelles.

Résumé

L' ETHNOCENTRISME par *Claude Lévi-Stauss*

Il semble que la diversité des cultures soit rarement apparue aux hommes pour ce qu'elle est un phénomène naturel, résultant des rapports directs ou indirects entre les sociétés : ils y ont plutôt vu une sorte de monstruosité ou de scandale; […]

L'humanité cesse aux frontières de la tribu, du groupe linguistique, parfois même du village, à tel point qu'un grand nombre de populations dites primitives se désignent elles mêmes d'un nom qui signifie les « hommes » (ou parfois - dirons-nous avec plus de discrétion? - les « bons » les « excellents » les « complets »), impliquant ainsi que les autres tribus, groupes ou villages ne participent pas des vertus ou même de la nature humaines, mais sont tout au plus composés de « mauvais » de « méchants » de « singes de terre » ou d « oeufs de pou ».

On va souvent jusqu'à priver l'étranger de ce dernier degré de réalité en en faisant un faisant un « fantôme» ou une « apparition ». […]

Pris entre la double tentation de condamner des expériences qui le heurtent effectivement, et de nier des différences qu'il ne comprend pas intellectuellement, l'homme moderne s'est livré à cent spéculations philosophiques et sociologiques pour établir de vains compromis entre ces pôles contradictoires, et rendre compte de la diversité des cultures tout en cherchant à supprimer ce qu'elle conserve pour lui de scandaleux et de choquant.

Mais , si différentes et parfois si bizarres qu'elles puissent être, toutes ces spéculations se ramènent en fait à une seule recette, que le terme de faux évolutionnisme est sans doute le mieux apte à caractériser. En quoi consiste-t-elle? Très exactement, il s' agit d' une tentative pour supprimer la diversité des cultures tout en feignant de la reconnaître pleinement. […]

Cette définition peut paraître sommaire quand on a présentes à l'esprit les immenses conquêtes du darwinisme. Mais celui-ci n'est pas en cause, car l'évolutionnisme biologique et le pseudo évolutionnisme que nous avons ici en vue sont deux doctrines très différentes.

Texte source

L' ETHNOCENTRISME par *Claude Lévi-Stauss*

Il semble que la diversité des cultures soit rarement apparue aux hommes pour ce qu'elle est un phénomène naturel, résultant des rapports directs ou indirects entre les sociétés : ils y ont plutôt vu une sorte de monstruosité ou de scandale ; dans ces matières, le progrès de la connaissance n'a pas tant consisté à dissiper cette illusion au profit d'une vue plus exacte, qu'à l'accepter ou à trouver le moyen de s'y résigner.

L'attitude la plus ancienne, et qui repose sans doute sur des fondements psychologiques solides puisqu'elle tend à réapparaître chez chacun de nous quand nous sommes placés dans une situation inattendue, consiste à répudier purement et simplement les formes culturelles morales, religieuses, sociales, esthétiques, qui sont les plus éloignées de celles auxquelles nous nous identifions. « Habitudes de sauvages », « Cela n'est pas de chez nous », « on ne devrait pas permettre cela », etc., autant de réactions grossières qui traduisent ce même frisson, cette même répulsion en présence de manières de vivre, de croire ou de penser qui nous sont étrangères. Ainsi l'Antiquité confondait-elle tout ce qui ne participait pas de la culture grecque (puis gréco-romaine) sous le même nom de barbare; la civilisation occidentale a ensuite utilisé le terme de sauvage dans le même sens. Or, derrière

ces épithètes se dissimule un même jugement : il est probable que le mot barbare se réfère étymologiquement à la confusion et à l'inarticulation du chant des oiseaux, opposées à la valeur signifiante du langage humain; et sauvage, qui veut dire « de la forêt », évoque aussi un genre de vie animal, par opposition à la culture humaine. Dans les deux cas, on refuse d'admettre le fait même de la diversité culturelle ; on préfère rejeter hors de la culture, dans la nature, tout ce qui ne se conforme pas à la norme sous laquelle on vit.

Ce point de vue naïf, mais profondément ancré chez la plupart des hommes, n'a pas besoin d'être discuté puisque cette brochure - avec toutes celles de la même collection - en présenté justement la réfutation. Il suffira de remarquer ici qu'il recèle un paradoxe assez significatif. Cette attitude de pensée, au nom de laquelle on rejette les « sauvages » (ou tous ceux qu'on choisit de considérer comme tels) hors de l'humanité, est justement l'attitude la plus marquante et la plus distinctive dé ces sauvages mêmes. On sait, en effet, que la notion d'humanité, englobant, sans distinction de race ou de civilisation, toutes les formes de l'espèce humaine, est d'apparition fort tardive et d'expansion limitée. Là même où elle semble avoir atteint son plus haut développement, il n'est nullement certain - l'histoire récente le prouve -qu'elle soit à l'abri dès équivoques ou des régressions. Mais, pour de vastes fractions de l'espèce humaine et pendant des dizaines de millénaires, cette notion parait être totalement absente. L'humanité cesse aux frontières de la tribu, du groupe linguistique, parfois même du village ; à tel point qu'un grand nombre de populations dites primitives se désignent elles-mêmes d'un nom qui signifie les "hommes" (ou parfois - dirons-nous avec plus de discrétion? - les « bons », les « excellents », les « complets »), impliquant ainsi que les autres tribus, groupes ou villages ne participent pas des vertus ou même de la nature humaines, mais sont tout au plus composés de « mauvais », de « méchants », de « singes dé terre » ou d' « oeufs de pou ». On va souvent jusqu'à priver l'étranger de ce dernier degré dé réalité en en faisant un « fantôme" » ou une « apparition ». Ainsi se réalisent de curieuses situations où deux interlocuteurs se donnent cruellement la répliqué. Dans les grandes Antilles, quelques années après la découverte de l'Amérique, pendant que les Espagnols envoyaient des commissions s d'enquête pour rechercher Si les indigènes avaient ou non une âme, ces derniers s'employaient à immerger dés blancs prisonniers, afin de vérifier, par une surveillance prolongé, si leur cadavre était ou non sujet à la putréfaction.

Cette anecdote à la fois baroque et tragique illustre bien le paradoxe du relativise culturel (que nous retrouverons d'ailleurs sous d'autres formes) : c'est dans la mesure même où l'on prétend établir une discrimination entre les cultures et les coutumes que l'on s'identifie le plus complètement avec celles qu'on essaye de nier En refusant l'humanité à ceux qui apparaissent comme les plus « sauvages » ou « barbares » de ses représentants, on ne fait que leur emprunter une de leurs attitudes typiques. Le barbare, c'est d'abord l'homme qui croit à la barbarie.

Sans doute les grands systèmes philosophiques et religieux de l'humanité - qu'il s'agisse du bouddhisme, du christianisme ou de l'Islam, des doctrines stoïcienne, kantienne ou marxiste se sont-ils constamment élevés contre cette aberration. Mais la simple proclamation de l'égalité naturelle entre tous les hommes, et de la fraternité qui doit les unir sans distinction de race ou de culture, a quelque chose de décevant pour l'esprit, parce qu'elle néglige une diversité de fait qui s'impose a l'observation, et dont il ne suffit pas de dire qu'elle n'affecte pas le fond du problème pour que l'on soit théoriquement et pratiquement autorisé à faire comme si elle n'existait pas. Ainsi, le préambule à la seconde déclaration de l'Unesco sur le problème des races remarque judicieusement que ce qui convainc l'homme de la rue que les races existent, c'est l'évidence immédiate de ses sens quand il aperçoit ensemble un Africain, un Européen, un Asiatique et un Indien américain.

Les grandes déclarations des droits de l'homme ont, elles aussi, cette force et cette faiblesse d'énoncer un idéal trop souvent oublieux du fait que l'homme ne réalise pas sa nature dans une humanité abstraite, mais dans des cultures traditionnelles dont les changements les plus révolutionnaires laissent subsister des pans entiers, et s'expliquent eux-mêmes en fonction d'une situation strictement définie dans le temps et dans l'espace. Pris entre la double tentation de condamner des expériences qui le heurtent effectivement, et de nier des différences qu'il ne comprend pas intellectuellement, l'homme moderne s'est livré à cent spéculations philosophiques et sociologiques pour établir de vains compromis entre ces pôles contradictoires, et rendre compte de la diversité des cultures tout en cherchant à supprimer ce qu'elle conserve pour lui de scandaleux et de choquant.

Mais, si différentes et parfois si bizarres qu'elles puissent être, toutes ces spéculations se ramènent en

fait à une seule recette, que le terme de faux évolutionnisme est sans doute le, mieux apte à caractériser . En quoi consiste-t-elle ? Très exactement, il s'agit d'une tentative pour supprimer la diversité des cultures tout en feignant de la reconnaître pleinement. Car, si l'on traite les différents états où se trouvent les sociétés humaines, tant anciennes que lointaines, comme des stades ou des étapes d'un développement unique qui, partant du même point, doit les faire converger vers le même but, on voit bien que la diversité n'est plus qu'apparente. L'humanité devient une et identique à elle-même ; seulement, cette unité et cette identité ne peuvent se réaliser que progressivement, et la variété des cultures illustre les moments d'un processus qui dissimule une réalité plus profonde ou en retarde la manifestation.

Cette définition peut paraître sommaire quand on a présentes à l'esprit les immenses conquêtes du darwinisme. Mais celui-ci n'est pas en cause, car l'évolutionnisme biologique et le pseudo-évolutionnisme que nous avons ici en vue sont deux doctrines très différentes. La première est née comme une vaste hypothèse de travail, fondée sur des observations où la part laissée à l'interprétation est fort petite. Ainsi, les différents types constituant la généalogie du cheval peuvent être rangés dans une série évolutive pour deux raisons la première est qu'il faut un cheval pour engendrer un cheval la seconde, que des couches de terrain superposées, donc historiquement de plus en plus anciennes, contiennent des squelettes qui varient de façon graduelle depuis la forme la plus récente jusqu'à la plus archaïque. Il devient ainsi hautement probable que Hipparion soit l'ancêtre réel de Equus caballus. Le même raisonnement s'applique sans doute à l'espèce humaine et à ses races. Mais quand on passe des faits biologiques aux faits de culture, les choses se compliquent singulièrement. On peut recueillir dans le sol des objets matériels, et constater que, selon la profondeur des couches géologiques, la forme ou la technique de fabrication d'un certain type d'objets varie progressivement. Et pourtant une hache ne donne pas physiquement naissance à une hache, à la façon d'un animal. Dire, dans ce dernier cas, qu'une hache a évolué à partir d'une autre constitue donc une formule métaphorique et approximative, dépourvue de la rigueur scientifique qui s'attache à l'expression similaire appliquée aux phénomènes biologiques. Ce qui est vrai d'objets matériels dont la présence physique est attestée dans le sol, pour des époques déterminables, l'est plus encore pour les institutions, les croyances, les goûts, dont le passé nous est généralement inconnu. La notion d'évolution biologique correspond à une hypothèse dotée d'un des plus hauts coefficients de probabilité qui puissent se rencontrer dans le domaine des sciences naturelles ; tandis que la notion d'évolution sociale ou culturelle n'apporte, tout au plus, qu'un procédé séduisant, mais dangereusement commode, de présentation des faits.

D'ailleurs, cette différence, trop souvent négligée, entre le vrai et le faux évolutionnisme s'explique par leurs dates d'apparition respectives. Sans doute, l'évolutionnisme sociologique devait recevoir de l'évolutionnisme biologique une impulsion vigoureuse ; mais il lui est antérieur dans le temps. Sans remonter jusqu'aux conceptions antiques, reprises par Pascal, assimilant l'humanité à un être vivant qui passe par les stades successifs de l'enfance, de l'adolescence et de la maturité, c'est au XVIIIe siècle qu'on voit fleurir les schémas fondamentaux qui seront, par la suite, l'objet de tant de manipulations les "spirales" de Vico, ses "trois âges" annonçant les "trois états" de Comte, l' « escalier » de Condorcet. Les deux fondateurs de l'évolutionnisme social, Spencer et Tylor, élaborent et publient leur doctrine avant L'origine des espèces ou sans avoir lu cet ouvrage. Antérieur à l'évolutionnisme biologique, théorie scientifique, l'évolutionnisme social n'est, trop souvent, que le maquillage faussement scientifique d'un vieux problème philosophique dont il n'est nullement certain que l'observation et l'induction puissent un jour fournir la clef

Exemple 3

Résumé automatique de l'article « Privée d'émotions, la mémoire flanche»[51] *publié dans la revue Pour la Science, avec un profil du type « résumé-auteur » et un seuil de filtrage limité à 20 % du texte source. Les titres, les sections, les paragraphes et les phrases ont été automatiquement délimités. Pour ce type de profil, les titres de chaque section du texte original sont systématiquement placés dans le résumé, même si aucune phrase de la section n'a été sélectionnée. Le temps de traitement sur une station de type micro-ordinateur est d'environ 1 secondes. Aucune post-édition manuelle n'a été effectuée à l'exception de la mise aux normes rédactionnelles.*

Résumé

Privée d'émotions, la mémoire flanche

Lobotomies frontales.
Ainsi naquit la psychochirurgie, thérapeutique audacieuse consistant à ôter une partie du cerveau pour traiter les maladies mentales. [...].
Inadaptation émotionnelle.
Hippocampe.
[...] Ainsi découvrait-on que l'hippocampe, dont la fonction était jusqu'alors inconnue, était en fait nécessaire pour la formation des souvenirs nouveaux.
[...]En premier lieu, l'essor des neurosciences cognitives, tout au long du XXe siècle, a considérablement accru notre savoir sur le cerveau, fournissant ainsi les bases indispensables pour aborder la complexité des phénomènes affectifs. En second lieu, des perspectives entièrement nouvelles ont émergé grâce à de récents progrès techniques. [...]Enfin, plusieurs chercheurs contemporains, ouvrant la voie des neurosciences affectives, ont su réactualiser l'idée ancienne selon laquelle les émotions sont en réalité la cheville ouvrière du bon fonctionnement de nombre de nos facultés, adaptation sociale, raisonnement, prise de décision, ou mémoire.
Imagerie fonctionnelle.
[...]Mais le résultat important était le suivant : plus l'amygdale située du côté droit du cerveau avait été active pendant la présentation des films, meilleurs étaient les souvenirs des films négatifs. A l'inverse, l'activité de cette amygdale ne prédisait en rien la qualité des souvenirs pour les films neutres. Cette étude fournit donc la preuve d'un lien entre l'activité de l'amygdale droite pendant l'encodage d'informations riches en émotions et leur rétention ultérieure.

Texte étiqueté

Texte automatiquement balisé et étiqueté sémantiquement (avec les étiquettes utilisées pour la construction du résumé).Les étiquettessont baliséesavec la baliseXML <et> qui contient un ou plusieurs attributs.

<h1><t1> Privée d' émotions, la mémoire flanche
</t1> <a> Emotions et souvenirs se forment dans la même partie du cerveau.<a> Mais l' impact de cette découverte a été négligé durant le XXe siècle.<a> Leurs relations commencent seulement à être étudiées grâce à l' émergence des neurosciences affectives.
<a> Nos émotions jouent un rôle essentiel dans notre mémoire autobiographique.<a> Mais l' étude des mécanismes cérébraux qui les gouvernent a longtemps été négligée par les neurosciences.<a> Toute émotion affecte simultanément notre corps, notre comportement, nos sentiments et notre mémoire.<a> Autant d' aspects difficiles à mesurer objectivement et à évaluer simultanément.<a> Face à une même situation, les réponses émotionnelles varient en fonction de l' individu, de son tempérament et de son environnement physique et social.<a> Une versatilité qui complique encore leur évaluation.
<a> A la fin du XIXe siècle, Sigmund Freud attribuait pourtant aux émotions une influence déterminante dans le développement des individus.<a> Le psychologue William James soulignait déjà leur importance pour le bon fonctionnement de la mémoire.<a> " Se souvenir de tout serait aussi fâcheux que ne se souvenir de rien ", insistait -il.<a> Le cerveau doit effectuer une sélection.<a> Il le fait en fonction de la valeur affective qu' un événement revêt pour nous.<a> Tout au long du XXe siècle, les émotions ont conservé une place centrale au sein de la psychologie.<a> En revanche, la compréhension de leur organisation cérébrale ne s' est imposée comme un enjeu majeur pour les neurosciences qu' au cours de ces dernières années.<a> Cette lente évolution est bien illustrée par l' histoire de nos connaissances de deux régions présentes dans chaque

[51] Ce texte est extrait du jeu d'essai utilisé par G. Jacquet pour son mémoire réalisé dans le cadre de la Maîtrise de Sciences et Techniques de sociologie de l'Université Paris-Sorbonne.

hémisphère du cerveau, les lobes frontal * et temporal *.<a> Découverte vers le milieu du XIXe siècle, leur importance pour les émotions n' a longtemps suscité qu' un intérêt marginal.<a> A l' opposé, leur implication dans la mémoire, identifiée plus tard, a immédiatement suscité un nombre considérable de travaux.
<a> Dès 1848, John Harlow, médecin d' une petite ville de l' Est américain, décrivait le cas spectaculaire de Phineas Gage (voir l' encadré : " Comment la barre de Phineas Gage révéla le rôle du lobe frontal ") et remarquait le rôle des lobes frontaux dans le contrôle des émotions.<a> Mais la localisation cérébrale des fonctions mentales, idée largement acceptée aujourd'hui, suscitait alors de vives controverses.<a> Elle ne s' imposera lentement qu' après la démonstration, dans les années 1860-1870, du rôle de certaines aires de l' hémisphère gauche dans le langage.<a> En 1848, le cas de Phineas Gage fut donc plutôt perçu comme un encouragement pour la neurochirurgie balbutiante du moment.<a> Il s' avérait en effet possible d' ôter une grosse portion du cerveau (en cas de tumeur par exemple) sans provoquer la mort, ni altérer aucune des fonctions psychologiques " majeures ", perception, motricité, langage, intelligence, ou mémoire.<a> Dans les décennies qui suivirent, la neurochirurgie fit d' énormes progrès, aidée, paradoxalement, par la Première Guerre mondiale et ses nombreux blessés.<a> Le cas de Phineas Gage ne fut plus guère évoqué dans la littérature médicale.<a> Le lobe frontal fascinait les chercheurs de la première moitié du XXe siècle, mais en tant que siège des fonctions intellectuelles " supérieures " spécifiques aux primates.<a> Cette région du cerveau est en effet si développée chez l' homme qu' elle occupe à elle seule un tiers du cortex.
<h2><t2> Lobotomies frontales
</t2> <a> Lors du 2 e Congrès international de neurologie, à Londres, en 1935, le neuropsychologue Carlyle Jacobsen et le neurochirurgien John Fulton présentèrent leurs travaux sur les effets d' une ablation des lobes frontaux chez des chimpanzés.<a> Placées face à deux coupelles identiques, les deux femelles opérées, Becky et Lucy, étaient incapables de retrouver laquelle dissimulait une friandise, bien que la récompense ait été cachée sous leurs yeux quelques secondes auparavant seulement.<a> Cette étude pionnière ouvrit la voie vers la compréhension des relations entre lobe frontal et mémoire.<a> Jacobsen et Fulton mentionnèrent également des changements surprenants de comportement chez les animaux opérés.<a> Lucy, à l' origine calme et tempérée, devint plus coléreuse et violente.<a> A l' inverse, Becky, irascible avant l' opération, semblait d' une indéfectible bonne humeur après.<a> Bien qu' apportant une nouvelle preuve du lien entre lobe frontal et émotions, ces anecdotes eurent peu de répercussions sur la recherche fondamentale.<a> Le cas de Becky eut, en revanche, une conséquence inattendue en psychiatrie.<a> Le neurologue portugais Egas Moniz allait, dès son retour du congrès de Londres, pratiquer des lobotomies frontales chez des patients psychotiques.<a> Ainsi naquit la psychochirurgie, thérapeutique audacieuse consistant à ôter une partie du cerveau pour traiter les maladies mentales.<et a0=Conclusion_2 />
<a> En dépit de ses effets secondaires, ce traitement radical allait rapidement être appliqué à des milliers de patients dans le monde entier.<a> Et même valoir un prix Nobel à Moniz en 1949, avant que son usage abusif ne lui fasse une sinistre réputation, et que l' arrivée des neuroleptiques dans les années 1950 ne le rende obsolète (1).<a> Le cortex préfrontal (ou partie avant du lobe frontal, celle qui fut touchée chez Gage), cible des lobotomies, allait cependant rester la structure la moins bien connue du cerveau jusque dans les années 1970 !<a> Les nombreux travaux suscités par les déficits d' apprentissage rapportés par Jacobsen et Fulton établiront seulement après cette date le rôle de la partie latérale du cortex préfrontal dans la mémoire de travail, celle qui nous permet de garder en tête une information, un numéro de téléphone par exemple, juste le temps de l' utiliser.
</h2><h2><t2> Inadaptation émotionnelle
</t2> <a> En ce qui concerne les lobes temporaux, les premiers indices de leur implication dans les émotions remontent à des observations faites en 1888.<a> Mais ils tombèrent dans l' oubli jusqu' à la découverte du psychologue Heinrich Klüver et du neurologue Paul Bucy de l' université de Chicago en 1938.<a> Etudiant des singes porteurs de lésions des lobes temporaux, ces auteurs furent surpris par les comportements émotionnels inadaptés de ces animaux.<a> Ils approchaient, manipulaient ou portaient à la bouche, de façon compulsive, tout ce qu' on leur présentait.<a> Ils paraissaient également ne plus ressentir aucune peur, même face à un serpent.<a> Une attitude qui leur aurait été fatale dans leur milieu naturel.<a> En 1956, on établit que ce syndrome, dit de Klüver et Bucy, est principalement dû à l' atteinte de la région antérieure de la partie médiane du lobe temporal, celle qui contient l' amygdale (2), une petite structure en amande.<a> On ne s' interrogera plus guère ensuite sur les fonctions exactes des lobes temporaux pour les émotions.<a> Car commence alors la saga, toujours d' actualité, de leurs relations avec la mémoire.
<a> En 1957 et 1958, la psychologue Brenda Milner de l' institut neurologique de Montréal décrit les cas dramatiques de patients devenus amnésiques à la suite d' une ablation chirurgicale de l' un ou des deux lobes temporaux.<a> Elle a observé cet effet inattendu chez quatre patients, deux parmi les trente opérés par William Scoville aux Etats-Unis, et à nouveau deux parmi plus de quatre-vingt - dix patients opérés par Wilder Penfield au Canada.<a> La postérité retiendra l' un d' entre eux, qui deviendra célèbre sous les initiales H. M. Une large partie de nos connaissances actuelles sur l' organisation cérébrale de la mémoire repose sur

lui.<a> En 1953, ce jeune homme de 27 ans subit une ablation des deux lobes temporaux pour ôter le foyer d' une épilepsie très invalidante et rebelle à tout traitement médicamenteux.<a> L' opération soulagea l' épilepsie.<a> Mais elle provoqua une amnésie profonde qui perdure aujourd'hui.
</h2><h2><t2> Hippocampe
</t2> <a> Depuis près de cinquante ans maintenant, H. M. oublie au fur et à mesure tous les événements de sa vie quotidienne.<a> Or sa lésion, contrairement à celle de la plupart des patients opérés en même temps que lui, s' étendait au point d' inclure non seulement l' amygdale, mais aussi une large portion de l' hippocampe.<a> Ainsi découvrait -on que l' hippocampe, dont la fonction était jusqu' alors inconnue, était en fait nécessaire pour la formation des souvenirs nouveaux.<et a0=Conclusion_2/ ><a> Cette découverte allait motiver un nombre considérable de travaux expérimentaux.<a> Un intense effort qui a abouti aujourd'hui à une remarquable connaissance des bases cérébrales de la mémoire épisodique et sémantique, celles, respectivement, des événements personnellement vécus et des connaissances générales sur le monde.
<a> En revanche, H. M. n' a jamais été l' objet d' une évaluation approfondie sur le plan émotionnel.<a> Seules quelques anecdotes ont été rapportées à son sujet qui suggèrent un appauvrissement émotionnel, différent mais néanmoins proche de celui des singes de Klüver et Bucy, après le même type de lésion.<a> En dehors de quelques accès d' irritabilité, H. M. a en effet été décrit d' une humeur étonnamment placide, parlant sur un ton monotone, et témoignant d' une résistance inhabituelle à la douleur, la faim ou la fatigue.
<a> Les émotions sont aujourd'hui l' objet d' un intérêt grandissant en neurosciences, comme en témoigne la croissance exponentielle des publications dans ce domaine depuis la fin des années 1990.<a> Ce rebondissement s' explique par la convergence d' au moins trois facteurs.<a> En premier lieu, l' essor des neurosciences cognitives, tout au long du XXe siècle, a considérablement accru notre savoir sur le cerveau, fournissant ainsi les bases indispensables pour aborder la complexité des phénomènes affectifs.<a> En second lieu, des perspectives entièrement nouvelles ont émergé grâce à de récents progrès techniques.<a> Notamment, l' imagerie fonctionnelle nous donne aujourd'hui la possibilité de voir le cerveau humain normal en action, alors qu' autrefois nous devions nous contenter des indices fournis par le cerveau lésé.<a> Enfin, plusieurs chercheurs contemporains, ouvrant la voie des neurosciences affectives, ont su réactualiser l' idée ancienne selon laquelle les émotions sont en réalité la cheville ouvrière du bon fonctionnement de nombre de nos facultés, adaptation sociale, raisonnement, prise de décision, ou mémoire.<a> Les neurosciences affectives offrent déjà un aperçu des mécanismes cérébraux qui gouvernent l' influence des émotions sur la mémoire.
<a> Les travaux actuels concernent principalement les deux amygdales (situées chacune à l' avant de la partie médiane du lobe temporal).<a> Chez le rat, différentes équipes dont celles de Michael Davis à Yale et de Joseph LeDoux à New York, ont réussi à démonter la machinerie complexe qui contrôle les peurs conditionnées (3).<a> Il s' agit de ce phénomène, commun à nombre d' espèces, de l' escargot de mer à l' homme, par lequel un stimulus neutre associé à un événement désagréable acquiert ensuite le pouvoir de déclencher à lui seul une réaction de peur.<a> Parmi les différents noyaux composant l' amygdale, le noyau latéral reçoit des informations des régions sensorielles comme le cortex visuel.<a> Il les transmet au noyau central relié aux centres cérébraux qui déclenchent les réactions dites autonomes, comme l' accélération du rythme cardiaque.<a> Ce circuit assure l' apprentissage des peurs conditionnées.<a> Il influence des structures voisines comme l' hippocampe qui restituent les souvenirs liés à ces peurs.
</h2><h2><t2> Imagerie fonctionnelle
</t2> <a> Les expériences chez le rongeur ont ouvert la voie à l' exploration du comportement plus riche des primates.<a> Chez le singe, la destruction sélective des seules cellules des amygdales suffit à perturber l' utilisation de l' ensemble du savoir émotionnel et social des animaux (4).<a> Chez l' homme, leur importance pour la mémoire émotionnelle a été particulièrement bien démontrée par une étude utilisant l' imagerie fonctionnelle par TEP (5) (Tomographie par émission de positons, voir l' article de Francis Eustache).<a> Le neuropsychologue Larry Cahill et ses collègues de l' université de Californie ont mesuré l' activité du cerveau de huit volontaires pendant qu' ils regardaient des documentaires relatant soit des événements neutres, soit des images très négatives (de crimes violents, par exemple).<a> Trois semaines plus tard, les sujets se souvenaient beaucoup mieux des films négatifs que des films neutres, reflétant l' amélioration de la mémoire par les émotions.<a> Mais le résultat important était le suivant : plus l' amygdale située du côté droit du cerveau avait été active pendant la présentation des films, meilleurs étaient les souvenirs des films négatifs.<et a0=Recapitulation_1 /><a> A l' inverse, l' activité de cette amygdale ne prédisait en rien la qualité des souvenirs pour les films neutres.<a> Cette étude fournit donc la preuve d' un lien entre l' activité de l' amygdale droite pendant l' encodage d' informations riches en émotions et leur rétention ultérieure.<et a0=Annonce_thématique a1=Conclusion_2/ >
<a> En accord avec cette conclusion, les patients dont l' amygdale a été endommagée présentent une mémoire correcte mais insensible à l' effet accélérateur des émotions.<a> A l' inverse, cet effet reste présent chez les amnésiques dont l' amygdale est intacte, ainsi que chez les personnes atteintes de la maladie d' Alzheimer.<a> Ces patients oublient moins les événements à forte connotation émotionnelle que les

autres.<a> Une découverte qui pourrait se révéler utile pour améliorer le soutien quotidien apporté à ces malades.</p>
</h2><h2><t2> Martine Meunier
</t2></h2><h2><t2>
</t2></h2><h2><t2>
</t2></h2></h1>

Texte source

Privée d'émotions, la mémoire flanche

Emotions et souvenirs se forment dans la même partie du cerveau. Mais l'impact de cette découverte a été négligé durant le XXe siècle. Leurs relations commencent seulement à être étudiées grâce à l'émergence des neurosciences affectives.

Nos émotions jouent un rôle essentiel dans notre mémoire autobiographique. Mais l'étude des mécanismes cérébraux qui les gouvernent a longtemps été négligée par les neurosciences. Toute émotion affecte simultanément notre corps, notre comportement, nos sentiments et notre mémoire. Autant d'aspects difficiles à mesurer objectivement et à évaluer simultanément. Face à une même situation, les réponses émotionnelles varient en fonction de l'individu, de son tempérament et de son environnement physique et social. Une versatilité qui complique encore leur évaluation.

A la fin du XIXe siècle, Sigmund Freud attribuait pourtant aux émotions une influence déterminante dans le développement des individus. Le psychologue William James soulignait déjà leur importance pour le bon fonctionnement de la mémoire . *« Se souvenir de tout serait aussi fâcheux que ne se souvenir de rien »*, insistait-il. Le cerveau doit effectuer une sélection. Il le fait en fonction de la valeur affective qu'un événement revêt pour nous. Tout au long du XXe siècle, les émotions ont conservé une place centrale au sein de la psychologie. En revanche, la compréhension de leur organisation cérébrale ne s'est imposée comme un enjeu majeur pour les neurosciences qu'au cours de ces dernières années. Cette lente évolution est bien illustrée par l'histoire de nos connaissances de deux régions présentes dans chaque hémisphère du cerveau, les lobes frontal* et temporal*. Découverte vers le milieu du XIXe siècle, leur importance pour les émotions n'a longtemps suscité qu'un intérêt marginal. A l'opposé, leur implication dans la mémoire, identifiée plus tard, a immédiatement suscité un nombre considérable de travaux.

Dès 1848, John Harlow, médecin d'une petite ville de l'Est américain, décrivait le cas spectaculaire de Phineas Gage (voir l'encadré : « Comment la barre de Phineas Gage révéla le rôle du lobe frontal ») et remarquait le rôle des lobes frontaux dans le contrôle des émotions. Mais la localisation cérébrale des fonctions mentales, idée largement acceptée aujourd'hui, suscitait alors de vives controverses. Elle ne s'imposera lentement qu'après la démonstration, dans les années 1860-1870, du rôle de certaines aires de l'hémisphère gauche dans le langage. En 1848, le cas de Phineas Gage fut donc plutôt perçu comme un encouragement pour la neurochirurgie balbutiante du moment. Il s'avérait en effet possible d'ôter une grosse portion du cerveau (en cas de tumeur par exemple) sans provoquer la mort, ni altérer aucune des fonctions psychologiques « majeures », perception, motricité, langage, intelligence, ou mémoire. Dans les décennies qui suivirent, la neurochirurgie fit d'énormes progrès, aidée, paradoxalement, par la Première Guerre mondiale et ses nombreux blessés. Le cas de Phineas Gage ne fut plus guère évoqué dans la littérature médicale. Le lobe frontal fascinait les chercheurs de la première moitié du XXe siècle, mais en tant que siège des fonctions intellectuelles « supérieures » spécifiques aux primates. Cette région du cerveau est en effet si développée chez l'homme qu'elle occupe à elle seule un tiers du cortex.

Lobotomies frontales.

Lors du 2e Congrès international de neurologie, à Londres, en 1935, le neuropsychologue Carlyle Jacobsen et le neurochirurgien John Fulton présentèrent leurs travaux sur les effets d'une ablation des lobes frontaux chez des chimpanzés. Placées face à deux coupelles identiques, les deux femelles opérées, Becky et Lucy, étaient incapables de retrouver laquelle dissimulait une friandise, bien que la récompense ait été cachée sous leurs yeux quelques secondes auparavant seulement. Cette étude pionnière ouvrit la voie vers la compréhension des relations entre lobe frontal et mémoire. Jacobsen et Fulton mentionnèrent également des changements surprenants de comportement chez les animaux opérés. Lucy, à l'origine calme et tempérée, devint plus coléreuse et violente. A l'inverse, Becky, irascible avant l'opération, semblait d'une indéfectible bonne humeur après. Bien qu'apportant une nouvelle preuve du lien entre lobe frontal et émotions, ces anecdotes eurent peu de répercussions sur la recherche fondamentale. Le cas de Becky eut, en revanche, une conséquence inattendue en psychiatrie. Le neurologue portugais Egas Moniz allait, dès son retour du congrès de Londres, pratiquer des lobotomies frontales chez des patients psychotiques. Ainsi naquit la psychochirurgie, thérapeutique audacieuse consistant à ôter une partie du cerveau pour traiter les maladies mentales.

En dépit de ses effets secondaires, ce traitement radical allait rapidement être appliqué à des milliers de patients dans le monde entier. Et même valoir un prix Nobel à Moniz en 1949, avant que son usage abusif ne lui fasse une sinistre réputation, et que l'arrivée des neuroleptiques dans les années 1950 ne le rende obsolète(1). Le cortex préfrontal (ou partie avant du lobe frontal, celle qui fut touchée chez Gage), cible des lobotomies, allait cependant rester la structure la moins bien connue du cerveau jusque dans les années 1970 ! Les nombreux travaux suscités par les déficits d'apprentissage rapportés par Jacobsen et Fulton établiront seulement après cette date le rôle de la partie latérale du cortex préfrontal dans la mémoire de travail, celle qui nous permet de garder en tête une information, un numéro de téléphone par exemple, juste le temps de l'utiliser.

Inadaptation émotionnelle.

En ce qui concerne les lobes temporaux, les premiers indices de leur implication dans les émotions remontent à des observations faites en 1888. Mais ils tombèrent dans l'oubli jusqu'à la découverte du psychologue Heinrich Klüver et du neurologue Paul Bucy de l'université de Chicago en 1938. Etudiant des singes porteurs de lésions des lobes temporaux, ces auteurs furent surpris par les comportements émotionnels inadaptés de ces animaux. Ils approchaient, manipulaient ou portaient à la bouche, de façon compulsive, tout ce qu'on leur présentait. Ils paraissaient également ne plus ressentir aucune peur, même face à un serpent. Une attitude qui leur aurait été fatale dans leur milieu naturel. En 1956, on établit que ce syndrome, dit de Klüver et Bucy, est principalement dû à l'atteinte de la région antérieure de la partie médiane du lobe temporal, celle qui contient l'amygdale(2), une petite structure en amande.

On ne s'interrogera plus guère ensuite sur les fonctions exactes des lobes temporaux pour les émotions. Car commence alors la saga, toujours d'actualité, de leurs relations avec la mémoire.

En 1957 et 1958, la psychologue Brenda Milner de l'institut neurologique de Montréal décrit les cas dramatiques de patients devenus amnésiques à la suite d'une ablation chirurgicale de l'un ou des deux lobes temporaux. Elle a observé cet effet inattendu chez quatre patients, deux parmi les trente opérés par William Scoville aux Etats-Unis, et à nouveau deux parmi plus de quatre-vingt-dix patients opérés par Wilder Penfield au Canada. La postérité retiendra l'un d'entre eux, qui deviendra célèbre sous les initiales H.M. Une large partie de nos connaissances actuelles sur l'organisation cérébrale de la mémoire repose sur lui. En 1953, ce jeune homme de 27 ans subit une ablation des deux lobes temporaux pour ôter le foyer d'une épilepsie très invalidante et rebelle à tout traitement médicamenteux. L'opération soulagea l'épilepsie. Mais elle provoqua une amnésie profonde qui perdure aujourd'hui.

Hippocampe.

Depuis près de cinquante ans maintenant, H.M. oublie au fur et à mesure tous les événements de sa vie quotidienne.

Or sa lésion, contrairement à celle de la plupart des patients opérés en même temps que lui, s'étendait au point d'inclure non seulement l'amygdale, mais aussi une large portion de l'hippocampe. Ainsi découvrait-on que l'hippocampe, dont la fonction était jusqu'alors inconnue, était en fait nécessaire pour la formation des souvenirs nouveaux. Cette découverte allait motiver un nombre considérable de travaux expérimentaux. Un intense effort qui a abouti aujourd'hui à une remarquable connaissance des bases cérébrales de la mémoire épisodique et sémantique, celles, respectivement, des événements personnellement vécus et des connaissances générales sur le monde.

En revanche, H.M. n'a jamais été l'objet d'une évaluation approfondie sur le plan émotionnel. Seules quelques anecdotes ont été rapportées à son sujet qui suggèrent un appauvrissement émotionnel, différent mais néanmoins proche de celui des singes de Klüver et Bucy, après le même type de lésion. En dehors de quelques accès d'irritabilité, H.M. a en effet été décrit d'une humeur étonnamment placide, parlant sur un ton monotone, et témoignant d'une résistance inhabituelle à la douleur, la faim ou la fatigue.

Les émotions sont aujourd'hui l'objet d'un intérêt grandissant en neurosciences, comme en témoigne la croissance exponentielle des publications dans ce domaine depuis la fin des années 1990. Ce rebondissement s'explique par la convergence d'au moins trois facteurs. En premier lieu, l'essor des neurosciences cognitives, tout au long du XXe siècle, a considérablement accru notre savoir sur le cerveau, fournissant ainsi les bases indispensables pour aborder la complexité des phénomènes affectifs. En second lieu, des perspectives entièrement nouvelles ont émergé grâce à de récents progrès techniques. Notamment, l'imagerie fonctionnelle nous donne aujourd'hui la possibilité de voir le cerveau humain normal en action, alors qu'autrefois nous devions nous contenter des indices fournis par le cerveau lésé. Enfin, plusieurs chercheurs contemporains, ouvrant la voie des neurosciences affectives, ont su réactualiser l'idée ancienne selon laquelle les émotions sont en réalité la cheville ouvrière du bon fonctionnement de nombre de nos facultés, adaptation sociale, raisonnement, prise de décision, ou mémoire. Les neurosciences affectives offrent déjà un aperçu des mécanismes cérébraux qui gouvernent l'influence des émotions sur la mémoire.

Les travaux actuels concernent principalement les deux amygdales (situées chacune à l'avant de la partie médiane du lobe temporal). Chez le rat, différentes équipes dont celles de Michael Davis à Yale et de Joseph LeDoux à New York, ont réussi à démonter la machinerie complexe qui contrôle les peurs conditionnées(3). Il s'agit de ce phénomène, commun à nombre d'espèces, de l'escargot de mer à l'homme, par lequel un stimulus

neutre associé à un événement désagréable acquiert ensuite le pouvoir de déclencher à lui seul une réaction de peur. Parmi les différents noyaux composant l'amygdale, le noyau latéral reçoit des informations des régions sensorielles comme le cortex visuel. Il les transmet au noyau central relié aux centres cérébraux qui déclenchent les réactions dites autonomes, comme l'accélération du rythme cardiaque. Ce circuit assure l'apprentissage des peurs conditionnées. Il influence des structures voisines comme l'hippocampe qui restituent les souvenirs liés à ces peurs.

Imagerie fonctionnelle.

Les expériences chez le rongeur ont ouvert la voie à l'exploration du comportement plus riche des primates. Chez le singe, la destruction sélective des seules cellules des amygdales suffit à perturber l'utilisation de l'ensemble du savoir émotionnel et social des animaux(4). Chez l'homme, leur importance pour la mémoire émotionnelle a été particulièrement bien démontrée par une étude utilisant l'imagerie fonctionnelle par TEP(5) (Tomographie par émission de positons, voir l'article de Francis Eustache). Le neuropsychologue Larry Cahill et ses collègues de l'université de Californie ont mesuré l'activité du cerveau de huit volontaires pendant qu'ils regardaient des documentaires relatant soit des événements neutres, soit des images très négatives (de crimes violents, par exemple). Trois semaines plus tard, les sujets se souvenaient beaucoup mieux des films négatifs que des films neutres, reflétant l'amélioration de la mémoire par les émotions. Mais le résultat important était le suivant : plus l'amygdale située du côté droit du cerveau avait été active pendant la présentation des films, meilleurs étaient les souvenirs des films négatifs. A l'inverse, l'activité de cette amygdale ne prédisait en rien la qualité des souvenirs pour les films neutres. Cette étude fournit donc la preuve d'un lien entre l'activité de l'amygdale droite pendant l'encodage d'informations riches en émotions et leur rétention ultérieure.

En accord avec cette conclusion, les patients dont l'amygdale a été endommagée présentent une mémoire correcte mais insensible à l'effet accélérateur des émotions. A l'inverse, cet effet reste présent chez les amnésiques dont l'amygdale est intacte, ainsi que chez les personnes atteintes de la maladie d'Alzheimer. Ces patients oublient moins les événements à forte connotation émotionnelle que les autres. Une découverte qui pourrait se révéler utile pour améliorer le soutien quotidien apporté à ces malades.

Martine Meunier

Exemple 4

Résumé

Résumé automatique de l'article de sociologie « Pourquoi le monde va changer» avec un profil du type « résumé-auteur » et un seuil de filtrage limité à 20 % du texte source. Les titres, les sections, les paragraphes et les phrases ont été automatiquement délimités. Pour ce type de profil, les titres de chaque section du texte original sont systématiquement placés dans le résumé, même si aucune phrase de la section n'a été sélectionnée. Le temps de traitement sur une station de type micro-ordinateur est d'environ 1 secondes. Aucune post-édition manuelle n'a été effectuée à l'exception de la mise aux normes rédactionnelles.

Pourquoi le monde va changer

avec les NTIC ?

La techno-science et ses impacts : Pour un nouveau contrat social

ERIE - Enseignement et Recherches Interdisciplinaires en Éthique

Université de Lausanne

12-13 novembre 1998

Comment une technique peut -elle " changer le monde " ?

[...] Qu' est -ce qui fait que certaines techniques soient considérées par tous comme des ouvreuses de chemin pour de nouvelles sociétés, voir même d' ères, et d'autres pas ?

(On ne parle pas, par exemple, de l'ère de la pénicilline, ...) Qu'est -ce qui fait alors que certaines techniques soient " révolutionnaires " et d'autres pas ?

En vertu de quoi une technique, et une technique seule, aurait -elle le pouvoir de révolutionner l' organisation sociale, alors que celle -ci est perçue par notre esprit comme étant le fruit de notre pensée rationnelle et de notre libre arbitre ?

Qu' est -ce qui fait qu' on attribue aujourd'hui aux NTIC le pouvoir d' accoucher d' une " nouvelle " société dite " de l' information " ? voire d' une nouvelle ère post Gütenberb ?

Je ferai le postulat que, pour qu' une technique soit " révolutionnaire ", elle doit d' une part remplacer avantageusement un usage antérieur, comme elle doit aussi, d' autre part, engendrer de nouvelles dynamiques socio-productives.

Ainsi, par exemple, si l' imprimerie est révolutionnaire parce qu' elle remplace avantageusement les moines copistes, elle crée aussi le roman littéraire qui était impensable sans elle.

[...] J' aimerais essayer de mettre en lumière ici ce qui est révolutionnaire dans les NTIC et répondre à la question : " qu' est -ce qui, dans ce dispositif technique, remplace avantageusement quoi ? " et " qu' est -ce que cela peut provoquer ? ".

Je pense que cette question est une question de base, à laquelle il faut donner une réponse si l' on veut pouvoir croire que les NTIC pourraient " changer le monde ".

Une révolution dans la raison graphique

[...] A partir du moment où l' électricité devient le support privilégié, tout ce qui est digitalisé sur un ordinateur connecté au réseau n' a plus besoin d' être imprimé sur du papier pour ensuite être transmis : c' est l' intéressé qui va le chercher ou le poser sur le réseau.

(Sauf que, dans cette ville virtuelle, on est affranchi de la distance, mais ceci au prix de l' abandon du corps. Le corps biologique ne rentre pas)

Comment le monde change

[...] Comment notre vie pourrait -elle changer à partir du moment où l' électricité détrône le papier, sans pour autant le remplacer ?

Posée ainsi, la question est vertigineuse.

[...] Donc, ce n' est pas parce que l' on ne vote pas sur un bout de papier, mais qu' on presse sur un bouton, que les votants seront forcément moins renseignés ou auront moins l' occasion de discuter du sujet.

[...] Si il en va ainsi, le remplacement du papier par le document numérisé dans nos bureaucraties locales nous laissent entrevoir la possibilité de leur disparition.

[...] Derrick de Kerkhove souligne ainsi que les " réseaux abolissent les repères traditionnels de l' identité, individuelle et collective ", et Olivier Abel, professeur de philosophie et d' éthique à la faculté de théologie protestante de Paris, écrit que " l' informatique augmente la complexité et, par -là même, étend l' espace de choix, donc de liberté.

Il faut en conséquence, penser à une citoyenneté complexe ".

[...] Ainsi les NTIC ne sont pas entièrement dépendantes de la logique et des besoins de la mondialisation du capital.

Conclusion

[...] Si le monde est d' accord aujourd'hui de juger Pinochet, c' est que les États n' ont plus beaucoup de pouvoir (est -il utile de rappeler que l'ensemble des États du monde ne contrôle que 4% du capital en circulation ?).

Si la logique de la mondialisation demande aux NTI de réduire de façon drastique l'entropie des administrations nationales, alors nos État " sont en passe de devenir des reliques du passé.

Il devient alors urgent d'encourager l'émergence de nouvelles solidarités, de nouveau systèmes de confiance et de reconnaissance pour imaginer les bases d' un nouveau contrat social.

Texte source

Pourquoi le monde va changer
avec les NTIC?

Communication au colloque :
La techno-science et ses impacts: Pour un nouveau contrat social ERIE - Enseignement et Recherches Interdisciplinaires en Éthique
Université de Lausanne
12-13 novembre 1998

Comment une technique peut-elle "changer le monde"?

Les sociologues ont beaucoup de peine à donner à la technique son juste rôle quant à sa capacité à transformer la société. Et ils ont raison, parce qu'il est difficile de se placer sur un axe allant de : "c'est le social qui crée la technique" à "c'est l'environnement technique qui crée la société". C'est beaucoup plus complexe. D'un côté en effet, une technique, aussi révolutionnaire soit elle, n'a d'existence historique que si elle est l'objet d'une appropriation sociale massive. Mais d'un autre côté on ne peut pas nier l'évidence de la réalité qui faisait dire à Jacques Neirynck au cours de sa leçon terminale à l'EPFL que la technique était une sorte de "rouleau compresseur que nous ne dirigions pas". C'est-à-dire que la technique aurait, en quelque sorte, une vie et un développement autonome.

Généralement les sociologues préfèrent opter pour une attitude plutôt proche de la première : "c'est quand même plutôt le social qui crée la technique". Et je suis d'accord avec eux. Mais on ne peut pas nier la réalité de l'outil et de ses limites, et nous sommes obligés de prendre en compte l'aspect accidentel et hasardeux des développements techniques. Parce que sans ces accidents, il n'y aurait pas de nouvelles technologies, de l'information ou autre. Sans moteur à explosion, pas d'avions et pas de voitures, et sans voiture pas de maison de week-end.

Je me poserai ici la question du lien de causalité entre le développement d'une technique et les changements sociaux et culturels qu'on leur impute. Qu'est-ce qui fait que certaines techniques soient considérées par tous comme des ouvreuses de chemin pour de nouvelles sociétés, voir même d'ères, et d'autres pas? (On ne parle pas, par exemple, de l'ère de la pénicilline, ...) Qu'est-ce qui fait alors que certaines techniques soient "révolutionnaires" et d'autres pas? En vertu de quoi une technique, et une technique seule, aurait-elle le pouvoir de révolutionner l'organisation sociale, alors que celle-ci est perçue par notre esprit comme étant le fruit de notre pensée rationnelle et de notre libre arbitre? Qu'est-ce qui fait qu'on attribue aujourd'hui aux NTIC le pouvoir d'accoucher d'une "nouvelle" société dite "de l'information"? voire d'une nouvelle ère post Gütenberb?

Je ferai le postulat que, pour qu'une technique soit "révolutionnaire", elle doit d'une part remplacer avantageusement un usage antérieur, comme elle doit aussi, d'autre part, engendrer de nouvelles dynamiques socio-productives. Ainsi, par exemple, si l'imprimerie est révolutionnaire parce qu'elle remplace avantageusement les moines copistes, elle crée aussi le roman littéraire qui était impensable sans elle. Si le moteur à explosion remplace avantageusement la traction animale, il développe aussi toute l'industrie du pétrole et de ses dérivés Ou encore: si la radio remplace efficacement le clocher du village, elle crée une ouverture d'esprit qui va par-delà la portée sonore d'une cloche.

J'aimerais essayer de mettre en lumière ici ce qui est révolutionnaire dans les NTIC et répondre à la question : "qu'est-ce qui, dans ce dispositif technique, remplace avantageusement quoi?" et "qu'est-ce que cela peut provoquer?". Je pense que cette question est une question de base, à laquelle il faut donner une réponse si l'on veut pouvoir croire que les NTIC pourraient "changer le monde". Avant de spéculer sur l'avenir de la société de l'information, il est utile de baser la réflexion sur la réalité de l'outil pour mieux en saisir les limites socio-techniques, et, à partir de là, tenter de comprendre les alternatives qui nous sont offertes en matière d'appropriation sociale des NTIC.

Une révolution dans la raison graphique

L'ordinateur ne constitue pas, en lui-même, l'outil accoucheur de la société de l'information. Après tout, tout ce que l'on fait avec un ordinateur, on pourrait le faire sans. Ada Lovelace disait déjà que la "machine universelle" de Charles Babbage ne pouvait rien faire d'autre que l'homme ne sache déjà faire.
On a commencé à parler des NTIC (avant on parlait simplement "d'informatique") lorsque s'est accomplie la fusion entre l'ordinateur et les réseaux de communication. Certes, cette fusion était déjà présente dans la première et la deuxième informatique (cf. les réseaux SAGE ou SABRE), mais il fallait la démocratisation de l'informatique, avec l'avènement du microprocesseur, et le développement du langage HTML pour que la 4ème informatique, celle qui unifie les PCs avec le réseau, puisse devenir un phénomène de masse. (il fallait d'autres "choses" aussi, mais j'écourte ce passage)
Et ce ne sont pas non plus le Web, ni le e-mail, ni l'Internet, ni les satellites qui constituent l'élément technique révolutionnaire qui nous intéresse ici. Ceux-là ne sont en effet que des applications et des développements d'un fait historique qui leur a préexisté.
L'idée que je cherche à amener ici, c'est que nous sommes en train d'assister à un changement du support matériel de l'information. La société de l'information se met en place à partir du moment où l'électricité remplace avantageusement le papier dans son rôle antique de support privilégié d'information. C'est cela qui est "révolutionnaire". C'est une révolution dans notre raison graphique.
Depuis Hollerith (vers 1870), l'électricité a remplacé le papier dans son rôle de support privilégié de l'écrit. Hollerith avait a vu en effet qu'il n'était pas possible - pour une question de temps - d'effectuer le recensement des États-Unis en gérant des papiers; alors il a eu l'idée d'utiliser l'électricité et les compteurs mécaniques de l'époque pour lire automatiquement des cartes perforées représentant chacune un des 70 millions d'habitant des États-Unis.
A partir de ce moment là, le papier a progressivement perdu son statut de support universellement privilégié de l'écrit, de l'inscription, et de la mémoire. C'est le support électrique ou électronique qui va peu à peu le remplacer. Celui-ci, au contraire du papier, circule à la vitesse de la lumière. Le support électrique a volé la vedette au papier, et c'est pour cela que les NTIC vont changer, certainement pas le monde entier, mais le notre en tout cas.
Tous les anciens supports de l'écrit étaient faits d'atomes. Depuis que nous avons inventé l'écriture, nous avons toujours stocké, enregistré, transmis, diffusé et traité nos inscriptions sur un support matériel, fait d'atomes et soumis aux lois de la gravitation terrestre (tablettes d'argile, cire, parchemins, stèles, papier, films, cassettes vidéo, disques, etc.).
Si le support de l'information est électrique il n'est alors plus soumis aux lois du monde des atomes. On peut alors, techniquement parlant, produire, stocker, traiter, émettre et recevoir de l'information de toutes sortes, sous forme orale, écrite et visuelle, sans limites de temps, de distance et de volume, et, en plus, à peu de frais. Et tout ceci à l'échelle de la planète, parce que les électrons circulent à la vitesse de la lumière.
L'essentiel de l'activité sociale humaine consiste à produire des biens et des connaissances, et à se les échanger parmi. L'histoire des technologies de l'information nous montre volontiers comment l'humanité s'est toujours souciée de réduire le temps de l'échange de ces informations/objets/valeurs. Lorsque l'information était inscrite sur un support matériel, ce dernier circulait dans l'espace newtonien que nous avons habité jusqu'à présent presque exclusivement (je ne veux pas nier la réalité du rêve, de l'imaginaire et des mystiques). Dans ce monde des atomes, la vitesse est soumise à la gravitation terrestre et à la pression de l'air, et elle dépasse difficilement Mach II ou Mach III. Dans le monde des bits, la vitesse de référence est idéalement de 299'792 km par seconde, ce qui, à l'échelle planétaire est quasi instantané. La rapport au temps et à l'espace dans le cyberespace n'est pas celui que nous avons communément connu, vécu et construit jusqu'à présent. Il est d'un autre ordre, à cause de sa vitesse, et ce n'est pas pour rien que Swatch lui a créé une nouvelle unité de mesure adaptée aux réalités du Cyberespace.
A partir du moment où l'électricité devient le support privilégié, tout ce qui est digitalisé sur un ordinateur connecté au réseau n'a plus besoin d'être imprimé sur du papier pour ensuite être transmis : c'est l'intéressé qui va le chercher ou le poser sur le réseau. Et cela change beaucoup dans notre économie. Par exemple, une seule " copie " d'un article peut desservir, virtuellement, toute la planète; et sa " valeur " ne se mesure plus à son nombre d'exemplaires imprimés, mais au nombre de connexions effectuées sur son emplacement dans le réseau. Il en va de même pour tous les produits susceptibles d'être digitalisé.
Par leur dispositif technique, les NTI fusionnent en un seul lieu toutes les bibliothèques et tous les médias du monde, créant ainsi une sorte de " cerveau planétaire ", une forme de mémoire collective interactive.
L'interactivité des NTIC fait qu'elles ne créent pas seulement une bibliothèque ou une librairie à l'échelle planétaire, mais elles créent aussi et surtout un espace public artificielle. Et au travers de cet espace public on échange des informations/valeurs, on fait des rencontres et on se fait connaître, comme dans les villes d'autrefois sur la Place du marché, sur l'Agora comme sur les cheminements hasardeux qui mènent de l'un à l'autre.
(Sauf que, dans cette ville virtuelle, on est affranchi de la distance, mais ceci au prix de l'abandon du corps. Le corps biologique ne rentre pas)

Et si notre monde doit changer avec les NTI, c'est notamment parce que, dans les villes que nous habitons, nombre de fonctions qui relèvent de l'échange d'informations/valeurs vont progressivement migrer dans cette nouvelle ville virtuelle.
Les NTIC ne "changent pas le monde", elles lui rajoutent quelque chose. Elles lui rajoutent le cyberespace, le monde des bits. Au moment où la recherche spatiale butte contre les limites du temps humain, elles lui offrent comme un sixième continent qui va contribuer, lui, à changer la vie sur les cinq premiers.

Comment le monde change

Comment notre vie pourrait-elle changer à partir du moment où l'électricité détrône le papier, sans pour autant le remplacer? Posée ainsi, la question est vertigineuse. Elle implique la quasi totalité de nos actions : notre façon d'apprendre (télé-enseignement, distance learning, ...), notre façon de travailler (le télétravail sous toutes ses formes), notre économie (l'exclusion des intermédiaires dans les chaînes de distribution), nos modes de transport, notre identification au territoire, notre façon de concevoir la vie et l'aménagement des villes, notre rapport à l'État comme à notre voisin de pallier...
Au niveau de la vie quotidienne, des objets familiers qui étaient matériels, parce qu'on avait besoin des atomes pour les constituer, vont disparaître ou se transformer, avec et dans les réseaux (des livres, les 25 volumes de l'encyclopédie, des cassettes, des films, des disques, les bottins de téléphones, les horaires des trains, l'argent de poche, les fax,...). D'autres sont déjà apparus, comme les bancomats et les cartes à puces, les adresse http dans la publicité, toutes les revues d'informatique, les magasins spécialisés, les e-mail dans les organisations politiques, ...
Mais il m'est impossible de lister ici tout ce qui serait susceptible de changer dans la vie quotidienne. D'abord parce que ce serait beaucoup trop long, et ensuite parce que cela serait nécessairement incomplet: nous ne pouvons pas imaginer aujourd'hui tous les développements accidentels des NTI qui sont à venir demain. N'oublions pas que le téléphone a été inventé pour retransmettre l'opéra, et que les développeurs de l'Internet, dans les années soixante-dix, n'avaient jamais imaginé l'ampleur qu'allait prendre leur dispositif un quart de siècle plus tard.
Je me contenterai ici, pour rester dans la problématique qui a été posée pour ces journées, d'avancer quelques questions relatives à l'impact des NTI sur la nature de l'état et des villes comme sur l'exercice de la démocratie. Le fait que les activités constitutives de la vie urbaine (Agora, place du marché et autres espaces publics) puissent être déplacé dans l'espace des réseaux va faire que beaucoup de choses vont pouvoir se transformer, disparaître ou émerger.
On se rend difficilement compte de l'importance du papier dans notre vie. On la commence pourtant avec un "acte de naissance", qui est une inscription officielle déposée sur un papier dans les archives de la commune; et on la termine avec un certificat de décès que les héritiers doivent obtenir pour toute une série de démarches administratives. Entre les deux, on oublie à quel point notre existence sociale est jalonnée par des papiers, certificats, déclarations officielles, signatures, diplômes, passeports, livrets de famille, No d'AVS, etc. Nous vivons dans une civilisation qui privilégie l'écrit contre la parole, parce que scripta manent. L'inscription fait foi contre la parole. Et c'est dans des textes que nous codifions nos lois, nos règles, nos constitutions, notre manière d'être ensemble. Et ce sont aussi des écritures qui gèrent les relation entre l'individu et l'état.
On voit alors comment les NTI peuvent engendrer un nouveau rapport entre l'individu et l'administration. A partir du moment où tous ces actes d'écriture sont numérisés l'individu n'a plus besoin de courir d'un guichet à l'autre pour déplacer des documents dispersés dans les différents offices de la commune, du canton et de la confédération.
Les NTI se développent de plus en plus dans les diverses administrations de l'État comme aide aux institutions démocratiques, que ce soit au niveau exécutif, législatif ou judiciaire. Il est évident qu'il vaut la peine, pour un député par exemple, de ne pas s'encombrer de dizaines de kilos de papiers de rapports et de règlements divers, de procès-verbaux de commissions et d'ordres du jour. Trop d'information tue le message. On peut supposer qu'un intranet d'État, puisse faciliter le travail des fonctionnaires en leur permettant d'une part un accès immédiat et documenté à l'ensemble des textes nécessaires à l'exercice de leur fonction, et, d'autre part, un mode d'échange d'information entre confrères qui soit rapide et affranchi des contraintes spatio-temporelles (messagerie électronique et forums). Dans ce sens, les NTI se présentent comme un outil assez idéal pour renforcer la communication entre les hommes politiques et, par là même, renforcer l'appareil d'État.
Dans le cas de figure où cet Intranet d'État viendrait à fonctionner, non pas tant techniquement mais surtout en termes d'usage, ce qui impliquerait qu'il soit effectivement le média de tous et non pas seulement des quelques fonctionnaires "branchés", on en arrive presque naturellement à se demander pourquoi on n'ouvrirait pas, dans une large mesure, ce "Cyberespace d'État" à tous les citoyens. A partir de là, on arrive assez naturellement au concept très flou de "démocratie électronique": on pourrait voter par e-mail. L'idée a été lancée par le PSS au grand dam de la classe politique traditionnelle. Mais si on peut voter par e-mail, alors on peut imaginer, comme Jacques Neirynck, une véritable démocratie directe, où l'on pourrait se passer finalement de la médiation du conseil communal, du Grand Conseil et du Parlement Fédéral (mais pas de l'exécutif), et restaurer ainsi une nouvelle forme de Landsgemeinde.

Nos voisins français, grands défenseurs de la démocratie représentative, sont viscéralement horrifiés par cette éventualité semblant sortir du plus mauvais roman de science fiction ("trop de démocratie directe tue la démocratie"). Pour les tenants de la démocratie représentative, il est nécessaire, pour assurer le bon fonctionnement des institutions démocratiques, de mettre la distance temporelle requise pour favoriser la concertation et la réflexion et éviter des débordements irréfléchis. "L'instantanéité politique" dit Philippe Breton "c'est une porte ouverte vers l'affectif, le passionnel et la démagogie". Paul Virilio ne croit pas non plus à la "démocratie presse-bouton": "Je ne crois absolument pas à ce que j'appelle la démocratie automatique. Je crois à la réflexion, pas au réflexe. Les technologies nouvelles sont des technologies de conditionnement (...) La prétendue démocratie électronique sera la fin de la démocratie participative." Pour Jacques Neirynck, ce genre de propos relève de la mauvaise foi: "Ce n'est pas une démocratie de réflexe. Cela ne veut pas dire qu'on ne fasse pas de campagnes avant de voter. Tout le monde pourra s'exprimer, tout le monde pourra se renseigner. Les informations que les partis politiques impriment à grands frais sur du papier et nous distribuent, on pourra les afficher sur l'écran. Donc, ce n'est pas parce que l'on ne vote pas sur un bout de papier, mais qu'on presse sur un bouton, que les votants seront forcément moins renseignés ou auront moins l'occasion de discuter du sujet. C'est de la mauvaise foi."

Pierre Lévy renforce également ce sentiment, sans se prononcer sur la forme politique, en constatant, comme bien d'autres, que le Cyberespace " est le lieu d'une démocratie d'initiative et d'expérimentation directe, utilisant de nouveaux instruments techniques et sociaux d'expression collective qui n'écrasent pas - et même favorisent - les singularités ". Philippe Rosé et Jean-Marc Lamère voient volontiers dans les NTI une excellente occasion de relancer le débat public "qui autrefois vivifiait la démocratie et qui a progressivement été étouffé par la société moderne".

Si il en va ainsi, le remplacement du papier par le document numérisé dans nos bureaucraties locales nous laissent entrevoir la possibilité de leur disparition. Et cette éventualité nous incite à nous poser la question de la survivance future du pouvoir étatique traditionnel et, par-là même, à redéfinir les bases de l'identité des acteurs tant publics qu'individuels.

Jusqu'à présent, il était admis plus ou moins implicitement, par les juristes et les politologues, que le pouvoir politique s'exerçait sur un territoire donné et clairement délimité. Certes, depuis la Société des Nations, une superstructure juridique s'est peu à peu développée pour régler, à travers le droit international, les divers problèmes nationaux engendrées par les effets de la mondialisation du capital. Mais ces dispositifs s'exerçaient sur le principe de la souveraineté nationale qui, en aucun cas, était menacée de remise en question. Or, les potentialités de l'Internet d'aujourd'hui et du futur sont justement celles de renforcer et faciliter le développement de rapports sociaux et économiques déliés de leurs contingences spatiales.

La globalisation des échanges et l'avènement du nouvel espace public des réseaux posent un problème politique sans précédent dans notre histoire. C'est celui de la souveraineté et du pouvoir d'action réel de l'État-Nation. L'État est construit sur un territoire, avec ses frontières, et l'Internet, par quasi définition, ne connaît pas de frontières territoriales. Pas plus que le capital. L'État perd une partie de son pouvoir et de sa substance parce qu'il perd sa capacité à contrôler les flux de valeurs et d'informations à l'entrée et à la sortie de son territoire : il est impossible d'ériger des postes de douane dans le Cyberespace.

La notion de "Globalisation" tient pour certains du "prêt-à-porter idéologique", et dissimule, plutôt qu'elle ne révèle la complexité de ce nouvel ordre mondial. Bien que l'image de "village global" aie été lancé à la fin des années 60, cette représentation ne s'est imposée que dans les années 80 avec la globalisation des marchés, des circuits de la finance, des entreprises, ainsi que de l'ensemble des échanges immatériels. Ce mouvement a été rendu possible par une vague de déréglementations et de privatisations qui a fait du marché le régulateur de la société. Ce qui s'est traduit par le recul des forces sociales, le déclin de l'État Providence et de la philosophie de service public, et, d'autre part, par la montée en puissance de l'entreprise, de ses valeurs et de l'intérêt privé. La mise sur pied de la LAMI par l'Organisation Mondiale du Commerce montre à quel point nous allons vers la décomposition de l'espace public traditionnel par la privatisation progressive du politique.

Le développement des NTIC annonce ainsi la fin du monde, unique et panoramique : sous les assauts de la diversité, notre réalité recule ou s'émiette au profit de l'irréductible pluralité des mondes. C'est l'une des principales préoccupations du grand patron de Microsoft qui écrit que "l'apparition de communications et de liaisons informatiques pratiquement gratuites modifiera les rapports entre les nations et les groupes socio-économiques", et ceci au point "qu'il se peut que certaines nations se sentent agressées si leurs peuples s'intéressent d'avantage aux cultures ou aux problèmes mondiaux qu'aux questions traditionnelles locales".

Alain Touraine, se référant à Braudel, nous rappelle à quel point la Nation joue un rôle de gardien du marché pour le capitalisme local, pour la petite bourgeoisie dépendante de l'argent que l'État a mis à disposition pour la reconstruction après la guerre. Si le capitalisme a toujours été international, mondial, depuis que les Portugais ont ramené l'or des Incas, la Nation est devenue trop petite pour le grand capital international. Pour ce dernier, le coût d'une centralisation excessive est devenu trop élevé. La mondialisation, qui est un fait structurel propre au capitalisme, fait que ce dernier ne peut pas se passer aujourd'hui des autoroutes de l'information.

L'impuissance croissante des gouvernements fait craindre aux intellectuels du Monde Diplomatique une "domination politique mondiale d'un nouveau type", et à Philippe Breton un "retour au féodalisme", caractérisé par une recrudescence des nationalismes locaux, avec des milices privées, l'arbitraire, etc. Paul Virilio voit de même cet avenir de tribus et de barbarie : "On dépasse l'État-Nation au profit d'ensembles plus restreints. Il y a une déconstruction de l'État National qui ne va pas dans le sens d'un dépassement de l'État-Nation, mais d'une régression aux tribus, aux groupes de pression qui ont précédé l'État National...".

Dans la foulée, nombre d'auteurs moins alarmistes s'inquiètent toutefois des conséquences des développements de la globalisation des marchés sur l'identité et la citoyenneté. Derrick de Kerkhove souligne ainsi que les "réseaux abolissent les repères traditionnels de l'identité, individuelle et collective", et Olivier Abel, professeur de philosophie et d'éthique à la faculté de théologie protestante de Paris, écrit que "l'informatique augmente la complexité et, par-là même, étend l'espace de choix, donc de liberté. Il faut en conséquence, penser à une citoyenneté complexe". Dans un monde "globalisé" les identifications deviennent de plus en plus multiples, et de moins en moins centrées sur un sentiment d'appartenance territoriale ou nationale. Pour Jacques Attali aussi, "les nations anciennes vont exploser", et il faut penser à mettre en place des "démocraties à N dimensions".

Si les pessimistes craignent un contrôle social accru, une déconstruction des réseaux traditionnels de la fonction politique et économique nationale, ainsi qu'un fort affaiblissement de la solidarité sociale telle qu'elle était conçue dans l'État-Providence, les plus optimistes, eux, se réjouissent de cet état des choses qui permet de "réduire l'entropie du système" et de se rapprocher d'un monde plus proche de celui qu'espéraient Erwin Schumacher dans Small is beautiful ou Denis de Rougemont avec son Europe des Régions.

La formation politique de nos citoyens de demain ne peut faire aujourd'hui abstraction de ces problèmes. Une attention toute particulière devrait être portée non seulement sur les différents rôles que jouent ce nouvel espace public des réseaux dans le processus de mondialisation et de déliquescence des États, mais aussi, et peut-être surtout, sur ses implications au niveau des réseaux sociaux locaux.

En effet, l'étude des usages sociaux de l'Internet démontre le contraire de la thèse pessimiste de Paul Virilio qui voudrait que " sur l'Internet on aime plus son lointain que son prochain ". En effet, dans la pratique, les Netizens communiquent jusqu'à 10 fois plus avec leur "prochain" qu'avec leur lointain. L'usage social qui est fait des technologies de communication correspond rarement à l'idée que s'en faisaient leur créateurs. Ainsi les NTIC ne sont pas entièrement dépendantes de la logique et des besoins de la mondialisation du capital. Les usages sociaux qui en seront fait pourraient, paradoxalement, se retourner contre la logique de mondialisation qui a prévalu à leur développement : le dispositif technique conçu par les ingénieurs pour créer le village global, semble être appropriée par les usagers pour renforcer les dynamiques "locales" tout autant, si ce n'est plus, que les dynamique "globales". Ceci au détriment des dynamiques nationales.

Conclusion

Si le monde est d'accord aujourd'hui de juger Pinochet, c'est que les États n'ont plus beaucoup de pouvoir (est-il utile de rappeler que l'ensemble des États du monde ne contrôle que 4% du capital en circulation?).

Si la logique de la mondialisation demande aux NTI de réduire de façon drastique l'entropie des administrations nationales, alors nos État" sont en passe de devenir des reliques du passé. Il devient alors urgent d'encourager l'émergence de nouvelles solidarités, de nouveau systèmes de confiance et de reconnaissance pour imaginer les bases d'un nouveau contrat social. Ne serait ce que pour des raisons fiscales...

Si l'État veut survivre, il lui faut un nouveau Montesquieu pour savoir si, à terme, il est préférable de vivre dans une anarchie de mafias économiques, ou bien au sein d'un état mondial tentaculaire, ou bien dans une fédération de "tribus locales" plus ou moins grandes et plus ou moins barbares, ou bien encore dans quelque chose d'autre qui reste à inventer.