TALN 2001, Tours, 2-5 juillet 2001

Une typologie des énumérations basée sur les structures rhétoriques et architecturales du texte

Christophe Luc
Institut de Recherche en Informatique de Toulouse - Université Paul Sabatier
118 route de Narbonne
31062 Toulouse Cedex
mél : luc@irit.fr

Résumé

Cet article concerne la caractérisation et la représentation de la structure interne des énumérations. Pour ce faire, nous utilisons deux modèles de texte : d'une part la Théorie des Structures Rhétoriques (RST) qui fournit un cadre d'interprétation pour la structure discursive des textes et d'autre part le modèle de représentation de l'architecture textuelle qui est principalement dédié à l'étude et à la représentation des structures visuelles des textes.
Après une brève présentation des modèles, nous nous concentrons sur l'étude de l'objet "énumérations". Nous exhibons et commentons trois exemples d'énumérations spécifiques que nous appelons des énumérations *non-parallèles*. Nous analysons la structure de ces énumérations et proposons un principe de composition des modèles de référence pour représenter ces énumérations. Enfin, nous présentons une classification des énumérations s'appuyant sur les caractéristiques de ces modèles.

mots-clés : structures textuelles, énumérations, représentation, classification, modèles de texte.

Abstract

This paper is related to the study and the representation of the internal structure of enumerations. We use two types of text structure: on one hand, the Rhetorical Structure Theory (RST) which provides a framework for interpreting and representing the discursive structures of texts and on the other hand, the model of text architecture which is is mainly dedicated to the study and representation of visuo-spatial structures of texts.
After a brief presentation of the theories and models, we focus the paper on the study of enumerations. We exhibit three significant examples of particular enumerations which are called non-parallel enumerations. We analyse these examples and their characteristics and we propose a general principle for the representation of these structures with these two reference models. Then, we present a classification of the different types of enumeration in the light of these models.

Keywords : Text structures, enumerations, representation, classification, text models.

Introduction

Les structures visuelles associées à tout texte écrit sont souvent perçues comme le résultat d'un traitement aval à celui du contenu. L'approche soutenue par Virbel (Virbel, 1989) suggère qu'au contraire, elles sont fortement indissociables du contenu et que l'étude de ces dernières doit être menée en relation avec d'autres types de structures textuelles (sémantiques, syntaxiques, rhétoriques, etc.). Il apparaît encore que des objets textuels de types particuliers (tels que les titres, définitions, etc.) ont des comportements spécifiques qui appellent la mise au point de représentations locales propres : la notion classique de classes de documents recoupe ainsi celle de classes d'objets, ces classes de documents étant régies par leur logique (syntaxique, sémantique et visuelle) propre.
L'étude présentée ici est une illustration de ceci, à propos de l'objet "énumération". Le choix de l'objet énumération se justifie car il existe peu d'étude sur cet objet alors qu'il est très répandu dans certains types de texte (textes scientifiques, textes à consignes, etc.) devant supporter des opérations de contrôle (bonne formation, cohérence, etc.). Par ailleurs, comme nous le montrerons, un certain nombre de ces énumérations présentent des structures textuelles complexes dont la caractérisation nécessite la prise en compte de données de nature diverse. Nous choisissons de représenter ces structures en relation avec deux modèles : la RST (Théorie des Structures Rhétoriques) représentant le contenu discursif d'un texte et le Modèle de représentation de l'Architecture textuelle (MAT) représentant un aspect des structures visuelles du texte.
Cette communication expose un cas de composition de ces deux modèles sur une énumération spécifique et présente une classification pour les phénomènes textuels complexes rencontrés. Le travail effectué se situe dans le domaine de la linguistique textuelle et de la modélisation de phénomènes langagiers particuliers.

1 Présentation des modèles

1.1 Le modèle de représentation de l'architecture textuelle (MAT)

Ce modèle prend source dans la notion de Mise en Forme Matérielle (Virbel, 1989) (abréviation usuelle MFM) et dans l'hypothèse de l'existence d'une équivalence fonctionnelle entre des phénomènes typographiques, dispositionnels et lexico-syntaxiques. La MFM est un sous-ensemble de propriétés morpho-dispositionnelles du texte, propriétés possédant des équivalents langagiers. Ainsi, les constituants des textes, les objets textuels, sont perceptibles par le jeu de contraste de la MFM. On trouve parmi ces objets : les théorèmes, les définitions, les énumérations, les titres, etc. L'*architecture de texte* est l'ensemble des objets textuels et les propriétés qu'ils entretiennent entre eux. Un modèle formel de représentation de l'architecture textuelle a été élaboré dans (Pascual, 1991). Ce modèle permet de représenter les différents objets textuels (à l'aide de métaphrases) et formalise les propriétés entre ces objets textuels. L'architecture d'un texte est représenté par un ensemble d'instance de métaphrases (respectant certaines propriétés de cohérences et de cohésions) et regroupées dans un métadiscours.
L'intérêt d'utiliser ce modèle, dans le cadre de cette étude est que, d'une part, il fournit un cadre d'interprétation et de délimitation des objets textuels et que, d'autre part, il autorise la prise en compte de structures de textes assez complexes (i.e. non-hiérarchiques) car il est peu contraint. Pour un plus grande lisibité, nous utiliserons ici une notation sous forme de graphe.

1.2 Théorie des Structures Rhétoriques (RST)

La RST (*Rhetorical Structure Theory* - (Mann et Thompson, 1987)), est définie comme étant une théorie descriptive et fonctionnelle du texte, mêlant des aspects sémantiques et intentionnels. Les auteurs posent une vingtaine de relations rhétoriques permettant de lier deux segments de texte adjacents entre eux, dont l'un possède le statut de noyau - segment de texte primordial pour la cohérence - et l'autre celui de satellite - segment optionnel. Ces segments peuvent être de deux types : segment minimal (*Text Unit*) ou segment composé (*Text Span*). Les auteurs définissent les unités minimales comme des unités fonctionnellement indépendantes : elles correspondent généralement aux propositions. Les auteurs de la RST affirment clairement que les relations rhétoriques sont indépendantes de tout signe spécifique : la reconnaissance d'une relation repose sur une interprétation sémantico-pragmatique du contenu du segment. Cette interprétation s'effectue suivant les opinions de l'analyste du texte : l'analyse résultante est donc une analyse subjective. Par ailleurs, des *schémas rhétoriques* décrivant l'organisation structurelle d'un texte, quelque soit le niveau hiérarchique de ce dernier, permettent de lier un noyau et un satellite, deux ou plusieurs noyaux entre eux, et un noyau avec plusieurs satellites. La structure du texte est donc définie en termes de compositions d'applications de schémas, et ce de manière réitérative. La structure rhétorique finale d'un texte est strictement hiérarchique et se présente sous la forme d'un *arbre RST*.
Dans ce travail, nous nous appuyons principalement sur la distinction Noyau/Satellite prônée par la RST et qui, selon nous, instaure une relation de dépendance entre deux segments.

1.3 Relations de nature morpho-syntaxique

Pour caractériser les phénomènes syntaxiques des énumérations, et en particulier leurs différences avec les phénomènes rhétoriques (Luc et al., 2000), nous avons défini deux nouvelles relations : les relations *paradigmatique* et *syntagmatique*. Nous utilisons ces relations dans leur utilisation linguistique habituelle : l'axe paradigmatique est défini comme étant l'axe de substitution des unités linguistiques et l'axe syntagmatique comme l'axe de succession des unités linguistiques. Ainsi, une relation de type *syntagmatique* instaure une dépendance syntaxique entre un item et un autre item tandis qu'une relation paradigmatique pose une équivalence syntaxique entre deux items au sein de l'énumération (cette relation paradigmatique est donc une relation transitive).
Nous utilisons de telles relations car il nous semble qu'intuitivement, elles représentent convenablement les phénomènes que nous voulons représenter à savoir la fonction syntaxique des items au sein de la structure énumérative qui est parfois différente de la fonction rhétorique (comme nous le verrons sur les exemples de la section suivante).

2 Les énumérations

2.1 Énumération parallèle *vs.* non-parallèle

Bien qu'il n'existe pas de définition attestée sur la bonne formation des énumération, ces objets sont le plus communément appréhendés comme un moyen de mettre en relief une forme d'identité entre des objets ou des entités qui occupent la même fonction (syntaxique ou textuelle)

dans le texte ; les procédés de mise en relief venant renforcer la fonction coordinative des éléments traités. Nous appelons ces énumérations des *énumérations parallèles*.
Cependant, l'observation des énumérations dans des textes scientifiques (Virbel, 1999b) nous a permis d'en trouver un grand nombre ne répondant pas à la vision classique (i.e. parallélisme fonction/mise lyse (rhétorique et architecturale) des énumérations du recueil. en forme). Ainsi, dans l'exemple suivant :

Le Lindy Hop est :

- une danse swing,
- dont la naissance remonte aux années 20.

les deux constituants entretiennent une relation de dépendance (syntaxique) et ne peuvent être échangés. Nous avons aussi rencontré des cas d'énumération où les constituants occupent la même fonction au sein de l'énumération mais sont visuellement différents (par exemple, deux items ne sont pas sur le même plan structural dans le texte). Par opposition au terme précédent, nous appelons ces énumérations des *énumérations non-parallèles*.

2.1.1 Terminologie

Un **item** : c'est une entité énumérée (ou plutôt coénumérée) perceptible par variation de la MFM. Un item est caractérisé par diverses marques pouvant être typographiques (tiret, numérotation, etc.), dispositionnelles (espacement vertical ou horizontal), lexico-syntaxiques (organisateurs textuels, schémas syntaxiques des items, etc.), ou toute combinaison de ces marques.
Une **énumération** : c'est un ensemble d'items (au moins 2). Ces items peuvent entretenir entre eux des relations diverses.
Une **amorce** : c'est une phrase introductrice précédant l'énumération. Cette amorce est caractérisée par une (ou des) combinaison(s) de marques lexicales (par exemple, "les suivants"), typographiques ([**:**]), dispositionnelles (saut de ligne) ou syntaxiques.
La **structure énumérative** : elle comprend une amorce, une énumération (i.e. ensemble d'items) et parfois une conclusion.

2.2 Les exemples commentés

Cette section est consacrée à la présentation d'exemples commentés d'énumérations non-parallèles. Ces énumérations sont extraites d'un recueil (Virbel, 1999b) regroupant plus de soixante-dix cas de ces énumérations non-parallèles (français et anglais). Notons que les textes d'où sont issues ces énumérations comprennent aussi d'autres énumérations qui, elles, répondent au format classique des énumérations (parallélisme fonctionnel et présentationnel).

2.2.1 Exemple de la "lecture savante"

Sur ce premier exemple (fig. 1), la première phrase constitue l'amorce de l'énumération (elle "introduit" l'énumération). Les différents items entretiennent entre eux des relations de dépendance : le deuxième dépend du premier (c'est une proposition subordonnée au premier item), le

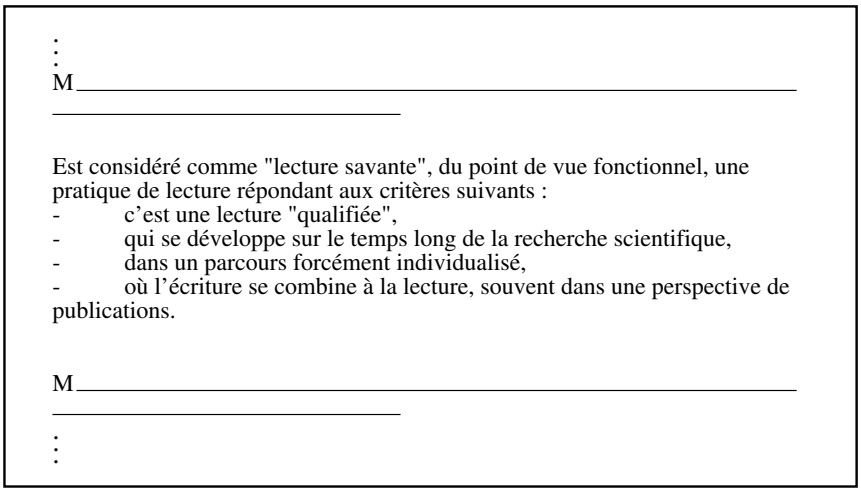

Figure 1: Exemple de la "lecture savante"

troisième dépend du second, etc. Ils ne peuvent donc pas être interchangés (à moins de modifications majeures dans la structure de l'énumération).
Cette énumération met les différents items sur le même plan d'équivalence visuelle (i.e. architecturale) alors qu'ils entretiennent des relations de dépendances les uns par rapport aux autres : les items n'occupent pas la même fonction syntaxique au sein de l'énumération.

2.2.2 Exemple du "web"

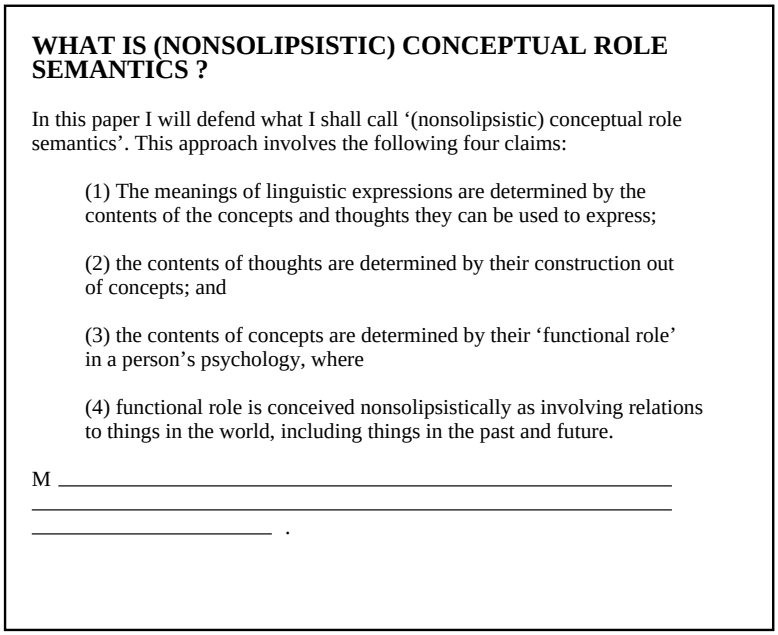

Figure 2: Exemple du "web"

D'un point de vue syntaxique, les 3 premiers items de cet exemple[1] (fig. 2) présente une structure identique. Cependant, d'un point de vue rhétorique (i.e. sémantico-pragmatique), le

[1]Cet exemple est extrait d'une page web.

premier item constitue une sorte d'introduction pour les deux items suivants (il introduit les notions de "contents of concepts" et de "contents of thoughts", reprises dans les deux items suivants). Enfin, le dernier item dépend directement du 3ème item (c'est une proposition subordonnée à celle qui constitue le 3ème item).
Ce cas est particulièrement intéressant car il met en évidence une triple structure dans l'énumération : une structure visuelle qui met au même plan tous les items (c'est la coénumérabilité), une structure rhétorique par le fait que le premier item introduit les deux suivants et enfin une structure syntaxique qui, elle, instaure une équivalence pour les trois premiers items, le dernier item dépendant du troisième item.

2.2.3 Exemple 3

Nous présentons cet exemple (fig. 3) sous forme d'image de texte[2] annotée par la délimitation et la numérotation des paragraphes (marge gauche) et des items (marge droite).

Cet exemple présente deux énumérations : une première dont les items courent sur plusieurs paragraphes (figure 3) et une seconde qui est comprise dans le dernier paragraphe. Cette deuxième énumération se compose de quatre éléments et est un résumé de la première énumération, comme ceci est clairement spécifié dans l'amorce ("...can be summarized..."). Ces deux énumérations sont donc liées : chaque item d'une énumération correspond à un item de l'autre énumération. Ces cas d'isomorphismes entre plusieurs énumérations sont plus fréquents dans le sens contraire : l'amorce d'une énumération est elle-même une énumération dont les items présentent les items de l'énumération principale.
Enfin, considérons les deux niveaux de structuration explicités dans l'image du texte : il y a d'une part une structure basée sur la segmentation en paragraphes et d'autre part, une autre structure basée sur l'énumération. Le cas le plus représentatif concerne le premier paragraphe : il se compose de l'amorce de l'énumération (repérée par l'expression "...a number of questions...") et du début du premier item (repéré par "One issue..."). Le suite du premier item correspond aux deux paragraphes suivants. Le deuxième item commence au quatrième paragraphe ("A second issue..").
Ces observations soulèvent la question du rôle des paragraphes dans les textes écrits. Heurley (Heurley, 1997) suggère de faire une distinction entre les paragraphes qu'il définit comme étant des unités visuelles, et les blocs informatifs, qui eux sont des unités structurelles ou sémantiques des textes. Ses suggestions reposent sur les résultats d'une expérimentation psycholinguistique visant à étudier la fonction du paragraphe dans les processus de lecture et d'écriture. Cette définition correspond bien à nos observations sur les interactions entre les paragraphes et les énumérations.

2.3 Représentation des structures énumératives

Il a été montré dans une étude précédente l'inadéquation des modèles pris séparément pour représenter les structures énumératives et la nécessité de recourir à une composition des modèles de référence (Luc et al., 1999). Le principe général de composition est d'attribuer le statut

[2]La notation par image de texte est inspirée des principes de la typographie aveugle (Pascual, 1991). On ne représente explicitement que les informations textuelles intéressantes pour notre propos.

Une typologie des énumérations...

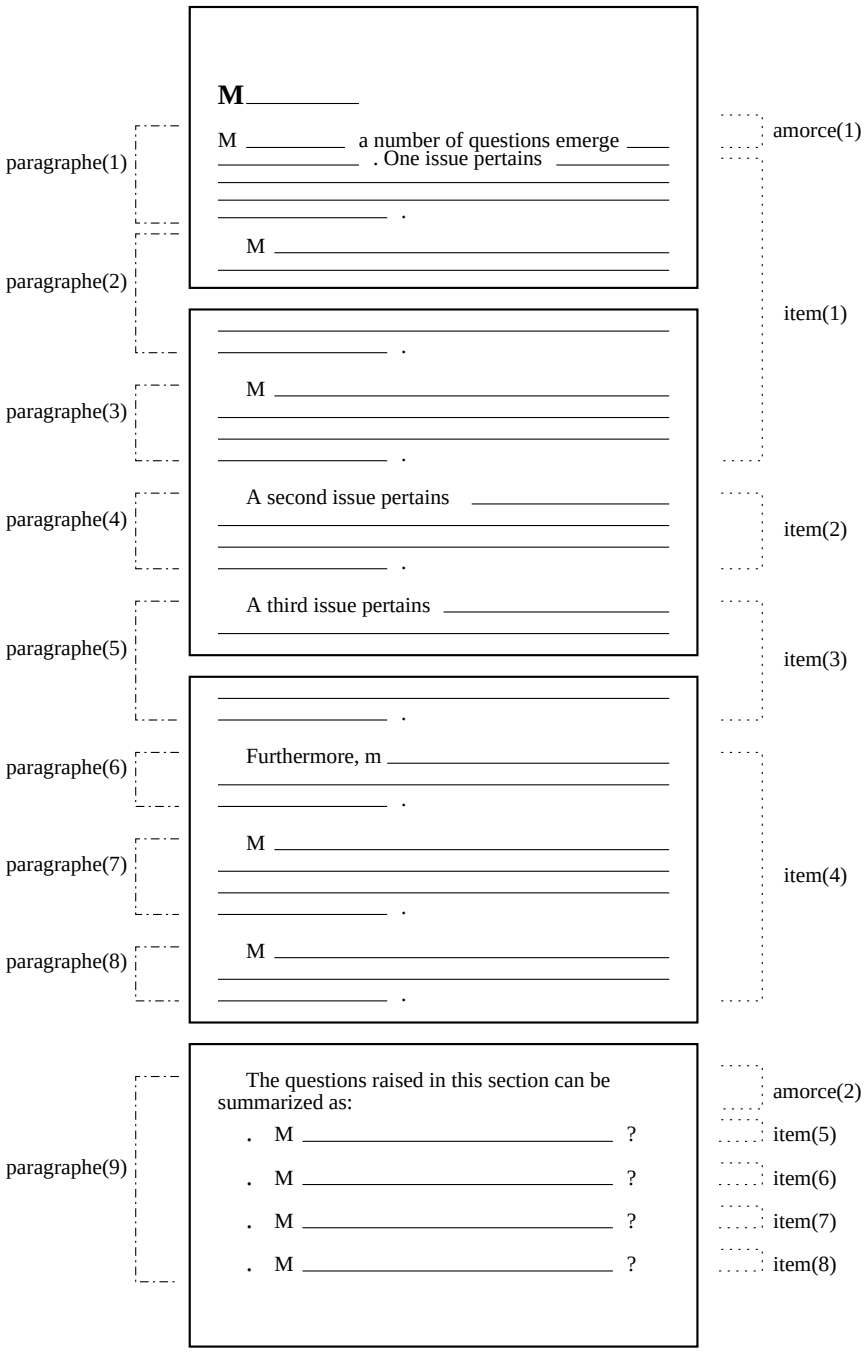

Figure 3: Exemple 3

d'objets textuels à certains segments de texte de la RST. Pour raffiner ce critère, nous avons défini deux classes d'objets textuels (Luc, 2000) :

- des **objets textuels "fonctionnels"** qui participent nécessairement à la structure rhétorique du texte. On trouve les objets suivants : titres, items, amorces, théorèmes, etc.

- des **objets textuels "structurels"** qui peuvent participer à la structure rhétorique des textes, mais pas systématiquement car il peut y avoir des problèmes de représentation liés à la structure hiérarchique de la RST. On trouve dans cette classe, les *parties* et les *paragraphes*.

Ces deux classes d'objets textuels ont été déterminées à partir des observations recueillies après l'analyse (rhétorique et architecturale) des énumérations du recueil.
Une façon optimale de déterminer une représentation mixte (RST/MAT) pour une structure de texte est : (1) de dresser la liste des objets fonctionnels, (2) de déterminer les relations rhétoriques entre ces objets, et (3) de compléter la représentation par les objets structurels et les liens architecturaux entre ces objets.

2.3.1 Exemple de composition

Nous présentons à la figure 4 une représentation mixte correspondant à l'exemple 2 (exemple du web) et contenant des informations rhétoriques, architecturales mais aussi les relations syntaxiques définies entre les items (les relations paradigmatiques et syntagmatiques).
Le structure rhétorique[3] est représentée en gras et lie les différents objets fonctionnels. Comme cela est défini par le MAT, l'objet textuel *énumération* est agencée en 4 items et liée à l'amorce. Par ailleurs, le premier paragraphe se compose de l'amorce et des 4 items.

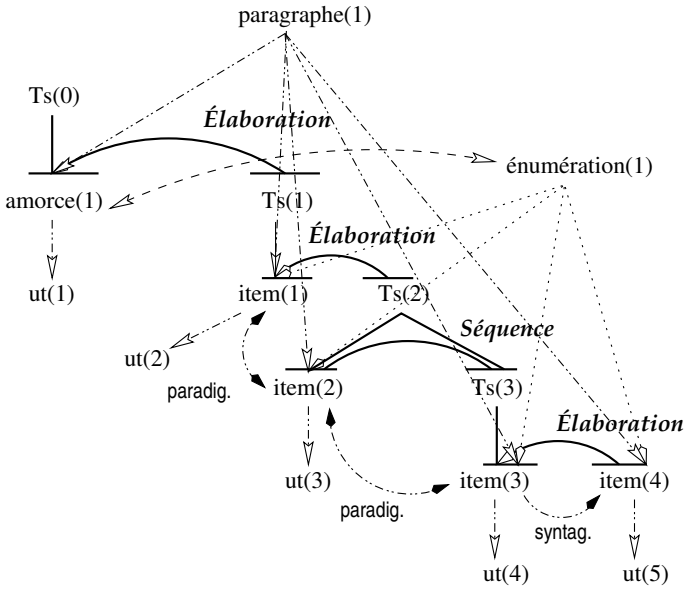

Figure 4: Graphe RST/MTA pour l'exemple 2

3 Classification des énumérations

Nous proposons ici une classification pour les énumérations, classification reposant sur les caractéristiques et propriétés des modèles de référence (RST et MAT). Nous posons 3 catégories pour les énumérations, les types d'énumérations sont exclusives à l'intérieur de chaque catégorie, mais une énumération possède 3 types appartenant à chaque catégorie.

Première catégorie

- *énumération syntagmatique* : tous les items entretiennent une relation de dépendance (syntaxique ou rhétorique) les uns par rapports aux autres.

[3] *Ts* correspond à un segment composé et *ut* à une unité minimale.

- *énumération paradigmatique* : les items sont fonctionnellement équivalents (syntaxiquement et rhétoriquement) au sein de l'énumération.
- *énumération hybride* : au moins deux items entretiennent une relation de dépendance (rhétorique ou syntaxique) et au moins deux autres items sont sur le même plan d'équivalence fonctionnelle (sémantique et syntaxique) dans l'énumération.

Deuxième catégorie

- *énumération hétérogène* : les items ne sont pas sur le même plan structural dans le texte.
- *énumération homogène* : tous les items sont sur le même plan structural dans le texte.

Troisième catégorie

- *énumération liée* : énumération dont l'amorce ou la conclusion associée contient une structure énumérative, ou bien, énumération qui est contenue dans l'amorce ou la conclusion d'une autre énumération.
- *énumération isolée* : énumération dont l'amorce ou la conclusion associée ne contient pas de structure énumérative, ou bien, énumération qui n'est pas contenue dans l'amorce ou la conclusion d'une autre énumération.

L'exemple de la "lecture savante" propose une énumération syntagmatique, homogène et isolée. L'exemple du web présente une énumération hybride, homogène et isolée. Enfin, la première énumération de l'exemple 3 est paradigmatique, hétérogène et liée et la seconde énumération est paradigmatique, homogène et liée.

À l'aide de cette typologie, nous posons les définitions suivantes :

- une *énumération parallèle* est une énumération paradigmatique, homogène et isolée.
- une *énumération non-parallèle* est une énumération qui n'est pas paradigmatique, ou qui est hétérogène ou liée.

Cette classification est complète vis-à-vis des objectifs initiaux, i.e. caractérisation à l'aide des modèles de référence (Luc, 2000). Il nous reste cependant à élucider d'autres types de phénomènes remarquables sur les énumérations non-parallèles comme, par exemple, les relations existantes entre l'amorce et l'énumération. Pour ce faire, il nous faudra considérer de nouveaux modèles car ceux actuellement utilisés semblent insuffisants.

Conclusion

Dans cette communication, nous nous sommes intéressé au cas de l'énumération, et spécialement aux énumérations "non-parallèles". Nous avons mis en évidence certains phénomènes textuels remarquables présents sur ces objets et nous avons montré la nécessité (1) de considérer plusieurs types d'indices, à savoir les indices visuels, syntaxiques et rhétoriques, (2) de

faire coopérer deux modèles qui sont la RST et le modèle de représentation de l'architecture textuelle. Nous avons proposé une représentation et une classification pour ces énumérations utilisant de manière conjointe des structures architecturales et rhétoriques ainsi que des relations de nature syntaxique.
Ce travail se poursuit actuellement dans deux directions. D'une part nous sommes engagés dans une collaboration avec des psycholinguistes pour tester l'impact des types d'énumérations proposées ici sur divers processus cognitifs (compréhension, mémorisation, etc.) et sur divers types de population. D'autre part, nous cherchons à valider et tester la fréquence des types d'énumérations sur un corpus "standard" composé de textes extraits du web (Bouraoui, 2000).
L'étude présentée ici suit les principes généraux de notre recherche à savoir l'examen approfondi des objets textuels pris séparément. Une première étude sur la définition (Pascual et Péry-Woodley, 1997) avait montré la compatibilité des modèles de la RST et du MAT. Le travail présenté ici va plus loin en donnant un principe général de composition des modèles.
Ce type de travail doit se prolonger sur d'autres types d'objets textuels complexes (comme les titres, les notes de bas de page, etc.) et en prenant probablement en compte de nouveaux modèles de texte. C'est le prix à payer pour obtenir une description la plus exhaustive possible des structures visuelles et discursives du texte.

Références

Bouraoui, J.-L. (2000). Les structures énumératives : caractérisation linguistique et reconnaissance automatique. Rapport de DEA, Université de Toulouse de Mirail.

Heurley, L. (1997). Processing units in written texts: Paragraphs or information blocks ? Dans Costerman, J. et Fayol, M., éditeurs, *Processing Interclausal Relationships*, Studies in Production and Comprehension of Text, pages 179–200. Lawrence Erlbaum Associates.

Luc, C. (2000). *Représentation et composition des structures rhétoriques et visuelles du texte. Approche pour la génération de textes formatés*. Thèse de doctorat, Université Paul Sabatier.

Luc, C., Garcia-Debanc, C., Mojahid, M., Péry-Woodley, M.-P. et Virbel, J. (1999). A linguistic approach to some parameters of layout: A study of enumerations. Rapport technique, AAAI, North Falmouth, Massachusets. From the Fall Symposium "Using Layout for the Generation, Understanding or Retrieval of Documents".

Luc, C., Mojahid, M., Péry-Woodley, M.-P. et Virbel, J. (2000). Les énumérations : structures visuelles, syntaxiques et rhétoriques. Dans *Actes du Colloque International sur le Document Électronique (CIDE'2000)*. Lyon.

Mann, W. C. et Thompson, S. A. (1987). Rhetorical Structure Theory: A theory of Text Organization. Rapport technique, ISI-RS-87-190, Information Sciences Institute, Marina Del Rey, Ca.

Pascual, E. (1991). *Représentation de l'Architecture Textuelle et Génération de Texte*. Thèse de doctorat, Université Paul Sabatier.

Pascual, E. et Péry-Woodley, M.-P. (1997). Définition et actions dans les textes procéduraux. Dans Pascual, E., Nespoulous, J.-L. et Virbel, J., éditeurs, *Le texte procédural : langage, action et cognition*, pages 223–248, Mons, Gers. Prescot.

Virbel, J. (1989). The contribution of linguistic knowledge to the interpretation of text structure. Dans André, J., Quint, V. et Furuta, R., éditeurs, *Structured Documents*, pages 161–181. Cambridge University Press.

Virbel, J. (1999b). Structures textuelles - planches. fascicule 1 : Énumérations. Rapport technique, IRIT. Version 1.