Annotation des relations causales dans un corpus de textes d'élèves d'école et collège

Myriam Bras, Maëlle Joret, Audrey Pépin-Boutin, Laure Vieu

Bras Myriam*, Joret Maëlle*, Pépin-Boutin Audrey*, Vieu Laure°
Université de Toulouse
* CLLE, UMR 5263, CNRS et Université Toulouse Jean Jaurès
° IRIT, UMR 5505, CNRS et Université Paul Sabatier

Annotation des relations causales dans un corpus de textes d'élèves d'école et collège

Nous proposons dans cette communication d'aborder la question de l'acquisition des relations causales à travers l'analyse de productions écrites d'élèves d'école et collège. Il s'agit d'élèves francophones dont le français est la langue de scolarisation. Le corpus analysé est issu du corpus RESOLCO constitué de textes d'élèves d'école primaire et de collège produits selon une même consigne d'écriture, une tâche-problème demandant aux élèves la production d'un texte narratif impliquant la résolution d'anaphores de divers types et imposant la présence de relations causales (Garcia-Debanc et Bonnemaison, 2014 ; Garcia-Debanc et al, 2017). Nous y avons sélectionné trois niveaux correspondant aux fins des cycles 2, 3 et 4 – CE2, 6ème et 3ème – afin d'observer d'éventuels paliers d'évolution. Ce corpus est en cours d'annotation dans l'objectif d'analyser la cohérence discursive à la réception de textes de scripteurs dont la compétence rédactionnelle est encore en cours d'acquisition, à partir de l'identification de relations de discours entre segments.

Nous procédons d'abord à une segmentation en Unités de Discours Elémentaires (UDE), établie sur la base de critères syntaxiques et sémantico-référentiels. L'annotation en Relations de Discours (RD) consiste ensuite à relier les UDE entre elles par des RD. Le jeu de relations choisi est proche de celui de la Segmented Discourse Representation Theory (Asher et Lascarides 2003), comme dans le projet ANNODIS (Afantenos 2012). La SDRT définit de façon formelle ce qu'est un discours cohérent et offre une méthode opératoire de construction de représentations du discours articulées par des relations rhétoriques. Parmi ces relations, certaines sont spécifiquement causales (Atallah 2014) : elles expriment les liens de causalité qui peuvent être interprétés à partir des connecteurs causaux ou inférés grâce à des informations diverses incluant la sémantique lexicale, la sémantique des temps verbaux et des connaissances extralinguistiques partagées par les locuteurs. Cette théorie est mise ici à l'épreuve pour la première fois sur des textes d'apprenants. Il s'agit de mettre au jour les données et les mécanismes à l'œuvre dans l'interprétation de discours dont la cohérence, en tant que propriété de la réception des discours (Charolles 1995), varie d'un scripteur à l'autre, et évolue tout au long de la scolarité. La construction du corpus annoté permet à la fois de tester la capacité explicative de la théorie face à divers types d'incohérence, pour lesquels un nouveau jeu de marques d'annotation est proposé, et à une structure globale des textes plus ou moins complexe et d'évaluer des hypothèses sur l'évolution de la cohérence entre la fin du cycle 2 et la fin du cycle 4.

Le corpus est en cours d'annotation. Nous prévoyons d'avoir annoté 24 textes pour la conférence. Nous donnons ici les tendances à mi-parcours de l'annotation.

Sur les 12 textes annotés, les RD majoritaires sont les relations de Narration et de Continuation avec plus de 40% au total. La proportion de RD causales s'élève à 15% environ (54 relations causales, une fois décomptées celles induites par les phrases imposées). Au sein des relations causales, on observe 8 fois plus de relations de Résultat que d'Explication en CE2, deux fois

plus en 6ème, et la tendance s'inverse en 3ème avec 1,3 fois plus de relations d'Explication que de Résultat.
Au-delà des évolutions dans les relations causales, on observe une complexification de la structure, ce que l'on peut apprécier par le nombre de segments complexes révélé par les relations coordonnantes de continuation : 24% en CE2, 45% en 6ème, 40% en 3ème.

Nous cherchons aussi à observer l'évolution des moyens utilisés par les élèves pour exprimer les relations causales, en particulier en répertoriant les RD causales explicites, i.e. marquées par un connecteur, et les RD causales implicites, i.e. inférées à partir d'autres sources d'informations. Avec seulement 54 relations causales dans 12 textes, et le fait que les relations peuvent être explicitées par divers connecteurs, les nombres d'occurrences par connecteur sont encore trop faibles pour dégager une tendance claire, et vérifier si la tendance au marquage des relations causales diminue ou augmente en fonction de l'âge, et donc si les relations causales sont comparables aux relations temporelles dont le marquage diminue, ou bien si elles bénéficient de l'impact de l'augmentation des connecteurs non-temporels dont la fréquence augmente avec l'âge (voir inter alia Schneuwly et al. 1989).

Références bibliographiques

Afantenos, S. ; Asher, N. ; Benamara, F. ; Bras, M. ; Fabre, C. ; Ho-Dac, M. ; Le Draoulec, A. ; Muller, P. ; Péry-Woodley, M.-P. ; Prévot, L. ; Rebeyrolle, J. ; Tanguy, L. ; Vergez-Couret, M. ; Vieu, L. (2012). « An empirical resource for discovering cognitive principles of discourse organisation: the ANNODIS corpus », In Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC)., 23-25 mai 2012, Istanbul, Turquie.

Asher, N., & Lascarides, A. (2003). *Logics of Conversation*. Cambridge : Cambridge University Press.

Atallah, C. (2014). *Analyse de Relations de Discours causales en corpus étude empirique et caractérisation théorique*. Thèse de doctorat, Université de Toulouse.

Charolles, M. (1995). « Cohésion, Cohérence et pertinence du discours », *Travaux de Linguistique,* 29 : 125-151.

Garcia-Debanc, C., Bonnemaison, K. (2014). « La gestion de la cohésion textuelle par des élèves de 11-12 ans : réussites et difficultés », Actes du 4e *Congrès Mondial de Linguistique Française (CMLF 2014),* Juillet 2014, Berlin, Allemagne.

Garcia-Debanc C., Ho-Dac, M., Bras, M., Rebeyrolle, J. (2017) « Vers l'annotation discursive de textes d'élèves », *Corpus* [En ligne], 16 | 2017.

Schneuwly B., Rosat M.C., Dolz J. (1989). « Les organisateurs textuels dans quatre types de textes écrits. Etude chez des élèves de 10, 12 et 14 ans », *Langue française*, 81 : 40-58.