L'identification des structures discursivesengendrées par les cadres organisationnels

Agata Jackiewicz, Jean-Luc Minel

L'identification des structures discursives engendrées par les cadres organisationnels

Agata Jackiewicz, Jean-Luc Minel

Laboratoire LaLICC
UMR 8139 (CNRS - Université Paris-Sorbonne)
Université de Paris-Sorbonne (ISHA)
96, bd Raspail, 75006 Paris - France
{prénom.nom}@paris4.sorbonne.fr

Résumé – Abstract

Cet article présente tout d'abord une analyse linguistique des cadres organisationnels et son implémentation informatique. Puis à partir de ce travail, une modélisation généralisable à l'ensemble des cadres de discours est proposée. Enfin, nous discutons du concept d'*indicateur* proposé dans le cadre théorique de l'exploration contextuelle.

To begin with, this paper outlines a linguistic analysis of textual enumerating frameworks and its computational making. Then, from this work, a modelling for all textual frameworks is suggested. Finally, we discuss the relevance of the concept of *clue* which is central in the theoretical framework of contextual exploration method.

Keywords – Mots Clés

Séries de cadres organisationnels, marqueurs d'intégration linéaire, cadres de discours, segmentation automatique de textes, méthode d'exploration contextuelle, filtrage automatique de textes.

Enumerating frameworks, linear integration markers, discourse frames, automatic text segmentation, contextual exploration method, automatic text filtering.

1 Introduction

Un grand nombre de travaux dans le domaine de l'extraction d'informations se sont concentrés sur le repérage d'informations pertinentes, avec pour résultat visé, la construction d'un texte (ou d'une fiche) indépendant du texte source traité (Pazienza 1997, MUC 2002). Les systèmes de résumé automatique (Mani 2001, Minel & al. 2001), et plus généralement les systèmes dédiés à la fouille de textes, qui cherchent par exemple à 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 définitions d'un concept, etc. illustrent bien ce type d'approche (Laublet & al. 2002). La plupart de ces systèmes s'appuient sur des algorithmes de reconnaissance de marques de surface, attribuant un score ou une « étiquette » aux phrases du texte, puis extrayant certaines de ces phrases pour construire un fragment textuel. Une des difficultés à laquelle sont confrontés ces systèmes est de fournir le contexte de validité (par exemple de vérité ou de prise en charge) de la phrase placée dans le fragment textuel. En effet, une information causale ou une description d'un événement n'auront pas le même intérêt pour un utilisateur donné, suivant qu'elles sont entièrement prises en charge par l'auteur du texte, attribuées à un autre énonciateur avec une marque de distanciation de la part de l'auteur, ou plus simplement placées à l'intérieur d'un univers spatial ou temporel limité (Jackiewicz 2000). Isoler une connaissance de son cadre de validité revient à occulter une part essentielle de son contenu, au risque de lui faire perdre sa valeur opératoire.

Plus généralement, le problème que rencontre aujourd'hui le traitement automatique des textes quelle que soit sa visée applicative ultérieure tient au fait que l'on ne sait pas comment les découper en unités de traitement. La prise en considération des indicateurs graphiques et typographiques constitue un préalable, mais les phrases se révèlent bien souvent des unités trop étroites et les paragraphes ou les sections, des unités trop vastes. Si certains textes sont structurés très finement par des moyens typographiques et dispositionnels (comme c'est le cas des textes rédigés pour les pages Web), d'autres sont entièrement linéaires et les idées qui y sont contenues sont organisées uniquement par le discours lui-même. Le fait de pouvoir segmenter les textes en prenant appui également sur l'organisation discursive des contenus représente un apport significatif, quel que soit l'objectif du traitement. Lors du filtrage des textes, par exemple, un fragment extrait peut parfaitement être intégré dans une série de segments discursifs corrélés entre eux. S'il n'est pas toujours pertinent de présenter à l'utilisateur l'ensemble de la structure qui comprend le segment ciblé, il est sans doute intéressant de lui offrir la possibilité d'y accéder dynamiquement et de l'assister dans le parcours des éléments liés au segment en question. Le problème des séries dans le discours intéresse également certains travaux sur le résumé de textes, où la lisibilité des extraits produits automatiquement est soit évaluée, soit censée être améliorée, dans des développements ultérieurs, par le recours aux liens de cohésion instaurés par les marqueurs d'intégration linéaire.

Les différents travaux menés dans le domaine de la fouille de textes font apparaître que l'on peut fonder cette activité sur deux grands types de traitement: (i) le repérage des thèmes traités dans le texte (Salton & al. 1996, Boguraev & al. 2000, Roussel 2002) et /ou des séquences thématiquement homogènes (Ferret & al. 1998), (ii) l'identification dans les textes de marqueurs linguistiques spécifiques qui correspondent à des points de vue et attentes d'un utilisateur potentiel (Laublet & al. 2002). Dans les deux cas, il est indispensable de pouvoir

s'appuyer sur la structure du texte, non seulement pour calculer les valeurs sémantiques des segments interprétés, mais aussi pour pouvoir parcourir le texte fouillé en suivant l'organisation de ces segments.
Toutes ces activités passent du point de vue du traitement automatique par la spécification d'un cadre méthodologique, la construction de modèles (de représentation des connaissances linguistiques, de représentation d'un texte), et la réalisation d'une plate-forme logicielle qui implémente ces modèles. C'est pourquoi, dans ce domaine, les recherches du laboratoire LaLICC se sont orientées selon plusieurs axes :

- le développement d'un cadre méthodologique fondé sur l'exploration contextuelle (Desclés & al. 1991, Desclés & al. 1997) ;
- la construction des modèles et des outils de représentation d'un texte (Minel & al. 2001, Ben Hazez 2002, Crispino 2003);
- l'étude fine de certaines structures discursives (Porhiel 2001, Jackiewicz 2002) ;
- le développement d'outils de navigation et de gestion des interactions entre l'utilisateur et le système de filtrage sémantique (Couto 2001, Minel 2003).

Le projet global qui sous-tend ces recherches vise à intégrer dans une même plate-forme les connaissances linguistiques qui permettent de construire une représentation du texte et les outils logiciels qui fournissent les moyens à l'utilisateur de se focaliser sur les fragments textuels qui répondent à ses besoins. Ce travail s'insère dans le courant qui vise à élaborer des outils d'analyse textuelle, comme la RST (Mann & al. 88) ou le modèle d'architecture textuelle (Luc & al. 2001 , Péry-Woodley 2000), mais il s'en démarque notamment sur deux points. Premièrement, nous ne cherchons ni à interpréter finement, ni à hiérarchiser les relations entre les structures discursives identifiées, à l'exception du traitement des enchâssements entre celles-ci. Notre modèle est donc plus simple que ceux que nous venons de citer. Mais, et c'est notre deuxième point, nous nous assignons une opérationalisation effective, fondée sur des connaissances linguistiques fines et supposées indépendantes du domaine dont traite le texte.

Afin d'illustrer la pertinence des recherches sur la structuration discursive des textes pour discuter et enrichir les cadres méthodologique et informatique des travaux sur la fouille des documents textuels, nous présenterons très brièvement les résultats d'une étude linguistique portant sur les séries dans le discours (Jackiewicz 2002), puis après une rapide description de l'implémentation de ces résultats dans la plate-forme ContextO, nous explicitons le modèle de référence associé à la structure en série, et proposons de généraliser cette démarche pour le traitement de plusieurs autres structures fondées sur la notion de cadre de discours.

2 Les séries dans le discours

2.1 Cadre théorique et démarche

Le travail linguistique a été mené dans la perspective textuelle de Michel Charolles (Charolles 1997, 2002) à qui nous empruntons les notions de *cadre de discours* et de *portée*. La notion générale de *cadre* sert à désigner les circonstances dans lesquelles il faut envisager un certain état ou une série d'événements. Les cadres de discours contribuent ainsi à partitionner l'information dans des rubriques homogènes. Ils instaurent un lien de cohésion textuelle (Halliday & al. 1976) que le lecteur reconstruit à partir de nombreux indices et en particulier en s'appuyant sur les *introducteurs de cadres*. La *portée* de ces marques, syntaxiquement non

intégrées à l'énoncé où elles figurent matériellement, généralement en position initiale, peut s'étendre sur plusieurs phrases créant ainsi une véritable unité textuelle, homogène thématiquement et relativement autonome par rapport au contexte. Parmi les expressions cadratives ayant un fort pouvoir intégrateur, une famille de marques de nature métalinguistique jouent le rôle d'organisateurs textuels par excellence : ce sont les marqueurs d'intégration linéaire (MIL). Les MIL ont la propriété caractéristique d'être indépendants des contenus sémantiques des segments qu'ils introduisent et relient entre eux sur le mode d'une série, créant ainsi une véritable structure textuelle.

Partant de l'observation de quelques MIL les plus fréquents, introducteurs des séries homogènes et régulières (*d'une part / de l'autre* ; *en premier lieu / en second lieu /… ; premièrement / deuxièmement* /…), nous avons cherché à répertorier l'ensemble des possibilités discursives disponibles en français pour marquer une série. Plus de 180 séries différentes ont été relevées dans le corpus[1] ; nous en avons étudié la longueur, la structure, la portée des items, le co-texte et les propriétés distributionnelles des introducteurs, ainsi que l'amorce (ou l'énoncé qui annonce la série). Une série est donc caractérisée par l'ensemble de ces éléments. De cette façon, nous avons mis en évidence les données linguistiques permettant (i) de repérer les MIL, les marques qui les supportent étant polysémiques et polyfonctionnelles dans le discours, (ii) de délimiter les segments textuels que ces organisateurs permettent d'introduire, (iii) de cadrer la totalité de la série (en trouver le début, c'est-à-dire délimiter le segment introducteur ou amorce et en fixer la fin, ce qui implique le repérage de la fermeture du segment textuel introduit par le dernier MIL de la série).
Le recueil et l'organisation de ces données ont été guidés par la méthode d'exploration contextuelle utilisée depuis de nombreuses années dans diverses applications de type TAL par l'équipe LaLICC (Desclés & al. 1991, 1997). Cette méthode consiste à identifier dans un texte, en premier lieu, les unités linguistiques appelées *indices pertinents* ou *indicateurs* qui sont significatifs pour la caractérisation (étiquetage) des segments en fonction d'une tâche donnée, et ensuite de rechercher, avec des règles spécifiques, des *indices complémentaires* dans le contexte des indices pertinents, indices complémentaires qui vont permettre la prise de la décision adéquate. Une partie des connaissances linguistiques relatives aux séries discursives est organisée dans un modèle conceptuel et exploitée au sein de la plate-forme informatique Context0 (Minel & al. 2001).

2.2 Des marqueurs d'intégration linéaire aux séries des cadres organisationnels

Les marqueurs d'intégration linéaire segmentent le texte en rendant perceptible sa configuration. L'instruction fournie par un MIL dit que le segment discursif qu'il introduit est à intégrer de façon linéaire dans une série (Turco & Coltier 1988). Cet enchaînement s'établit en général entre unités textuelles du même niveau ce qui, sur le plan sémantique, revient à dire que les constituants mis en rapport doivent être traités dans un même mouvement interprétatif. Chaque marqueur d'intégration linéaire constitue à lui tout seul une expression cadrative à même d'indexer plusieurs propositions (autrement dit, chaque MIL introduit un

[1] Le Monde Diplomatique sur CD-ROM (12520 articles et documents parus entre 1984 et 1998) et la base FRANTEXT (3650 œuvres, soit 214 millions de mots).

cadre organisationnel). A l'écrit, ce pouvoir intégrateur peut s'étendre sur plusieurs paragraphes successifs.

La longueur des séries balisées par les marqueurs d'intégration linéaire varie dans nos corpus entre deux et dix éléments, avec la distribution suivante : {2 items :70 séries ; 3 :70 ; 4 :25 ; 5 :6 ; 6 :5 ; 7 :3 ; 8 :1 ; 9 :1 ; 10 :1}. Ainsi, 75% de ces séries sont composées de deux ou de trois segments, ce qui nous renseigne sur la longueur typique de ces séries. Ce résultat dévoile à son tour une propriété saillante des séries courtes, à savoir l'hétérogénéité de leurs introducteurs. Cette caractéristique tient au fait que les introducteurs particuliers issus des séries d'origine temporelle, spatiale ou « numérique » peuvent se combiner entre eux (exemples : *premièrement / en deuxième lieu / enfin* ou *tout d'abord / deuxième facteur / en troisième lieu / enfin, dernier élément)*. Notons que le balisage des séries peut également être incomplet, avec un ou deux introducteurs manquants. En analysant l'ensemble des séries triées par la longueur, on constate que plus une série est longue, plus ses introducteurs tendent à être homogènes (appartenir à la même série originelle), ce qui explique le très faible nombre de possibilités effectivement exploitées pour introduire une série longue.

La mise en texte d'une série dans le discours n'appelle pas de disposition visuelle spécifique. Plusieurs cas de figure sont possibles, allant du paragraphe compact qui contient à la fois le segment amorce et l'ensemble des éléments de la série s'enchaînant linéairement, à une suite parallèle de paragraphes nettement distincts visuellement, introduits chacun par un MIL en position initiale et par une marque graphique (tiret, puce, numéro). Le marquage mixte (organisation « visuelle » + marquage discursif) peut prendre dans sa réalisation systématique deux formes différentes : (i) le marquage *redondant* (ou renforcé) où chaque marqueur discursif est doublé d'une marque typographique et/ou d'une marque dispositionnelle ; (ii) le marquage *complémentaire* où chaque famille de marques balise un niveau différent dans une structure complexe.

Si les énumérations structurées au moyen des marques typographiques et dispositionnelles peuvent être organisées à plusieurs niveaux, les séries balisées par les MIL semblent en privilégier un seul. Il existe toutefois des exceptions. Le corpus du MD fournit plusieurs séquences balisées par les MIL présentant un enchâssement d'items (*tout d'abord…(d'une part… d'autre part…) enfin…*). On constate par ailleurs que plusieurs séries courtes peuvent s'enchaîner directement l'une à la suite de l'autre. L'énoncé introducteur indexe alors explicitement l'ensemble des séries qui se suivent et précise la nature du lien existant entre elles. Il peut s'agir d'une opération de *parenthésage* qui marque la formation de sous-ensembles d'items ((1,2), (3,4)), ou d'une *reformulation* où une dichotomie entre items est reprise sous un éclairage différent ((1,2), (1',2')).

Les marqueurs d'intégration linéaire introduisent plusieurs segments textuels qui s'interprètent comme sémantiquement équivalents. L'idée fédératrice qui relie ces segments entre eux est généralement explicitée dans un segment textuel (ou amorce) précédant la série. Parmi les marques qui renvoient au principe fédérateur de la série, il y a des classifieurs *(éléments, étapes, causes, raisons…)*. Ces éléments, qui ne peuvent être définis qu'en extension, sont souvent repris au niveau de l'expression cadrative *(elle pourrait être précipitée par trois éléments : premièrement… deuxième élément… troisième élément…).* Le segment amorce donne également une indication sur le nombre d'items de la série. La longueur peut être indiquée (i) précisément (*deux, double…*) ou d'une manière approximative (*plus d'un*, *nombreux…*), ces deux types de marques se combinant avec le classifieur, (ii) ou plus implicitement, par des expressions telles que *un paradoxe, une dichotomie…* Il n'existe pas toujours d'annonce explicite, mais la séquence textuelle qui précède la série est réellement

indispensable à son interprétation (par exemple, quand elle exprime une thèse que les différents items de la série ont pour charge d'étayer).

La fermeture du dernier cadre organisationnel de la série, comme c'est le cas de tous les cadres de discours, n'est pas marquée explicitement. Des complexes de plusieurs indices (de nature thématique, graphique...) en général liés à l'ouverture d'une nouvelle structure, amènent le lecteur à fermer ce dernier segment et clore ainsi toute la série. Nous avons constaté toutefois que les séries discursives étaient fréquemment suivies par une évaluation rétrospective portant sur l'ensemble de leurs items ; cette évaluation est introduite par des marques comme (*l'ensemble, les deux, tous*…).

Nous avons ainsi dégagé une structure discursive composée d'une amorce et d'une suite de segments textuels formant une série, qui en dépit de certaines variations présente des caractéristiques stables permettant un traitement automatique de la dite structure.

3. Repérage et délimitation des séries discursives

3.1 Organisation des données

Les connaissances relatives aux séries discursives capitalisées dans la base ContextO se répartissent dans deux grands ensembles. Le premier est représenté par les introducteurs des items formant à l'origine 182 séries différentes. Dans chacune de ces séries, le premier MIL est considéré comme un indicateur, les suivants ont le statut d'indice complémentaire. Précisons que l'organisation informatique des MIL de rang supérieur ou égal à deux (MIL_2, MIL_3…) est fondée sur la constitution de paradigmes (classes) dont les éléments sont tous les MIL pouvant occuper respectivement le rang 2, le rang 3… dans la série ouverte par le MIL de rang 1 (indicateur). Cette organisation, illustrée dans le tableau 1 pour le marqueur *d'abord*, augmente considérablement la combinatoire des MIL pouvant s'enchaîner dans une série, sans pour autant s'autoriser à combiner librement tous les MIL_1 avec tous les MIL_2… jusqu'au MIL_n (avec n=10). L'union de l'indicateur et des indices constitue la signature de la série.

MIL1 indicateur	MIL2	MIL3	MIL4	MIL5
&d_abord	&d_abord2	&d_abord3	&d_abord4	&d_abord5
d'abord	*d'un autre côté, dans un second temps, …*	*dans un troisième temps, le troisième, …*	*quatrièmement, quatrième, …*	*cinquième*

Tableau 1 : extrait de la table des MIL pour les séries ouvertes par *d'abord*.

Le deuxième ensemble réunit les données impliquées dans la désambiguïsation des formes, ainsi que dans le cadrage de la totalité de la structure nécessitant d'une part l'identification de l'énoncé introducteur, et d'autre part le repérage de la fermeture du dernier item de la série. Tous ces marqueurs font partie des indices complémentaires. Les indices qui interviennent dans l'interprétation d'une occurrence de forme en tant que marqueur d'intégration linéaire ont trait à la position et aux éléments pouvant figurer dans le co-texte immédiat d'un MIL. Le critère de position spécifie qu'un MIL peut figurer soit à l'initiale de la phrase, soit dans une incise délimitée par deux virgules. Si ce critère n'est pas vérifié, on considère la nature de l'élément qui précède immédiatement le MIL. Parmi les indices autorisés à figurer dans le

contexte gauche d'un MIL citons (i) les signes de ponctuation tels que deux-points ou point-virgule, (ii) certains connecteurs (*mais, et, surtout…*) ou marqueurs de clôture (*enfin, en dernier lieu*), (iii) divers marqueurs appartenant à la liste {*si, c'est…}*, (iv) les marques graphiques d'énumération telles que tirets, puces, numéros. Le traitement de l'énoncé introducteur dans sa forme explicite s'appuie sur l'identification de ses éléments saillants : le classifieur et le cardinal (indiquant la longueur de la liste). Notre liste des classifieurs comprend 172 formes. Les indicateurs de longueur sont au nombre de 50; ils sont classés en deux sous-ensembles selon leur aptitude à se combiner ou non avec le classifieur (*deux* éléments / *une dichotomie* ∅). Les indices lexicaux de la fermeture du dernier item de la série sont quant à eux très peu nombreux (*L'ensemble, Les deux…, Tous les*…), leur position (dans la phrase et par rapport au dernier MIL) est fortement contrainte. Le processus de l'identification de l'ensemble de la série des cadres organisationnels fait appel à plusieurs règles heuristiques qui doivent respecter les contraintes imposées par le modèle sous-jacent à cette structure. Ainsi, le repérage de l'indicateur d'une série déclenche le processus de recherche des indices situés nécessairement en aval. L'étendue de cette recherche est bornée par l'observation empirique que la portée d'un MIL n'excède généralement pas deux paragraphes. Toutes ces heuristiques sont implémentées sous la forme de règles d'exploration contextuelles (Minel & al. 2001). La reconnaissance d'une série entraîne la construction d'une instance de la classe *Cadre* qui vient enrichir le modèle de représentation du texte. Une description détaillée de ce modèle est donnée dans (Crispino 2003).

3.2 Modèle de référence

Le repérage des structures discursives nécessite deux niveaux d'analyse, ce qui le différencie des travaux jusqu'alors menés dans le cadre de l'exploration contextuelle. Un premier niveau vise à identifier la fonction discursive (ou sémantique) d'un marqueur. Par exemple, l'adverbe *tout d'abord,* suivant sa position dans la phrase peut appartenir ou non à une structure discursive plus complexe (exemples 1 et 2).

1) *Raskolnikov ne la reconnut pas* ***tout d'abord****. C'était Sonia Semionovna Marmeladova* (Dostoïevski, Crime et châtiment…).

2) *La dramatique situation du Mexique montre aujourd'hui que le problème de la dette demeure entier. Avec cette circonstance aggravante que le dispositif imaginé entre 1982 et 1984 arrive à expiration. Pour trois raisons:*
- ***tout d'abord,*** *la récession envisagée comme méthode universelle pour ramener les compteurs d'un pays à zéro n'est plus acceptée ni acceptable par les pays du tiers- monde qui n'en voient pas la fin. (…) ;*
- de plus, l'approche "au cas par cas", telle que le FMI l'a conçue et pratiquée, a également tourné à l'échec. (…);
- enfin, troisième et dernier point, les rééchelonnements pluriannuels sont devenus inopérants. (Le Monde diplomatique)

Ce premier niveau mobilise des connaissances linguistiques qui s'expriment sous forme de règles d'exploration contextuelle dont l'espace de recherche est limité à la phrase. Un deuxième niveau d'analyse va conjuguer le repérage issu du premier niveau avec des contraintes exprimées dans un modèle propre à la structure discursive que l'on cherche à identifier. Par exemple, pour les séries des cadres organisationnels, ce modèle stipule les contraintes suivantes :

- Les éléments de la structure obéissent à une organisation séquentielle (*en second lieu* apparaît nécessairement après en *premier lieu* et avant en *troisième lieu*) sachant que certaines marques lexicales peuvent être absentes. L'ensemble de ces éléments constitue une série paradigmatique qui possède une signature. La signature permet de distinguer deux séries paradigmatiques qui ont des introducteurs communs (par exemple, l'introducteur *deuxièmement* appartient à plusieurs séries).
- Deux séries de même signature sont nécessairement disjointes ;
- Deux séries de signature différente sont soient disjointes, soient enchâssées.

Ces contraintes vont être utilisées par le processus de reconnaissance pour calculer la portée et l'enchâssement des structures discursives notamment afin de pallier l'absence de certaines marques typographiques ou lexicales comme l'illustre l'exemple (3) :

3) Cette proposition amène plusieurs remarques. En premier lieu, ... En second lieu ..., on peut ainsi distinguer différents cas ... D'une part,.... D'autre part, ... Enfin,... En troisième lieu, ... Enfin ...

La première occurrence de *Enfin* peut appartenir à la structure discursive ouverte par en *Premier lieu* ou par *D'une part.* Faute de marques typographiques spécifiques (comme par exemple des « puces »), cette indétermination peut être résolue en se référant aux contraintes exprimées dans le modèle. En effet, *En troisième lieu* appartient nécessairement à une série déjà ouverte ce qui implique que la première occurrence de *Enfin* ferme la série ouverte par *D'une part*. C'est le modèle qui stipule qu'une ouverture de cadre ferme nécessairement le cadre précédemment ouvert.

A chaque structure discursive est donc associé un modèle qui décrit son organisation paradigmatique, les contraintes qui régissent les éléments qui appartiennent à la structure et celles qui régissent les structures entre elles. Ces contraintes sont propres à chaque modèle. Ainsi, l'étude menée sur les cadres thématiques (Porhiel 2001) semble montrer que ce type de structure n'est jamais enchâssé, alors que le niveau d'enchâssement des cadres spatiaux ou temporels peut être d'un niveau supérieur à deux (Charolles 1997). Précisons enfin que ces contraintes n'ont pas de finalité normative. Ainsi, il n'est pas exclu de rencontrer des textes dans lesquels, l'auteur pour des raisons stylistiques ou rhétoriques ou tout simplement à la suite d'une erreur, ne respecte pas ces contraintes. Nous postulons simplement que ces exceptions sont des artefacts qu'un système de traitement automatique peut exclure de son champ d'analyse.

Ces notions de modèle et de contraintes associées à celui-ci ont été implémentés dans ContextO pour l'identification des séries de cadres organisationnels dans un module spécifique. Nous travaillons actuellement à la description d'un langage de description qui permettrait de spécifier les modèles des différentes structures discursives (cadre temporel, cadre thématique, cadre de connaissances, etc.) afin de proposer un outil de description générique.

4. Discussion méthodologique

Sur le plan méthodologique, le processus d'identification des structures discursives interroge les principes fondateurs de la méthode d'exploration contextuelle. D'une part, la notion d'indicateur telle qu'elle a été jusqu'à présent considérée dans la méthode d'exploration contextuelle n'a plus la même pertinence. En effet, la méthode postule qu'il existe un

indicateur (marqueur le plus saillant) qui est un élément lexical, grammatical ou typographique. Dans l'étude sur les MIL (Jackiewicz 2002), c'est le premier élément de la série qui a été considéré comme un indicateur, ce qui engendre nécessairement des silences. En fait, aucun élément ne peut être choisi comme indicateur, car l'étude montre que chaque élément pris séparément peut être absent d'une réalisation textuelle d'une série de cadres organisationnels. C'est donc un complexe composé de différentes marques linguistiques, ce complexe pouvant être différent pour diverses réalisations, qui constitue l' « indicateur ». D'autre part, comme l'avait déjà montré les travaux menés dans le projet Régal (Ferret & al. 2001) l'identification de la portée d'un introducteur de cadre peut rarement s'appuyer sur des marques linguistiques de surfaces fiables. La solution adoptée dans ce projet a consisté à rechercher des ruptures d'ordre « thématique », mesurées à partir de la variation de la fréquence lexicale entre deux segments textuels. En revanche, l'étude sur les séries dans le discours montre que des marques lexicales permettent parfois d'identifier leur fermeture. Nous travaillons actuellement à une modélisation qui permette de spécifier pour chaque structure discursive le type de calcul qui doit être mis en œuvre.

Références

Ben Hazez, S. (2002). *Un modèle d'exploration contextuelle des textes : filtrage et structuration d'informations textuelles, modélisation et réalisation informatique (système SEMANTEXT)*, Université Paris-Sorbonne, Paris.

Boguraev, B. C., Neff M. (2000). Lexical cohesion, Discourse Segmentation and Document Summarization, *RIAO 2000,* p. 962-979.

Charolles, M. (1997). L'encadrement du discours - Univers, champs, domaines et espace , *Cahier de recherche linguistique*, 6.

Charolles, M. (2002). Organisation des discours et segmentation des écrits », in *Inscription Spatiale du Langage : structures et processus*, IRIT, Toulouse, 2002.

Couto J. (2001). *ContextO, Los sistemas de exploracion contextual de cara al usuario*, Mémoire de Master, Université de la République, Uruguay.

Crispino, G. (2003). *Thèse de doctorat,* en cours, Université de Paris-Sorbonne, Paris.

Desclés, J.-P., Jouis C., Oh H-G., Maire Reppert D. (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*, D. Herin-Aime, R. Dieng, J-P. Regourd, J.P. Angoujard (éds), Amsterdam, p. 371-400.

Desclés, J.-P., Cartier E., Jackiewicz A, Minel J.-L. (1997). Textual Processing and Contextual Exploration Method in *CONTEXT 97*, Universidade Federal do Rio de Janeiro, Brésil, p. 189-197.

Ferret O., Grau B., Masson N. (1998). Thematic segmentation of texts : two methods for two kind of texts , *Actes ACL-COLING'98*, Montréal, Canada, volume 1: 392-396.

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

Halliday M. et Hasan R. (1976). *Cohesion in English* Longman, New York.

Jackiewicz, A. (2000). « Causalité et prise en charge énonciative », in *Etudes Cognitives*, n°2, Académie Polonaise des Sciences, Varsovie.

Jackiewicz, A. (2002). Repérage et délimitation des cadres organisationnels pour la segmentation automatique des textes, *CIFT'02,* Hammamet*,* Tunisie, p. 95-107.

Luc Ch. et Virbel J. (2001). Le modèle d'architecture textuelle, fondements et expérimentation, *Verbum*, t. XXIII, n°1, « Cohérence et relations de discours à l'écrit ».

Laublet, P., Naït-Baha L., Jackiewicz A., Djioua B. (2002). Collecte d'informations textuelles sur le Web selon différents points de vue in *Interaction Homme-Machine et Recherche d'Informations*, Editions Hermès, Paris

Mani, I. (2001). *Automatic Summarization*, John Benjamins Publishing Company, Amsterdam.

Mann, W. C., Thompson S. A. (1988). Rhetorical Structure Theory : Toward a functional theory of text organization, *Text*, 8(3) p. 243-281.

Minel J.-L., Cartier E., Crispino G., Desclés J.-P., Ben Hazez S., Jackiewicz A. (2001). Résumé automatique par filtrage sémantique d'informations dans des textes, Présentation de la plate-forme FilText, *Technique et Science Informatiques,* n° 3, Paris, p. 369-396.

Minel J.-L. (2003). *Filtrage sémantique. Du résumé à la fouille de textes*. Hermès, Paris.

MUC (2002). Message Understanding Conferences
www.itl.nist.gov/iaui/894.02/related_projects/tipster/muc.htm.

Pazienza, M.T. (1997). *Information Extraction, a multidisciplinary approach to an emerging information technology*, *SCIE'97,* Springer Verlag, Notes in Computer Science.

Péry-Woodley, M.-P. (2000). *Une pragmatique à fleur de texte : approche en corpus de l'organisation textuelle*, Mémoire d'habilitation, Université de Toulouse- le Mirail.

Porhiel S. (2001). Linguistic expressions as a tool to extract thematic information , *Corpus Linguistic 2001*, Lancaster University, 2001.

Roussel. (2002). Navigation dans l'information par recomposition de documents et cartographie, *CIFT 02*, Hammamet, Tunisie, p. 27-42.

Salton G., Singhai A., Buckley C., Mitra M. (1996). Automatic text decomposition using text segments and texts themes. *In Seventh ACM Conference on Hypertext,* Washington D.C.

Turco G. et Coltier D., (1988). Des agents doubles de l'organisation textuelle, les marqueurs d'intégration linéaire, *Pratiques*, n°57.