ANNODIS : une ressource pour l'identification de systèmes de marqueurs du discours

Lydia-Mai Ho-Dac, Marie-Paule Péry-Woodley

ANNODIS : une ressource pour l'identification de systèmes de marqueurs du discours

Lydia-Mai Ho-Dac et Marie-Paule Péry-Woodley
CLLE-ERSS, CNRS et Université de Toulouse (UTM)

A la recherche des "marqueurs" impliqués dans la signalisation de l'organisation discursive, et des interactions ou jeux de contraintes entre différents systèmes de marqueurs, de nombreux travaux visent à définir des combinaisons ou faisceaux d'indices discursifs. L'étude que nous présentons s'inscrit dans cette lignée, mais de manière descriptive et empirique à travers l'application de techniques de fouille à un corpus annoté manuellement. Nous décrivons brièvement ce corpus, puis la méthode qui nous permet de passer d'abord des traits (pré-marqués automatiquement) aux indices (annotés manuellement), puis des indices aux combinaisons que nous appelons "cuesets".

1 Présentation de la ressource

Notre corpus de travail, issu de la ressource ANNODIS ([3], `redac.univ-tlse2.fr/corpus/annodis.html`), est composé d'articles à visée informative représentant une certaine diversité en terme de genre textuel, de type dominant et de structure de document [4]. Nous avons en effet cherché à intégrer d'entrée de jeu l'hypothèse de la variation dans les réalisations discursives en fonction de variations extra-linguistiques. Trois sources ont été sélectionnées afin de constituer trois sous-corpus : **WIKI** – articles encyclopédiques longs (type expositif) publiés sur l'encyclopédie en ligne Wikipedia (version datant de l'été 2009) ; **LING** – articles scientifiques provenant du 1er Congrès Mondial de Linguistique Française (CMLF-08) ; **GEOP** – articles et rapports (type expositif/argumentatif) publiés par l'Institut Français des Relations Internationales. Ces trois sous-corpus présentent de fortes variations de structuration visuelle, les articles encyclopédiques étant beaucoup plus structurés visuellement que les autres : plus de (niveaux de) titres et de listes formattées, des sections et paragraphes plus courts, etc.
Ces textes ont été annotés manuellement selon un modèle définissant deux structures multi-échelle : les structures énumératives et les chaînes topicales. L' annotation a impliqué à la fois la délimitation de segments (d'au moins deux phrases et pouvant aller jusqu'à plusieurs sections) et l'indication des traits de surface signalant ces structures, les traits annotés devenant alors des **indices** de ces structures. Une structure énumérative (**SE** en abrégé) est un segment de texte défini par une structuration interne mettant en jeu les sous-segments suivants : une **amorce**, segment optionnel qui introduit l'énumération ; plusieurs **items** qui constituent le noyau de l'énumération (au moins deux items doivent être identifiés pour qu'une structure soit annotée) ; et une **clôture** segment optionnel qui résume ou clôt l'énumération.

corpus	Nbmots	Nbtextes	SE	amorce	item	clôture	CT
WIK2	218756	28	401	297	1651	36	266
LING	169310	25	297	228	848	46	88
GEOP	226877	30	293	209	863	49	234
total	614943	83	991	734	3362	131	588

Table 1 : Répartitions des structures et segments annotés dans la ressource ANNODIS

Les chaînes topicales (**CT** en abrégé), quant à elles, sont définies comme des segments de texte regroupant des phrases reliées par un référent commun. Ce référent doit être exprimé en position pré-verbale (i.e. potentiellement topicale) dans plusieurs des phrases du segment.
Au total, 83 textes entiers ont été annotés en structures multi-échelle. La répartition par sous-corpus du nombre de segments annotés est donnée dans le tableau 1.

2 Des traits aux indices

Cette étude s'intéresse particulièrement à l'ensemble de traits considérés par les annotateurs comme des indices des différents segments annotés (structure énumérative, amorce, item, clôture, et chaîne topicale). L'association entre un indice et un segment a été réalisée soit par validation de traits prémarqués soit par identification de nouveaux traits dont la nature était alors renseignée. Selon le segment associé et la nature de l'indice, différentes catégories d'indices se dégagent. Au niveau des chaînes topicales, 3298 indices ont été annotés dont la plupart sont des descriptions co-référentielles (2266) ou des pronoms de 3e personne (959). Parmi les descriptions co-référentielles, nous pouvons distinguer une majorité de descriptions définies et noms propres répétés (1990) des autres types de descriptions (151 descriptions possessives, 125 descriptions démonstratives, 74 SN indéfinis).
Du côté des structures énumératives, 4543 indices ont été annotés : 817 dans des amorces, 101 dans des clôtures et 3625 dans des items. Ces indices sont beaucoup plus diversifiés que pour les chaînes topicales, notamment au niveau de la signalisation des items. La table 2 donne les effectifs des 7 principaux types d'indices d'items.

1047	puces et numérotations
639	adverbiaux circonstanciels de temps (428), lieu (97) et ce qui a été annoté comme des adverbiaux "notionnels" (114)
616	séquenceurs (388) et connecteurs (228)
592	parallélismes syntaxiques ou sémantiques
383	titres de section
246	ponctuations : virgules, points-virgules et coordinations
102	autres : reprise d'annonce (39), "entité nommée" (11), expressions co-référentielle (13), apposition (32), autres circonstants (7)

Table 2 : Effectifs des différents types d'indices annotés dans les items

3 Des indices aux cuesets

Dans un premier temps, l'analyse de la composition et de la signalisation des structures énumératives a conduit à une première typologie en lien avec la structure de document ([1]) : SE dont les items sont des sections titrées ; listes formatées ; SE s'étendant sur plusieurs paragraphes ; et finalement SE intra-paragraphiques. Pour aborder à présent l'interaction des indices dans ces structures, nous avons proposé la notion de "cueset": le cueset d'une SE est l'ensemble des types d'indices qui lui sont associés. Une SE dont les items sont signalés par un parallélisme syntaxique et des marqueurs d'intégration linéaire aura ainsi le cueset (Paral. + Seq.). Les cuesets entretiennent entre eux des relations d'inclusion, tout indice dans un cueset donné appartenant également au cueset plus spécifique. Cette approche nous permet de caractériser les énumérations (liste d'items) des quatre types de SE. Les résultats obtenus sont particulièrement intéressants en ce qui concerne les associations (attraction et répulsion) entre marqueurs d'intégration linéaire, adverbiaux cadratifs, et parallélismes syntaxiques. D'autres observations concernent l'association entre cueset (de l'énumération) et la présence/absence d'amorce. Cette approche empirique à grande échelle nous semble présenter un intérêt à la fois pour l'étude linguistique de la signalisation de l'organisation textuelle, et pour le développement d'applications du TAL.

References

[1] Ho-Dac L.-M., Péry-Woodley M.-P. & Tanguy L. Anatomie des structures énumératives. *TALN 2010*, ATALA, Université de Montréal, Montréal, Juillet 2010.

[2] Ho-Dac L.-M., Fabre C., Péry-Woodley M.-P.-P., Rebeyrolle J. & Tanguy L. An empirical approach to the signalling of enumerative structures. *Discours*, 10, à paraître 2012.

[3] Péry-Woodley M.-P., Afantenos S.D., Ho-Dac L.-M. & Asher N. La ressource ANNODIS, un corpus enrichi d'annotations discursives. *TAL*, 52(3), à paraître 2012.

[4] Power R., Scott D., Bouayad-Agha N. Document Structure. *Computational Linguistics*, 29(2), p. 211-260, 2003.