Résumé automatique par filtrage sémantique d'informations dans des textes

Jean-Luc Minel, Jean-Pierre Desclés, Emmanuel Cartier, Gustavo Crispino, Slim Ben Hazez, Agata Jackiewicz

Résumé automatique par filtrage sémantique d'informations dans des textes

Jean-Luc Minel* — Jean-Pierre Desclés* — Emmanuel Cartier*
Gustavo Crispino — Slim Ben Hazez* — Agata Jackiewicz***

** Équipe Lalic du CAMS*
CNRS/ Université Paris-Sorbonne, 75 006 Paris
Jean-Luc.Minel@paris4.sorbonne.fr

*** Université de la République*
J. Herrera y Reissig, 565
11300 Montevideo – Uruguay
crispino@fing.edu.uy

RÉSUMÉ : *L'originalité de notre projet est de se donner les moyens d'accéder au contenu sémantique des textes, pour mieux les cibler et en extraire certaines séquences particulièrement pertinentes. A cet effet, nous nous proposons d'exploiter un savoir de nature purement linguistique, et plus précisément sémantique, en nous appuyant sur la technique d'exploration contextuelle. Le modèle conceptuel et le langage de description des connaissances linguistiques de la plate-forme FilText sont présentés, ainsi que son implémentation logicielle ContextO. Un exemple d'utilisation de ContextO, la production de résumé automatique, est détaillé.*

ABSTRACT : *Our project aims to provide means to identify semantics in texts in order to extract relevant sequences. We present the contextual exploration method which exploits this kind of linguistic knowledge. The conceptual model and the descriptive language used in FilText are presented as well as the workstation ContextO. As an example of the use of ContextO, automatic summarization by semantic labeling is detailed.*

MOTS-CLÉS : *Ingénierie linguistique, étiquetage sémantique, résumé et filtrage automatique, exploration contextuelle, connaissances causales, énoncés structurants.*

KEY WORDS : *Computational linguistics, semantic labeling, automatic summarization, contextual exploration method, causal knowledge, structuring phrases.*

1. Introduction

Les grandes entreprises, les grandes administrations (Ministères, services publics...), les laboratoires et services de développement sont confrontés à un défi : gérer la masse des documents textuels saisis sur des supports électroniques. Comment les classer ? Comment les stocker pour y retrouver rapidement les informations qu'ils contiennent ? Comment diffuser ces informations à ceux qui sauront les utiliser ? Comment filtrer une information pertinente parmi toutes les informations contenues dans les documents stockés ? En effet, ce qui est jugé pertinent pour l'un ne l'est pas nécessairement pour un autre. Les critères traditionnels, plus ou moins efficaces pour les textes saisis et manipulés par les supports imprimés, comme l'emplacement physique du document dans les archives, la mémoire du documentaliste..., ne sont pas applicables aux documents électroniques. D'autres critères, plus rigoureux, doivent alors être trouvés pour s'adapter aux possibilités du traitement informatique. 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. En effet, elles sont souvent trop bruyantes : trop d'informations non pertinentes.

La notion de résumé automatique devient un des grands thèmes du Traitement Automatique des Langues. Plutôt que de diffuser les documents entiers, n'est-il pas préférable de diffuser seulement les résumés qui contiendraient les informations vraiment pertinentes ? En effet, il est plus facile de lire quelques lignes ou quelques pages susceptibles d'apporter l'information cherchée que de lire des centaines de pages pour s'apercevoir qu'aucune information nouvelle ne s'y trouve. 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. Mais pourquoi ne pas se contenter d'un résumé rédigé par un résumeur professionnel ? D'abord parce que tous les textes ne sont pas accompagnés d'un résumé, notamment les textes qui circulent sur le réseau Internet, et surtout parce que le coût de production d'un résumé par un professionnel est très élevé. A titre d'exemple, pour un texte source d'une dizaine de pages, un résumeur professionnel, lorsqu'il est spécialiste du domaine, produit un résumé en une dizaine de minutes mais il lui faut presque une heure lorsque le domaine traité ne relève pas de sa compétence. Enfin, la fiabilité de ce type de résumé est très controversée. Ainsi une expérience [RAT 61] montre des résultats surprenants ; 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 %. Et 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 %.

Les techniques statistiques fondées sur des fréquences d'occurrences et de cooccurrences [SAL 83] ont été largement utilisées, avec des résultats mitigés, du point de vue des utilisateurs. Il apparaît de plus en plus qu'il faut arriver à maîtriser

la gestion sémantique des informations : tous les mots n'ont pas la même pertinence informationnelle, une expression peu utilisée dans un document (un hapax par exemple) peut être un indicateur pertinent pour la veille technologique et les mesures d'innovation. C'est pourquoi, depuis quelques années, un certain nombre de modèles exploitent des connaissances ou des ressources linguistiques. Ainsi, les modèles de [MII 94, MAR 97] s'appuient sur la « Rhetorical Structure Theory » [MAN 88] et sur l'analyse des connecteurs pour construire des arbres rhétoriques qui hiérarchisent l'importance des parties textuelles. De leur côté, Paice [PAI 81], Lehmam [LEH 95] et Teufel [TEU 97] repèrent des fragments textuels sur la base de scores calculés pour chaque phrase, en fonction de termes préétablis. Quant à Berri [BER 96a], il cherche à attribuer des étiquettes sémantiques à certaines phrases, afin de les sélectionner ou non dans un résumé. Masson [MAS 98] reconnaît partiellement des structures thématiques dans un texte et Ellouze [ELL 98] exploite différents types d'objets textuels pour produire des schémas de résumé.

Les évaluations réalisées sur certains systèmes [MIN 97, JIN 98] ainsi que les travaux menés en collaboration avec les résumeurs professionnels [END 95] ou en comparaison avec les résumés produits par ces professionnels [SAG 98], ont néanmoins montré la difficulté à réaliser des résumés standard, c'est-à-dire construits sans tenir compte des besoins des utilisateurs. En effet, il n'existe pas de critères précis [SPA 93] pour déterminer ce que serait un bon résumé. 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. Par ailleurs, l'activité résumante des humains [FAY 89] a été fort peu étudiée par la psychologie. Les résumés seront 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.

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.

Cette expérience du résumé automatique acquise par la réalisation de la maquette SERAPHIN, et les résultats obtenus, évalués sur le terrain par deux protocoles [MIN 97], nous ont amenés à élargir le champ de nos recherches en visant non plus de

simples résumeurs automatiques non ciblés, mais des *systèmes automatiques de filtrage d'informations extraites de textes*. D'une part, ces systèmes sont fondés sur l'utilisation de critères sémantiques, donnant ainsi à un utilisateur la possibilité de définir un *profil de filtrage* en fonction de son objectif. D'autre part, ils font essentiellement appel à des ressources linguistiques générales, c'est-à-dire indépendantes des domaines particuliers (chimie, ingénierie, pharmacie, médecine...).

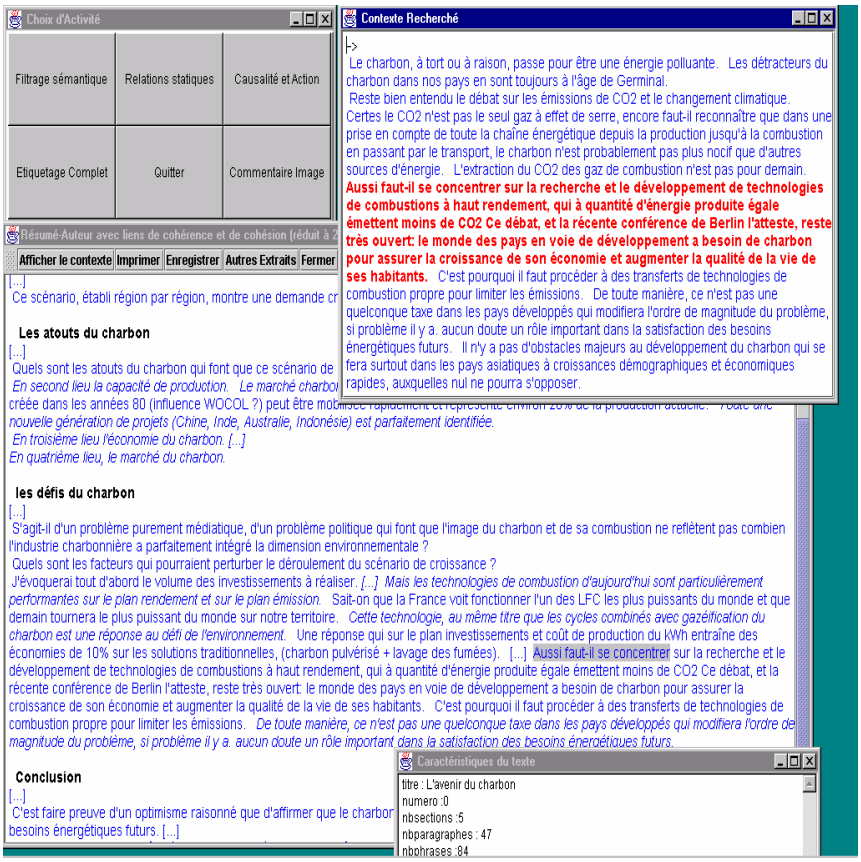

Figure 1. *Exemple de résumé et de navigation entre un extrait et le texte source*

La figure 1 illustre ce qu'un tel système apporte à un utilisateur en présentant un extrait produit par la plate-forme logicielle ContextO (cf. §3.4). En haut à gauche de l'écran, une interface présente sous la forme de boutons les différentes choix offerts

à l'utilisateur. Dans l'exemple, celui-ci a déclenché la tâche *Filtrage sémantique (*bouton en haut à gauche*)*, avec un profil du type *Résumé-Auteur*, qui n'apparaît pas sur l'exemple. Trois fenêtres présentent les résultats de ce filtrage. La fenêtre principale, au centre sur la figure, affiche le résumé produit ; les titres sont systématiquement placés dans le résumé, ils apparaissent en gras ; les phrases sélectionnées grâce à l'étiquetage sémantique sont affichées en police normale, alors que celles qui ont été placées dans le résumé pour en améliorer la cohérence et la cohésion sont en italique. C'est le cas de la phrase « *En troisième lieu l'économie du charbon* » qui a été reconnue comme appartenant au segment textuel qui débute par la phrase « *En premier lieu..* » ; ce segment est lui même lié à la phrase interrogative « *Quels sont les atouts du charbon… »* étiquetée comme une *Annonce Thématique.* L'utilisateur peut également afficher le contexte d'une phrase présente dans l'extrait en sélectionnant (avec la souris) une partie de la phrase - c'est le cas de la phrase qui commence par « *Aussi faut-il se concentrer..* » - qui apparaît surlignée dans la fenêtre principale. Ce contexte, ici le paragraphe du texte source contenant la phrase, est visualisé dans la fenêtre placée en haut à droite. Enfin, la troisième fenêtre qui apparaît en bas à droite affiche certaines caractéristiques du texte traité comme le nombre de phrases, le nombre d'indicateurs identifiés, etc.

Nous allons présenter la méthodologie utilisée pour organiser les ressources linguistiques utilisées et nécessaires à la production de ce type d'extraits, ainsi que la plate-forme qui accueille et exploite ces ressources.

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. Les ressources linguistiques sont constituées de marqueurs discursifs explicites (morphèmes, mots, expressions et locutions...) caractéristiques d'une intention pragmatique de l'auteur du texte, intention que le système doit être capable d'identifier et d'interpréter sémantiquement en fonction du contexte, comme doit le faire un lecteur peu averti du domaine traité.

D'une part, nous cherchons à exploiter directement l'organisation textuelle des propos de l'auteur. Le jugement d'importance est fondé essentiellement sur ce que l'auteur a lui-même explicitement mis en valeur dans son texte. Plusieurs segments textuels tels que *: voici ce qui doit être noté…, j'insiste sur…, mon hypothèse principale est…, nous allons traiter dans cet article…,* sont autant d'indicateurs pertinents employés par l'auteur pour orienter l'attention du lecteur vers certains segments textuels.

D'autre part, nous nous intéressons aux manifestations textuelles de certaines relations organisatrices de connaissances (relations définitoires, causales,

spatiales…). 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. Ce savoir ne se réduit pas à une nomenclature (objets, propriétés, événements, etc.). Il est notamment structuré par un certain nombre de relations entre concepts (cf. § 4.1).

La méthode d'exploration contextuelle est issue d'une réflexion initiale sur le traitement informatique des valeurs aspecto-temporelles dans les langues avec une première réalisation informatique SECAT [DES 91] 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 indications présentes dans le contexte, pour un calcul des valeurs sémantiques relevant de différentes tâches. 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 97b] qui sont déclenchés, dans un premier temps, 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. Comme exemple de champ grammatical, donnons l'identification des valeurs sémantiques de morphèmes grammaticaux comme ceux de l'aspect en 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...

L'identification d'un marqueur (grammatical ou discursif) n'est cependant pas suffisante pour déterminer complètement la valeur sémantique du 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 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. Ceux-ci permettront, soit de lever l'indétermination sémantique attachée *a priori* au marqueur analysé, par conséquent une étiquette sémantique pourra être attribuée à un segment linguistique (syntagme, phrase, paragraphe selon les cas), soit d'invalider les hypothèses sémantiques qui pouvaient être envisagées à propos du marqueur analysé dans son contexte. 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.

Un exemple clair d'indétermination sémantique liée à un marqueur grammatical, même s'il ne relève pas directement de la problématique du résumé automatique, est le morphème de l'imparfait *-ait*, facilement identifiable. 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, on peut 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*, *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 référentielle. On remarquera, à propos de cet exemple, que d'un côté les marqueurs et indices complémentaires sont des indices généraux, totalement indépendants des ontologies et que, d'un autre côté, le traitement informatique ne nécessite pas au préalable des analyses et des représentations très complexes.

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

- d'un ensemble d'*indicateurs linguistiques* - c'est-à-dire de marqueurs linguistiques 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.

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. Il faut souligner que les règles d'exploration contextuelle sont des heuristiques 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 (cf. § 3.1.2 et § 3.2.2).

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. Il va de soi qu'une connaissance du domaine externe [PAZ 97] ne peut

qu'améliorer le traitement et la compréhension du texte. La prise en compte du contexte linguistique peut ainsi être complétée par une prise en compte du contexte externe (connaissances des domaines, stratégies de communication entre les interlocuteurs...). Cette technique d'analyse ne fait pas appel à des analyses syntaxiques préalables approfondies, une catégorisation morpho-syntaxique des unités linguistiques étant en général suffisante.

Les avantages de la technique d'exploration contextuelle sont :

- l'*indépendance* entre les connaissances linguistiques nécessaires au système et les connaissances accumulées sur un domaine particulier ; de ce fait, les Systèmes d'Exploration Contextuelle sont généraux et adaptables à de nombreux domaines d'application ;

- la *compatibilité* de l'exploration contextuelle avec une utilisation de connaissances contextuelles externes (ontologies, connaissances des stratégies argumentatives...) ;

- l'*orientation* vers des problèmes précis : les indicateurs linguistiques sont jugés pertinents en fonction du filtrage voulu ;

- l'*extensibilité incrémentale*, obtenue en complétant les listes déjà établies (recherche d'indices plus fins) et en affinant les règles d'exploration ; un Système d'Exploration Contextuelle est donc plus ou moins performant selon la richesse des indices pris en compte et la finesse de l'exploration ; il peut s'approcher par approximation d'une solution idéale, l'approximation devenant meilleure lorsque le système devient plus riche.

3. Plate-forme FilText

Le concept de plate-forme d'ingénierie linguistique dédiée au traitement textuel conçue comme « une boîte à outils » n'est pas nouveau ; par exemple [HER 96, MEU 98] proposent de définir un modèle conceptuel puis de développer des tâches spécialisés 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]. La plate-forme FilText reprend en partie ce paradigme, en s'appuyant sur les techniques d'exploration contextuelle [DES 91, 97a, 97b] présentées précédemment. Elle présente l'avantage d'une part, de rendre le travail linguistique relativement indépendant de toute implémentation informatique et d'autre part, d'articuler effectivement dans une même architecture logicielle analyse linguistique et traitement informatique. Cette méthode n'est donc pas limitée à des traitements spécifiques mais offre un cadre de travail réaliste comme le souligne M. Charolles :

« La délimitation des univers d'énonciation mettant en jeu (…) une pluralité d'indicateurs, je plaiderai en faveur d'une approche du type " exploration contextuelle "… cette approche offre aux linguistes non pas un modèle général d'interprétation des discours – objectif à mon avis impossible à atteindre et qu'elle ne prétend du reste pas atteindre – mais un cadre théorique réaliste aussi bien pour le développement d'applications pratiques comme le filtrage ou le résumé automatique de textes que pour celui d'hypothèses psycholinguistiques locales, hypothèses dont certaines ont été encore peu explorées. » [CHA 98]

Nous avons élaboré un modèle conceptuel et un langage de description de ces données linguistiques. Celles-ci sont gérées par un système, indépendamment des applications qui les utilisent, cela en vue d'assurer une pérennité et une capitalisation de ces connaissances. Les règles d'exploration contextuelle sont écrites dans un langage formel indépendant de tout langage de programmation.

3.1. ***Organisation et gestion des données linguistiques***

Le travail préalable du linguiste consiste à étudier systématiquement un corpus de textes pour y rechercher des régularités discursives dont l'emploi est représentatif de la catégorie sémantique considérée. D'après l'hypothèse de travail qui jusqu'ici s'est révélée féconde, les modes d'expression associés à ces catégories discursives dans les corpus sont en nombre fini. Par conséquent, cela n'exige ni le repérage de structures syntaxiques spécifiques, comme peut le faire [REB 98] qui s'appuie sur les travaux de Z. Harris, ni la construction d'ontologies du domaine en vue d'en énumérer les concepts ou plus modestement les éléments thématiques.

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. Nous allons illustrer son utilisation par quelques exemples.

Dans la table 1, le linguiste déclare soit des formes lexicales significatives (des indices ou des indicateurs) qu'il organise en classes[1] non nécessairement disjointes, soit 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.* Bien que des outils d'aide à l'acquisition permettent de produire automatiquement toutes les formes fléchies ou dérivées, le linguiste doit ne retenir que certaines formes fléchies, car pour une tâche donnée, seules certaines flexions d'un verbe sont

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

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.

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

Table 1. *Déclaration des formes*

Dans la table 2, ces formes sont déclarées comme des indices ou des indicateurs, qui peuvent être associés à une ou plusieurs tâches. Rappelons qu'un indicateur linguistique est un marqueur linguistique d'une valeur sémantique jugé pertinente pour la tâche à résoudre et qu'un indice permet de résoudre, en contexte, l'éventuelle polysémie de l'indicateur (cf. §2).

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

Table 2. *Déclaration des indices et indicateurs*

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. La première ligne de la table 3 déclare une 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.

D'une manière plus générale, une règle peut déclencher différents types de décision (cf. § 3.1.2). Des outils d'aide à la gestion de la cohérence et à l'intégration des connaissances issues d'autres travaux linguistiques permettent de répondre à l'objectif d'une acquisition incrémentale et capitalisable des connaissances. Le linguiste peut ainsi voir 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

Table 3. *Déclaration des règles*

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. L'intérêt de cette notion 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 est important de pouvoir exprimer des contraintes qui prennent en compte la dimension textuelle. Chaque règle comprend 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.

Nom de la règle : Rthematique **;**
Tâche déclenchante : Thématique ;
Commentaire : capte un schéma du type : *Dans les lignes qui suivent … nous présentons …*
Classe de l'Indicateur : &verbe_presentatif ;
E1 := Créer_espace(voisinage Indicateur)
L1:= &partie_document3
L2:= &partie_document1
Condition : Il_existe_un_indice x appartenant_à E1 tel_que classe_de x appartient_a (L1) ;
Condition : Il_existe_un_indice y appartenant_à E1 tel_que classe_de y appartient_a (L2) ;
Precede (x,y, contrainte) ;
Actions :
1 : Attribuer(PhraseParent « Annonce_Thematique »)

Figure 2. *Un exemple de règle écrite dans le langage formel*

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. La partie *Condition* explicite les conditions que doivent vérifier les indicateurs et les indices complémentaires. Le langage permet d'exprimer différentes conditions, comme l'existence, la position, l'agencement des indices et d'exprimer des contraintes sur les attributs des unités lexicales ou sur les morphèmes qui les composent. La partie *Action* indique le type d'actions réalisées par la règle. Actuellement, deux actions possibles sont : attribuer une étiquette à un segment textuel ou déclencher une autre tâche.

La figure 2 présente un exemple de règle écrite dans ce langage. La règle de nom *Rthematique*, attribuée à la tâche *Thématique*, est déclenchée si un indicateur appartenant à la classe *&verbe_presentatif*; qui regroupe des formes lexicales des verbes présentatifs, est présent dans l'espace de recherche *E1*. Celui-ci est construit avec les formes lexicales de la phrase dans lequel apparaît l'indicateur. Une première *condition* exprime qu'une occurrence d'un indice de la classe *&partie_document3*, qui regroupe certains substantifs comme *article, lignes*, *rapport*, etc., doit être présente dans la phrase et une seconde *condition* précise une contrainte sur la présence, dans le même espace de recherche, d'un indice de la classe *&partie_document3*. Enfin, le prédicat *Precede* impose un ordre de précèdence entre les deux indices. La partie *Action* indique que l'étiquette *Soulignement_Auteur* est attribuée à la phrase en question.

Le système de gestion des données linguistiques est totalement spécifié dans un modèle objet et implémenté dans un système de gestion de base de données relationnelles. Les règles écrites en langage formel sont actuellement traduites en JAVA mais un compilateur de règles est en cours de spécification.

3.2. *Moteur d'exploration contextuelle*

Le moteur d'exploration contextuelle exploite 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'exécuteur.

3.2.1. *Analyseur de texte*

L'analyseur de textes construit une première représentation qui reflète l'organisation structurelle du texte. Le traitement textuel nécessite 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 peut s'appuyer sur des balises d'un langage du type SGML, HTML ou XML, lorsqu'elles existent. En l'absence d'un tel balisage, il est fait appel à une tâche spécialisée, un segmenteur développé par Mourad [MOU 99] au sein de notre équipe. Ce segmenteur applique des règles heuristiques pour reconnaître les sections avec leurs titres, les paragraphes, les phrases et les citations. L'analyseur construit ainsi une structure hiérarchique, totalement spécifiée par une grammaire illustrée par

la figure 4 en utilisant la notation UML[2], qui est utilisée par les opérations spécifiées dans les règles d'exploration contextuelle. Cette structure est ensuite enrichie en vue de modéliser partiellement les chaînes de liage, les segments textuels, les cadres de discours [CHA 88, ADA 90].

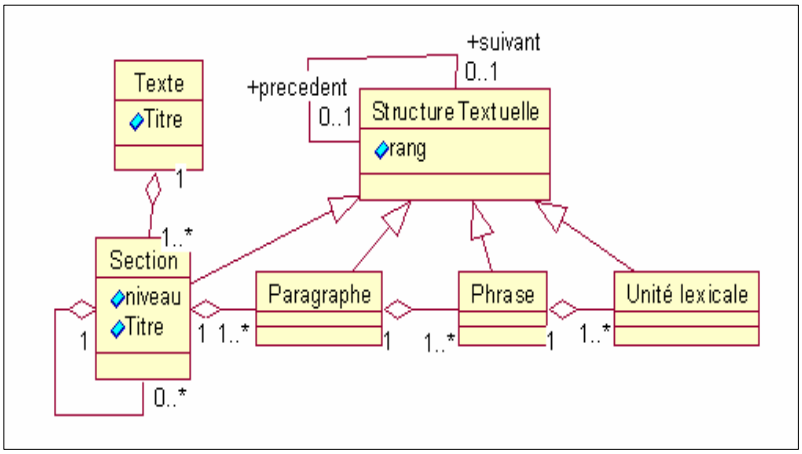

Figure 3. *La structure hiérarchique (simplifiée) d'un texte*

3.2.2. *Exécuteur*

L'exécuteur déclenche, pour toutes les tâches choisies par l'utilisateur, toutes les règles associées à celles-ci. Les règles sont 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 ». D'autre part, 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. Enfin, il convient de noter que le langage formel de déclaration des règles ne présume en rien des outils informatiques (automates, moteur spécialisé, code dédié, etc.) utilisés pour les implémenter.

[2] UML est la norme de modélisation objet adoptée par *l'Object Management Group*.

3.3. *Agents spécialisés*

Les agents spécialisés ont pour tâche d'exploiter les « décorations sémantiques» du texte en fonction des objectifs définis par l'utilisateur. Il existe ainsi un agent résumeur qui construit un résumé composé de phrases qui correspondent à un profil type. Il existe également un agent filtreur qui construit différents extraits de textes en fonction de profils choisis par l'utilisateur. Il existe enfin un agent chargé de l'extraction et de la capitalisation des connaissances à partir de documents textuels.

Ce dernier système [LEP 99] permet d'extraire des relations sémantiques entre concepts. Les relations identifiées (hiérarchies classe / sous-classe, attribut / valeur, instances de classe, 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. Par exemple, à partir de l'énoncé « *Le 3 TC est un type de médicament anti-VIH appelé " inhibiteur nucléosidique " de la transcriptase inverse* », extrait d'un corpus sur le VIH et de la liste des termes de référence associée, le système propose deux relations sémantiques :

- une relation d'inclusion entre « 3 TC » et « médicament anti-VIH » ;
- une relation d'identification entre « inhibiteur nucléosidique de la transcriptase » et « médicament anti-VIH ».

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].

3.4. *Plate-forme logicielle ContextO*

La plate-forme FilText vise à accueillir des connaissances linguistiques et à les exploiter relativement indépendamment du matériel. Cet impératif nous a conduits à utiliser le langage JAVA pour implémenter FilText dans la plate-forme logicielle ContextO (figure 4) que nous avons développée. En réponse à des appels d'agents spécialisés, le moteur d'exploration contextuelle déclenche, pour une ou plusieurs tâches spécialisées (repérage d'annonces thématiques, de relations statiques, etc.), le processus de reconnaissance des indicateurs et des indices présents dans un segment textuel. Ce processus est réalisé par le système de gestion des données linguistiques qui, en retour, fournit au moteur d'exploration contextuelle les règles potentiellement déclenchables. Les agents spécialisés ont pour tâche d'exploiter les résultats du traitement en fonction des objectifs qu'ils souhaitent atteindre.

La plate-forme logicielle ContextO est actuellement opérationnelle, et les ressources linguistiques issues de systèmes antérieurs [CAR 97, GAR 98, JAC 98, JOU 93], soit environ 11 500 marqueurs et 250 règles d'exploration contextuelle, sont intégrées progressivement. Les performances de traitement sont de l'ordre d'une page à la seconde avec une limite supérieure d'environ 250 pages, sur une plate-forme matérielle de type micro-ordinateur.

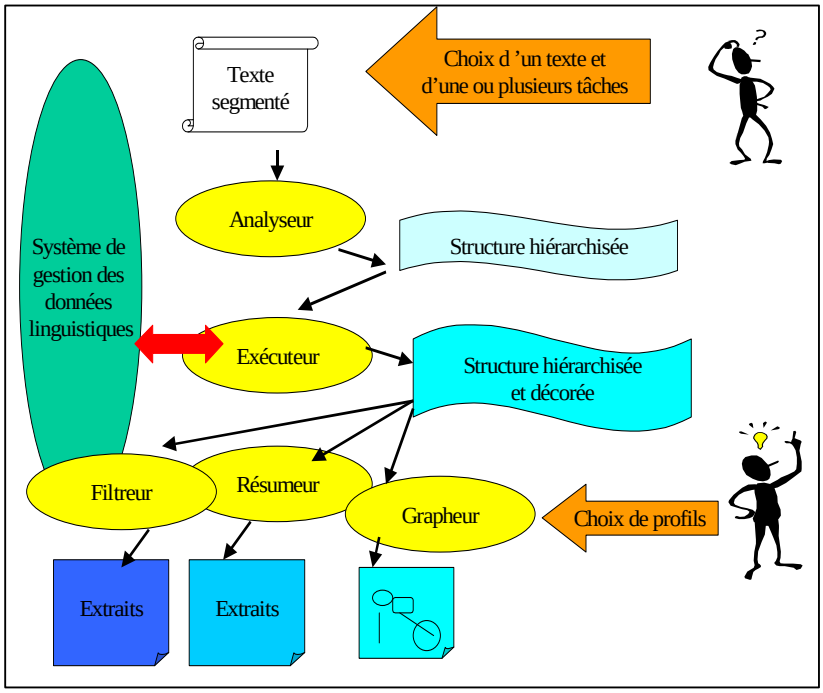

Figure 4. *Les différents composants de ContextO*

4. Un exemple d'agent spécialisé : le résumeur-filtreur

Cet agent exploite les connaissances des systèmes SERAPHIN [LER 94, BER 96a], fruit d'une collaboration avec la Direction des Études et des Recherches (DER) de la société EDF et SAFIR [BER 96b]. Pour une entreprise comme EDF, les besoins en matière « d' activité résumante » sont nombreux et portent sur des domaines variés, stratégiques pour l'entreprise comme le véhicule électrique, le traitement des déchets, l'effet de serre, l'environnement, la domotique, l'opinion publique et le nucléaire, etc. Les applications envisagées s'inscrivent dans un contexte de développement où différents outils de traitement textuel doivent coopérer.

Pour répondre à ces besoins, un cahier des charges fixait plus précisément les contraintes auxquelles devait répondre le système SERAPHIN :

- les textes à résumer traitent de domaines très divers mais relèvent tous d'un genre que l'on peut qualifier d'informatif. Avec cette contrainte de départ, il était hors de question que le système puisse dépendre d'une description d'un domaine quelconque, d'un dictionnaire du domaine, où d'une ontologie.

- ces textes présentent une structuration très variable ; certains n'en possèdent aucune (style journalistique), d'autres à l'inverse sont fortement structurés (rapports techniques) ;

- 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 ;

- les résultats attendus doivent être produits en quelques minutes ;

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*

Les principales étiquettes sémantiques relèvent de deux types d'identifications : la reconnaissance des énoncés structurants, d'une part, et les connaissances causales et les définitions d'autre part.

4.1.1. *Enoncés structurants*

L'identification des énoncés structurants regroupent plusieurs types d'informations.

L'étiquette « annonce thématique » [CAR 97] 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[3]. L'annonce thématique par excellence est le plan du document, exprimé en tête de texte et/ou, plus rarement, en conclusion du texte. 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...*),

[3] Voir [Fay 89] pour une justification de l'importance de ce type d'information, notamment dans l'objectif d'un résumé. Nous faisons aussi reposer notre analyse sur l'existence d'une norme textuelle propre aux textes à visée informative et argumentative qui recommande de disposer des repères permettant au lecteur d'identifier sa structure thématique et les enseignements généraux, les points importants du propos.

d'un présentatif (*commencer l'étude, 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 comprend pas de plan de document explicité en tête ou repris en fin de document, ni de titres et de sous-titres, le système doit repérer des annonces thématiques au fil du texte. Nous avons ainsi identifié trois formes d'annonce locale : 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.

L'étiquette « récapitulation/conclusion thématique» est attribuée aux phrases explicitant les conclusions et enseignements généraux du texte ; il s'agit là encore d'une information capitale, puisqu'elle correspond à ce qu'il faudra retenir de la démarche textuelle. Cette classe d'informations comporte deux sous-classes : les récapitulations et les conclusions. Les énoncés récapitulatifs sont aisément identifiables au moyen de locutions comme *: 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. En voici quelques exemples[4] :

(3) ***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.(...)*

(4) *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. (...)*

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

(6) ***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 (*donc, il faudrait donc*) ou reformulatifs (*de toute façon*), nous devons vérifier la co-présence d'un marqueur de soulignement (*est devenu évident*) ou d'une modalité (*il faudra, il sera … nécessaire*) et la position du connecteur dans la phrase. De plus, pour n'extraire que les conclusions globales, nous nous limitons à certaines positions de la phrase dans le texte.

---

4 Tous ces exemples proviennent du corpus de textes utilisé dans le projet SERAPHIN, identifiés par le nom de l'auteur.

4.1.2. *Définitions*

Les énoncés définitoires constituent un autre type d'information recherchée car c'est un moyen de capter le ou les thèmes d'un article (exemples 7 et 8). Nous avons également élaboré un ensemble de règles qui, sur la base d'un groupe nominal du titre, extraient certaines propositions qui concernent ce syntagme nominal ; ce type d'énoncés présente l'intérêt de pallier l'absence des annonces thématiques classiques.

(7) *(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 à 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)*

(8) ***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)*

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... Il se trouve que la notion même de causalité est très présente dans les textes mais avec des acceptions diverses et d'importantes nuances de signification. De plus, elle est impliquée dans une très grande diversité d'actions humaines, scientifiques et techniques. Aussi, dans la perspective du filtrage et de l'extraction d'informations pertinentes, la causalité ne peut-elle être abordée d'une manière unique et pour un seul type de besoin ou encore dans un seul mode d'utilisation. Dans la plate-forme FilText, nous avons choisi d'offrir plusieurs possibilités de manipulation de la causalité à partir de diverses connaissances qui la caractérisent.

Partant d'une vaste étude linguistique de l'expression des rapports causaux dans des textes, nous disposons maintenant [JAC 98] d'un ensemble de connaissances relatives : (i) à une certaine organisation sémantique du lexique verbal de la causalité (avec des verbes comme *faciliter, aider, gêner, augmenter, diminuer, freiner, limiter, stopper, empêcher* ...) selon quatre approches des rapports causaux : qualitative, fonctionnelle, analytique et synthétique ; (ii) aux relations étroites (sources nombreuses d'ambiguïtés) que la causalité entretient, de par son expression linguistique, avec les notions d'action agentive ("un agent fait et contrôle que ...") et d'argumentation (explication, justification ...) ; (iii) à la trace discursive de celui qui

prend en charge la relation causale énoncée, ce dernier pouvant prendre une certaine distance par rapport à la relation causale, cette prise de distance étant manifestée par des jeux de modalité et par des modalités d'action. Chacune de ces trois facettes de la causalité se matérialise par un vaste réseau hiérarchique de valeurs sémantiques de plus en plus spécifiques auxquelles sont associés des indices linguistiques identifiables dans un texte. Par ailleurs, des règles heuristiques ont été formulées, qui, en fonction de certaines caractéristiques contextuelles, associent des valeurs sémantiques spécifiques, extraites du réseau des valeurs, aux occurrences des marqueurs linguistiques de la causalité rencontrées dans le texte.

A titre d'exemple, dans l'énoncé qui suit, l'indice déclencheur *aura ... un effet sur* et l'indice complémentaire *être certain que* illustrent le lien causal établi et la modalité qui l'accompagne :

(9) *Qui peut être certain, en effet, que l'absorption à haute dose des niaiseries qui constituent 90% des programmes, n'**aura** pas, à moyen ou long terme, **un effet** délétère irréversible sur la conscience même des téléspectateurs ?* (*Le Monde diplomatique, juin 1996, p. 32)*

Cet ensemble de connaissances, une fois intégré dans la base de données de la plate-forme FilText, peut être exploité de différentes manières. L'application principale vise l'identification d'énoncés causaux pertinents de façon à les extraire des corpus textuels analysés et, ultérieurement, à les intégrer éventuellement dans les résumés construits avec un certain objectif. Dans le cas, plus général, du filtrage sémantique, il s'agit d'offrir à l'utilisateur une grande marge de manœuvre et un maximum de souplesse dans la sélection et dans les associations des critères filtrants retenus. L'association de critères permet de jouer sur la co-présence ou sur la présence exclusive dans un espace de recherche de certaines informations identifiées. Ainsi, pour résumer un article scientifique, il est préférable de privilégier les approches analytique et fonctionnelle de la causalité (exemple 10), car elles permettent de rendre plus finement compte des déterminations causales multiples ou complexes. En revanche, dans le résumé d'une étude d'opinion, certaines valeurs de la prise en charge énonciative s'imposeront comme étant des éléments contextuels indispensables qui doivent nécessairement accompagner l'expression des relations causales extraites et insérées dans le résumé (exemple 11).

*(10) Selon Andrew A. Monjan, directeur de la section Neurobiologie du vieillissement à l'Institut national de la vieillesse, à Bethesda (Maryland), « l'étude montre que l'âge n'**est** pas **le facteur** critique **dans** la baisse de production de mélatonine par le cerveau. Il faut plutôt **impliquer d'autres facteurs comme** des maladies, l'usage de médicaments et toute une série de problèmes qui accompagnent parfois la vieillesse ».*

*(11) Beaucoup d'Européens (52 %) croient que l'Union européenne **jouera un rôle** plus important **dans** leur vie quotidienne au début du siècle prochain et 32 % pensent qu'elle **jouera le** même **rôle**. En outre, la plupart des Européens souhaiteraient que l'Union **joue un rôle** plus important (48 %) ou identique (27 %) **dans** leur vie quotidienne. Très peu de personnes considèrent que l'UE va **jouer un***

***rôle** moins important (7 %) **dans** leur vie quotidienne au XXI^e^ siècle et 14 % le souhaitent. (« Le processus d'intégration européenne», rapport Eurobaromètre.)*

4.2. *Filtrage des phrases étiquetées*

Certaines phrases étant étiquetées, il devient possible de construire des extraits qui répondent aux besoins spécifiques d'un utilisateur en appliquant différentes stratégies de sélection. Pour cela, des stratégies de sélection paramétrées par P, SE et SF ont été définies.

Le profil de filtrage P est soit prédéfini, soit déterminé par l'utilisateur. Il précise 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 alors sous la forme d'une liste hiérarchisée d'étiquettes, ce qui permet aussi d'ignorer un certain type d'informations. La profondeur d'exploration permet 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.

La stratégie d'exploration SE précise l'ordre d'exploration des sections et la profondeur d'exploration (et donc de sélection) 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. Pour chaque section du texte, il est possible de préciser un profil de filtrage P.

Le seuil de filtrage SF permet de produire un ensemble de phrases qui correspondent à un nombre fixé de phrases relativement à la taille du texte source, par exemple le seuil de sélection peut être de 10 % ou 20 % du texte source.

La construction d'extraits par extraction de phrases issues du texte source est confrontée à plusieurs problèmes. Le premier concerne la présence de termes anaphoriques, comme dans la phrase « *Ceci est notre principale conclusion.* », étiquetée phrase *conclusive*, mais dont l'apport informationnel est faible. Ce problème est résolu par l'utilisation d'une heuristique simple mais robuste : la détection de termes potentiellement anaphoriques dans une phrase P comme *ceci, cette, il, etc.*, combinée avec l'utilisation d'une liste d'exceptions comme *il semble, il se peut, etc.*, déclenche l'insertion, dans l'extrait, de la phrase précédant la phrase P. L'application de cette heuristique permet de résoudre plus de 80% des cas [MIN 97]. 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. Nous atteignons là une des limites des systèmes de résumé automatique par simple extraction de phrases [MIN 00].

5. Conclusion

La difficulté théorique du résumé automatique vient de ce qu'il n'existe pas une typologie générale des textes et que, pour un texte donné, la notion de « résumé idéal » qui pourrait satisfaire tous les utilisateurs, n'a finalement pas grande signification. C'est pourquoi certaines recherches sur l'automatisation du résumé se sont orientées vers un filtrage sémantique destiné à synthétiser des informations extraites des textes, ce filtrage étant guidé par l'utilisateur. Cependant, cette extraction brise la cohérence du texte source et peut même introduire des contresens. Le développement d'interfaces de communication qui permettent à l'utilisateur de naviguer entre l'extrait et le texte source partiellement étiqueté constitue une des réponses offertes par ContextO. 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. Nous travaillons actuellement à l'intégration d'autres agents spécialisés et à l'exploitation des textes annotés pour pouvoir utiliser les informations morpho-syntaxiques.

Remerciements

La réalisation de la plate-forme logicielle est le fruit d'une collaboration entre l'équipe LaLIC du CAMS (UMR 8557 du CNRS, EHESS, Université Paris-Sorbonne) et l'Université de la République (Uruguay) ; elle a reçu le soutien du programme ECOS-Sud (Action n° U97E01).

6. Bibliographie

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

[BER 96a] BERRI J., 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, 1996.

[BER 96b] BERRI J., CARTIER E., DESCLES J-P., JACKIEWICZ A., MINEL J-L., SAFIR, système automatique de filtrage de textes, *Actes du colloque TALN'96*, p. 140-149, Marseille, 1996.

[CAR 97] CARTIER E., La définition dans les textes scientifiques et techniques : présentation d'un outil d'extraction automatique des relations définitoires, *TIA'97*, p. 41-48, Toulouse, 1997.

[CHA 88] CHAROLLES M., Les plans d'organisation textuelle ; période, chaînes, portées et séquences, *Pratiques*, n° 57, p. 3-13, 1988.

[CHA 98] CHAROLLES M., L'organisation du texte, le filtrage et le résumé, *RIFRA'98, Rencontre Internationale sur l'Extraction, le Filtrage et le Résumé Automatiques*, p. 20-21, Sfax, Tunisie, 1998.

[CHA 99] CHAROLLES M., Phrase, texte, discours, *Langue Française*, n° 121, p. 76-116, 1999.

[DES 91] DESCLES J-P., JOUIS C., OH H-G., MAIRE REPPERT D.,.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, (ed. D. Herin-Aime, R. Dieng, J-P. Regourd, J.P. Angoujard), Amsterdam, 1991.

[DES 97a] DESCLES J-P, CARTIER E., JACKIEWICZ A., MINEL J-L., Textual Processing and Contextual Exploration Method, *CONTEXT'97*, p. 189-197, Rio de Janeiro, Brésil, 1997.

[DES 97b] DESCLES J-P., *Systèmes d'exploration contextuelle. Co-texte et calcul du sens*. (ed. Claude Guimier), Presses Universitaires de Caen, p. 215-232, 1997.

[ELL 98] ELLOUZE, M., BEN HAMADOU A., 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*, p. 108-115, Sfax, Tunisie, 1998.

[END 95] ENDRES-NIGGEMEYER B., MAIER E., SIGEL A., How to implement a naturalistic model of abstracting : four core working steps of an expert abstractor, *Information Processing & Management*, 31(5), p. 631-674, 1995.

[FAY 89] FAYOL M., Le résumé : un bilan provisoire des recherches en psychologie cognitive, *Actes du colloque international de linguistique (aspects linguistiques, sémiotiques, psycholinguistiques et automatiques)*, (ed. Charolles, Petitjean), Pont-à-Mousson, Paris, Klincksieck, p. 163-182, 1989.

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

[HER 96] HERVIOU M-L., QUATRAIN R., MONTEIL M-G., Construction de terminologies : une chaîne de traitement supportée par un atelier intégrant outils linguistiques et statistiques, *TALN'96*, p. 130-139, Marseille, 1996.

[JAC 98] JACKIEWICZ A., 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, 1998.

[JIN 98] JING H., BARZILAY R., MCKEOWN K., Summarization Evaluation Methods : Experiments and Analysis in *Symposium on Intelligent Text Summarization ACL*, Stanford, CA, 1998.

[JOU 93] JOUIS C., Contribution à la conceptualisation et à la modélisation des connaissances à partir d'une analyse linguistique de textes, Thèse de doctorat, EHESS, Paris, 1993.

[LEH 95] LEHMAM A., Le résumé de textes techniques et scientifiques, aspects linguistiques et computationnels, Thèse de doctorat, Université de Nancy 2, 1995.

[LEP 99] LE PRIOL F., A data processing sequence to extract terms and semantics relations between terms, *10th mini EURO Conference Human Centered Processes HCP'99*, p. 241-248, Brest, 1999.

[LER 94] LEROUX D., MINEL J-L., BERRI.J., SERAPHIN project, *First European Conference of Cognitive Science in Industry*, p. 275-283, Luxembourg, 1994.

[MAN 88] MANN W C., THOMPSON S. A., Rhetorical Structure Theory : Toward a functional theory of text organization, *Text*, 8(3), p. 243-281, 1988.

[MAR 97] MARCU D., From discourse structures to text summaries in *Workshop Intelligent Scalable Text Summarization ACL*, p. 82-88, Madrid, Espagne, 1997.

[MAS 98] MASSON N., Méthodes pour une génération variable de résumé automatique : Vers un système de réduction de textes, Thèse de Doctorat, Université Paris-11, 1998.

[MEU 98] MEUNIER J-G., La gestion des connaissances et les intelliciels, *RIFRA'98, Rencontre Internationale sur l'Extraction, le Filtrage et le Résumé Automatiques*, p. 4-5, Sfax, Tunisie, 1998.

[MII 94] MIIKE E S., ITOH E., ONO K., SUMITA K., A full-text retrieval system with a dynamic abstract generation function, *Proceedings Sigir'94*, p. 152-161, Springer-Verlag, Dublin, 1994.

[MIN 97] MINEL J-L., NUGIER S., PIAT G., How to appreciate the Quality of Automatic Text Summarization, *Workshop Intelligent Scalable Text Summarization, EACL*, p. 25-30, Madrid, Espagne, 1997.

[MIN 00] MINEL J-L., DESCLES J-P., *Résumé Automatique et Filtrage des textes*, in *Ingénierie des langues*, Paris, Editions Hermès, 2000.

[MOU 99] MOURAD G., La segmentation des textes par l'étude de la ponctuation, *CIDE'99*, p. 155-171, Damas, Syrie, 1999.

[PAI 81] PAICE C. D., The automatic generation of literature abstracts : an approach based on the identification of self indicating phrases, *Information retrieval research*, p. 172-191, 1981.

[PAZ 97] PAZIENZA M.T., (éd.), Information extraction (a multidisciplinary approach to an emerging information technology*), International Summer School, SCIE'97*, Springer Verlag (Lectures Notes in Computer Science), 1997.

[RAT 61] RATH G.J., RESNICK A., SAVAGE T., The formation of abstracts by the selection of sentences, *American Documentation*, 12, (2), p. 139-143, 1961.

[REB 98] REBEYROLLE J., PERY-WOODLEY M-P., 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*, p. 19-30, Sfax, Tunisie, 1998.

[SAL 83] SALTON G.M., *Introduction to Modern Information Retrieval*, Mac Graw Hill Book Co, New York, 1983.

[SAG 98] SAGGION H., LAPALME G., Where does information come from ? Corpus Analysis for Automatic Abstracting, *RIFRA'98, Rencontre Internationale sur l'Extraction, le Filtrage et le Résumé Automatiques*, p. 72-83, Sfax, Tunisie, 1998.

[SPA 93] SPARCK JONES K., What might be in a summary ?, *Information Retrieval 93*, p. 9-26, Universitates Verlag Konstanz, 1993.

[TEU 97] TEUFEL S., MOENS M., Sentence extraction as a classification task, *Workshop Intelligent Scalable Text Summarization, EACL*, p. 58-65, Madrid, Espagne, 1997.

Annexe

Résumé automatique de l'article 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 de la revue TSI.

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.

Article reçu le 7 juin 1999
Version révisée le 3 janvier 2000
Rédacteur responsable : Daniel KAYSER

Jean-Luc Minel *est ingénieur de recherche en informatique au CNRS. Il poursuit actuellement ses recherches dans l'équipe LaLIC du CAMS ; celles-ci portent sur l'élaboration de méthodes et de modèles de représentation des connaissances linguistiques pour l'extraction d'informations dans des corpus informatisés. Il s'intéresse plus particulièrement au résumé automatique et au filtrage sémantique des textes.*

Jean-Pierre Desclés *est professeur (informatique et linguistique) à l'Université de Paris-Sorbonne (Paris IV). Ses travaux de recherche concernent la linguistique informatique, la linguistique théorique, la sémantique cognitive, et l'extraction des connaissances dans les textes par des méthodes sémantiques. Il est responsable de l'équipe LaLIC (Langages, Logiques, Informatique et Cognition) rattachée au CAMS.*

Emmanuel Cartier *termine un doctorat en linguistique à l'Université Paris-Sorbonne sur le repérage des définitions dans les textes. Il est actuellement ingénieur linguiste à Lexiquest. Ses travaux portent sur la construction de grammaires et sur l'extraction d'informations textuelles.*

Gustavo Crispino *est titulaire d'un M.Sc. en informatique de l'Université de la République en Uruguay. Il est actuellement professeur adjoint à l'Université de la République. Ses travaux portent sur le filtrage sémantique de textes.*

Slim Ben Hazez *termine un doctorat en informatique linguistique à l'Université Paris-Sorbonne. Ses travaux portent sur la modélisation des ressources linguistiques, le filtrage sémantique et l'extraction d'informations dans les textes.*

Agata Jackiewicz*, docteur en linguistique informatique, est maître de conférences à l'Université de Paris-Sorbonne. Ses travaux portent sur le filtrage automatique de textes et l'extraction d'informations par des méthodes sémantiques.*