Un Indice de Structuration de Texte Combinant Finesse et Disponibilité au Niveau Global et Local

Nicolas Hernandez

Un Indice de Structuration de Texte Combinant Finesse et Disponibilité au Niveau Global et Local

Nicolas Hernandez
LIMSI-CNRS LIR, Université de Paris-Sud, 91403 ORSAY, France
Nicolas.Hernandez@limsi.fr

1 Introduction

De fait, l'outil informatique offre la possibilité d'automatiser de nombreuses tâches manuelles fastidieuses ainsi que la capacité de traiter des grandes masses d'information. Néanmoins, le passage d'un support papier à un support numérique a de nombreuses conséquences sur la visualisation des documents et nos manières de les appréhender. D'une part bon nombre d'informations résultant des propriétés physiques du document ne sont plus accessibles (structure globale, genre, etc.), et d'autre part les moyens d'accès (manipulations du document) sont contraints par les caractéristiques matérielles et logicielles des interfaces intermédiaires.
Les travaux que nous menons sont directement dirigés vers la résolution de ces problèmes. Ils visent à enrichir les documents numériques avec des informations descriptives sur leur structure et contenu afin de soutenir de nouveaux moyens de visualisation et de navigation plus en adéquation avec les dispositifs d'interaction actuels.
Dans ce papier, nous focalisons notre attention sur les marques discursives des textes de genre scientifique afin de mettre en place une détection automatique d'une structure hiérarchique de ces textes. Posé que les relations inter-énoncés[1] d'une structure hiérarchique sont soit de type *subordination* (relation de dépendance entre énoncés), soit de type *coordination* (relation d'égalité entre énoncés), l'enjeu que nous soulevons est d'identifier les marques qui soutiennent tel ou tel type de relation.
Les méthodes de structuration existantes (Marcu, 1997; Choi, 2002) exploitent essentiellement des indices locaux (connecteurs et/ou relations lexicales entre énoncés adjacents). Les méthodes de segmentation fondées sur la cohésion lexicale portent des informations de structuration globale, mais (Hernandez *et al.*, 2004) montrent que ces méthodes entraînent des questions sur la granularité des segments considérés ainsi que des problèmes de cohésion abusive (e.g. une annonce d'énumération et son premier item) ou restrictive (e.g. un énoncé et son exemple lexicalement différent).
C'est dans ce cadre que nous rapportons les résultats d'une analyse préliminaire de structures décrivant des propriétés de *symétrie* linguistique et de formatage visuel au sein de ses énoncés constituants. Nous avançons que ce phénomène peut être considéré comme un des principes majeurs de la structuration des textes, qu'il est présent à tous les niveaux d'intégration (du local au global) et de granularité fine. Cette approche étend certains aspects des travaux de (Luc *et al.*, 1999) et de (Jackiewicz, 2002) sur les énumérations.

[1] Nous gardons volontairement une notion ambiguë d'énoncé ; nous l'utilisons lorsque son propos ne concerne pas une unité textuelle particulière. Exemples : syntagme, proposition, phrase, paragraphe, segment, etc.

[. . .] The three structures can be informally defined as follows:
Ideational Structure (propositional meaning conveyed by the discourse) - Two discourse units are ideationally related if their utterances in the given context entails the speaker's commitment to the existence of the relation in the world described by the discourse. Examples: cause, contrast, temporal, and so forth.
Rhetorical Structure (hierarchy of intentions in the discourse) - Two discourse units are rhetorically related if the illocutionary force of one unit is subserviant to that of the other. Examples: justification, motivation, evidence, and so forth.
Sequential Structure (coordination and subordination of discourse segments) - The sequential structure describes paratactic or hypotactic relations between adjacent discourse segments that are ideationally and rhetorically only loosely or indirectly related. A paratactic sequential relation is a transition between issues or topics that either follow a preplanned list or is locally occasioned, as for instance in conversation. Hypotactic sequential relations are those leading into or out of, for instance, a commentary, correction, paraphrase, digression, or interruption segment.
Usually one of the three components is more salient than others for anchoring an utterance in its context.[. . .]

Figure 1: Exemple de symétrie extrait de (Redeker Rapport Technique'00)

2 La *Symétrie*, observation du phénomène en corpus

Nos observations ont été menées sur un corpus de 42 structures "symétriques" relevées tout au long de nos lectures bibliographiques (en anglais et en français). Notre critère de sélection reposait sur la rencontre d'une nouvelle forme ou d'un aspect encore inconnu.

Intuitivement, nous qualifions d'énoncés symétriques les énoncés qui présentent des similarités linguistiques et/ou typo-dispositionnelles indiquant l'intention de l'auteur de considérer ces énoncés sur un même plan. Ces similarités peuvent aussi bien être formelles que de surface, et ne concerner que tout ou partie des énoncés.

On trouve ce phénomène à différents niveaux de structuration des textes : de la structure logique en titres à l'énumération d'une suite de syntagmes nominaux dans une phrase, en passant par les découpages en paragraphes, les phénomènes d'ellipses verbales ou autres, les énumérations avec organisateurs textuels (e.g. premièrement, deuxièmement, etc.), celles avec marqueurs typo-dispositionnels (ponctuation et marques de mise en forme et en page), ou celles combinant les deux ou ne possédant aucun.
Les énumérations sont les structures qui exemplifient le plus clairement ce phénomène. (Luc *et al.*, 1999; Luc, 2000; Jackiewicz, 2002) les ont étudiées. Principalement ils ont cherché à décrire les marques récurrentes de l'objet textuel "énumération", et ce d'une part en identifiant et typant ses constituants, et d'autre part en relevant les indices typo-dispositionnels et lexicaux permettant de les repérer. Leur attention s'est ainsi essentiellement portée sur les énumérations ayant une mise en forme matérielle explicite. Notre travail complète leurs analyses en s'intéressant particulièrement à l'étude des traits symétriques des énoncés-items et ceux de leurs constituants. Bien que syntaxiquement distincts deux énoncés peuvent comporter des syntagmes internes symétriques qui permettent de les considérer comme deux items d'une même structure énumérative.

Les figures 1 et 2 sont différents exemples de structures symétriques extraites de diverses parties de document. Ils ont été sélectionnés pour illustrer le caractère d'équivalence et de com-

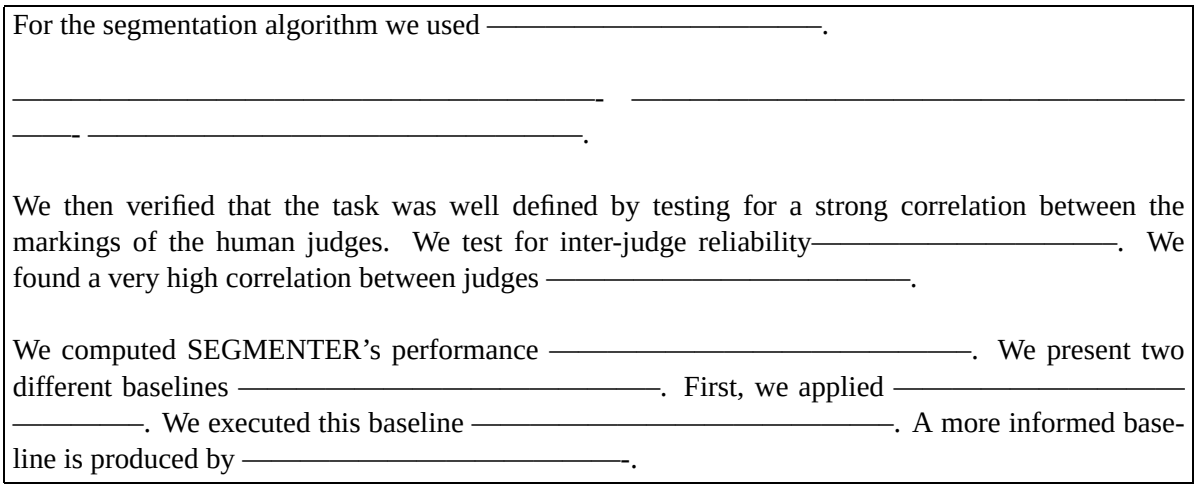

Figure 2: Exemple de symétrie extrait de (Kan et al. WVLC'98)

plémentarité fonctionnelle entre formatage visuel et formulation langagière, relevé par (Virbel, 1989). En l'absence de formatage typo-dispositionnel, les expressions linguistiques héritent de traits symétriques afin de marquer la relation de coordination entre les énoncés. Le schéma de symétrie des deux premiers items de la figure 1 correspond à :

> [*ADJECTIF+al Structure (X the discourse) - Two discourse units are ADVERB+ly related if Y . Examples: NOM, NOM, NOM, and so forth.*]

Ce schéma montre entre autre que la symétrie peut concerner plusieurs constituants de l'item et que les symétries peuvent être de nature typo-dispositionnelle, lexicale, grammaticale. Le troisième item peut être rapproché des deux premiers par un patron plus souple. Nous remarquons qu'il contient lui-même une sous-structuration marquée par des éléments symétriques.
Dans cet exemple, la symétrie traduit une structuration globale signalant une relation de type coordination entre les différents items.

La structure décrite à la figure 2 se compose de trois items, un pour chaque paragraphe. Le schéma de symétrie correspond à :

> [*we X VERBE+prétérit*]

Bien que beaucoup plus simple que le précédent, il est difficile de repérer les items de la structure globale de ceux des niveaux de structuration inférieurs du fait de leur symétrie apparente (dans certains cas on constate [*we VERBE+présent*]). L'information typo-dispositionnelle (début de paragraphe) permet d'identifier la symétrie qui est au niveau le plus global et ainsi de reposer le problème d'identification des items symétriques au niveau inférieur.
Cet extrait illustre le problème majeur des structures symétriques, à savoir le cas où la structure symétrique présente un lien anaphorique. Dans ce genre de situation, on peut se demander si la symétrie souligne un nouvel aspect d'une entité thématique (type coordination) ou bien si elle décrit la présence d'une élaboration (type subordination) ? Dans notre corpus, l'utilisation d'une heuristique fondée sur la cohésion lexicale ou bien le suivi rhème vers thème permet d'identifier si deux items sont une élaboration ou pas.
En l'absence de liens anaphoriques insidieux, nos observations nous amènent à penser que la symétrie rend compte d'une relation de type coordination.

3 Synthèse

Ce phénomène de symétrie porte un certain nombre de propriétés (liste non exhaustive) :

- comparativement aux connecteurs et à l'information lexicale, il constitue un indice de structuration à la fois fin (i.e. repérable avec précision) et observable à différents niveaux de granularité contextuelle (i.e. du niveau local au niveau global). La relation discursive portée par la symétrie est préférentiellement de type coordination ;
- il concerne aussi bien des traits spécifiques de l'énoncé-item pris en tant que tel, que des traits de ses constituants. Suivant l'unité textuelle considérée comme item, ses constituants peuvent être des phrases, des propositions, des syntagmes ou bien des tokens[2]. Le phénomène ne touche pas seulement les marques de signalisation explicite d'items (e.g. d'une part, d'autre part, etc.), il touche également les éléments du propos des énoncés ;
- les traits symétriques sont de nature typo-dispositionnelle, lexicale, grammaticale (fonction sujet, verbe, etc. ; catégorie nom, déterminant, etc. ; genre, nombre ; temps verbal ; etc.) et/ou sémantique (avec certaines classes privilégiées, e.g. les organisateurs). Les relations de symétrie s'observent suivant l'axe paradigmatique (axe de substitution des unités linguistiques) et l'axe syntagmatiques (axe de succession) ;
- il permet d'apporter des informations descriptives sur le contenu de la structure symétrique par distinction entre le matériel linguistique transversal et celui spécifique à chaque item.

Ce travail est novateur car jusqu'à présent aucune technique de structuration n'exploite cette caractéristique discursive. Dans la perspective d'une détection automatique de structures de texte, nos travaux à venir consisteront d'une part à apporter un soutien quantitatif à notre analyse qualitative, et d'autre part à concevoir une méthode pour le repérage de symétries entre énoncés.

Références

CHOI F. Y. Y. (2002). *Content-based Text Navigation*. PhD thesis, Department of Computer Science, University of Manchester.

HERNANDEZ N., VIGIER D., CHAROLLES M. & DESCLES J.-P. (2004). Text organisation by combining fine-grained linguistic markers with global statiscal measures. In *Document Design*, Netherlands.

JACKIEWICZ A. (2002). Repérage et délimitation des cadres organisationnels pour la segmentation automatique des textes. In *CIFT'02*, Hammamet, Tunisie.

LUC C. (2000). *Représentation et composition des structures visuelles et rhétoriques du texte. Approche pour la génération de textes formatés*. PhD thesis, Université Paul Sabatier – Toulouse III.

LUC C., MOJAHID M., VIRBEL J., GARCIA-DEBANC C. & PÉRY-WOODLEY M.-P. (1999). A linguistic approach to some parameters of layout: A study of enumerations. In *AAAI*, Massachusets.

MARCU D. (1997). *The Rhetorical Parsing, Summarization, and Generation of Natural Language Texts*. PhD thesis, Dept. of Computer Science, University of Toronto.

VIRBEL J. (1989). The contribution of linguistic knowledge to the interpretation of text structure. In J. ANDRÉ, V. QUINT & R. FURUTA, Eds., *Structured Documents*, p. 161–181. Cambridge University.

[2]Les symboles et les caractères de ponctuation sont considérés au même titre que les mots.