ANNODIS: une approche outillée de l'annotation de structures discursives

Marie-Paule Péry-Woodley (1), Nicholas Asher (2), Patrice Enjalbert (3), Farah Benamara(2), Myriam Bras (1), Cécile Fabre (1), Stéphane Ferrari (3), Lydia-Mai Ho-Dac (1), Anne Le Draoulec (1), Yann Mathet (3), Philippe Muller (2), Laurent Prévot (4), Josette Rebeyrolle (1), Ludovic Tanguy (1), Marianne Vergez-Couret (1), Laure Vieu (2), Antoine Widlöcher (3)

(1) CLLE-ERSS – Université de Toulouse UTM
{pery,bras,cecile.fabre,hodac,draoulec,rebeyrol,vergez}@univ-tlse2.fr
(2) IRIT – Université de Toulouse UPS
{asher,benamara,philippe.muller,vieu}@irit.fr
(3) GREYC – Université de Caen
{patrice.enjalbert,stephane.ferrari,mathet,awidloch}@info.unicaen.fr
(4) Laboratoire Parole et Langage – Université de Provence
laurent.prevot@lpl-aix.fr

Résumé.

Le projet ANNODIS vise la construction d'un corpus de textes annotés au niveau discursif ainsi que le développement d'outils pour l'annotation et l'exploitation de corpus. Les annotations adoptent deux points de vue complémentaires : une perspective ascendante part d'unités de discours minimales pour construire des structures complexes *via* un jeu de relations de discours ; une perspective descendante aborde le texte dans son entier et se base sur des indices pré-identifiés pour détecter des structures discursives de haut niveau. La construction du corpus est associée à la création de deux interfaces : la première assiste l'annotation manuelle des relations et structures discursives en permettant une visualisation du marquage issu des prétraitements ; une seconde sera destinée à l'exploitation des annotations. Nous présentons les modèles et protocoles d'annotation élaborés pour mettre en œuvre, au travers de l'interface dédiée, la campagne d'annotation.

Abstract

The ANNODIS project has two interconnected objectives: to produce a corpus of texts annotated at discourse-level, and to develop tools for corpus annotation and exploitation. Two sets of annotations are proposed, representing two complementary perspectives on discourse organisation: a bottom-up approach starting from minimal discourse units and building complex structures *via* a set of discourse relations; a top-down approach envisaging the text as a whole and using pre-identified cues to detect discourse macro-structures. The construction of the corpus goes hand in hand with the development of two interfaces: the first one supports manual annotation of discourse structures, and allows different views of the texts using NLP-

based pre-processing; another interface will support the exploitation of the annotations. We present the discourse models and annotation protocols, and the interface which embodies them.

Mots-clés :

annotation de corpus, structures de discours, interface d'annotation

Keywords:

corpus annotation, discourse structures, annotation tools

1. ANNODIS : annotation discursive de corpus

Le projet ANNODIS[1] vise la constitution d'un corpus de français écrit enrichi d'annotations concernant le niveau discursif. Tout en se situant dans le sillage de projets d'annotation de relations et structures de discours existants pour l'anglais – e.g. Penn Discourse Treebank (Prasad et al., 2006) – il s'en démarque sur plusieurs points : sa principale originalité est d'aborder l'organisation discursive à partir de deux approches complémentaires – ascendante et descendante (macro-structures) ; deuxièmement, les annotations s'appliquent à un corpus diversifié pour permettre la prise en compte de réalisations discursives variées ; ce corpus fait l'objet de prétraitements automatiques pour guider les annotations ; enfin, le développement d'outils d'aide à l'annotation et à la navigation constitue un objectif majeur du projet.

Les sections 2 et 3 présentent les deux points de vue théoriques et les choix méthodologiques sur lesquels se fondent les annotations ascendantes et descendantes respectivement. La section 4 est consacrée à l'interface, au cœur du dispositif puisqu'elle donne corps aux deux modèles d'annotation, et permettra leur mise en relation.

2. Annoter des relations de discours : l'approche ascendante

Le point de vue « ascendant » s'inscrit dans le cadre des théories du discours qui s'attachent à construire une structure complète d'un discours, essentiellement vu comme un ensemble d'unités élémentaires reliées par des relations de cohérence (dites rhétoriques). Cette vision théorique du discours rassemble principalement les travaux de la RST (Mann et Thompson, 1987), la LDM (Polanyi et al., 2004 ; Wolf et Gibson, 2005), DLTAG (Forbes et al, 2003), le PDTB cité ci-dessus, et la SDRT de (Asher et Lascarides, 2003), qui a servi de point de départ à cette partie du projet. Les relations de cohérence considérées se divisent généralement en plusieurs groupes (causalité, structuration, discours rapporté, élaboration, etc.). La plupart de ces théories permettent de définir des structures hiérarchiques en créant des segments complexes à partir de segments élémentaires de façon récursive. La SDRT permet de plus d'avoir une structure de graphe plus expressive.

---

[1] Projet financé pour 3 ans par l'ANR Programme Sciences Humaines et Sociales Appel 2007 *Corpus et outils de la recherche en sciences humaines et sociales.*

Le point de vue ascendant est focalisé sur la représentation précise et complète de la structure d'un texte et est donc plus adapté sur des textes ou des unités textuelles de petite à moyenne taille, et vise donc à s'articuler avec la représentation descendante présentée dans la section suivante.

L'annotation dans le point de vue « ascendant » commence avec la segmentation d'un texte en unités de discours élémentaires. Nous avons développé un manuel d'annotation qui indique comment trouver ces unités élémentaires. Ces segments correspondent aux propositions, mais également aux syntagmes prépositionnels, adverbiaux détachés à gauche (par ex. les adverbiaux temporels et spatiaux) et incises. Sémantiquement, il est important que chaque segment contienne la description d'une éventualité (évènement ou état). Pour les relations de base, nous avons pris un ensemble restreint, c'est à dire les relations à peu près communes à toutes les théories de discours. Nous avons utilisé les résultats de travaux antérieurs sur ces relations et leurs déclencheurs, ainsi que sur les adverbiaux temporels, les temps verbaux, les changements aspectuels, pour guider le choix de la relation et du point d'attachement, et préciser les relations et leurs indices. Notre manuel d'annotation contient des descriptions pour chacune de ces relations. Nous avons commencé une campagne préliminaire d'annotation avec deux étudiants en linguistique de l'université de Toulouse pour raffiner notre manuel d'annotation et faire des premiers tests d'accord inter-annotateurs.

Nous avons aussi commencé à développer un segmenteur automatique fondé sur des indices lexicaux et syntaxiques, pour préparer le travail des annotateurs et à terme les soulager complètement de cette partie de la tâche. D'un point de vue procédural, on peut séparer le problème de l'annotation ascendante en trois tâches : le repérage des segments, la détermination des paires de segments à relier, et le typage de la relation entre ces paires de segments reliés (sous certaines contraintes pragmatiques globales).

3. Annoter des structures discursives : l'approche descendante

L'approche « descendante » vise l'annotation de structures discursives de haut niveau. Elle s'ancre dans les recherches sur les *indices de discontinuité* qui délimitent des segments à différents niveaux de grain, segments qui sont susceptibles d'être mis en relation avec une organisation fonctionnelle. Les textes sont envisagés dans leur dimension de *document* (cf. Péry-Woodley et Scott, 2006) et nous nous intéressons spécifiquement à leur *mise en texte* (y compris dans ses aspects visuels, ou mise en espace (cf. Luc et Virbel, 2001)). La mise en texte est premièrement abordée à partir d'une caractéristique incontournable : la séquentialité. Du point de vue de la séquentialité, deux stratégies sont possibles à tout moment du texte (pour le scripteur, et partant, pour le lecteur) : continuation avec ce qui précède ou rupture/transition (*shift*). La continuation étant la stratégie par défaut, les ruptures/transitions doivent faire l'objet d'une signalisation. Ainsi les indices typographiques et dispositionnels mettent à part une citation, un élément de discours rapporté, un titre… Plus intégrés dans le texte, les indices de discontinuité lexico-syntaxiques (*marqueurs de segmentation* (Bestgen, 2000) ; *shift signals* (Goutsos, 1996)) avertissent le lecteur d'un changement thématique ou rhétorique : adverbiaux détachés à l'initiale (cf. notion de *cadre de discours* (Charolles et Vigier, 2005)), redénominations (Schnedecker, 1997).

Le modèle d'annotation pour l'annotation descendante est centré essentiellement sur une méta-structure : la structure énumérative. Stratégie de base de la mise en texte à différents niveaux

de grain, la structure énumérative ne peut se définir et s'interpréter qu'à partir des indices qui la signalent (Luc et Virbel, 2001). Elle est pourtant loin d'être une simple question de formatage, et la multiplicité de ses réalisations et de ses rôles fonctionnels en fait un bon point d'entrée dans la complexité de l'organisation discursive[2] : listes formatées, découpage en sections, en paragraphes, listes « plates » à l'intérieur des paragraphes, autant de déclinaisons d'une structure où chaque item (unité énumérée) est caractérisé par une continuité, et se trouve en discontinuité par rapport à ses items voisins. A un niveau plus global, une structure énumérative constitue un segment caractérisé par une continuité tant au plan de la mise en texte qu'au plan thématique.

L'annotation manuelle se base sur une visualisation des textes enrichis d'indices potentiels des structures recherchées : indices typographiques et dispositionnels (titres de sections, listes, changement de paragraphe) ; indices lexico-syntaxiques (expressions co-référentielles, adverbiaux ou connecteurs détachés à l'initiale des phrases, paragraphes ou sections) ; redénominations démonstratives (y compris métadiscursives e.g. *cette analyse*) ; marqueurs d'énumération ; parallélismes structurels. Ce marquage est fondé sur différentes procédures de repérage automatique qui s'appuient sur la structure de document, ainsi que sur les sorties d'un étiquetage morpho-syntaxique (TreeTagger) et d'une analyse syntaxique (SYNTEX (Bourigault, 2007)). L'annotateur doit pouvoir dans un premier temps « scanner » le texte dans une approche globale, à la recherche de zones de concentration d'indices de discontinuité, pour, dans un second temps, délimiter précisément les segments par une lecture plus ou moins précise.

4. L'interface d'annotation

Il découle de ce qui précède qu'un impératif du projet ANNODIS est de disposer d'un outil d'annotation manuelle du discours permettant de travailler à différents niveaux de grain sur un même document, et assez souple pour s'adapter à différentes campagnes d'annotation. Aucun outil ne répondant à notre connaissance à ce cahier des charges, nous avons alloué une partie de nos ressources à la conception et au développement d'un tel logiciel.

La plateforme Glozz qui en résulte est suffisamment générique pour s'adapter aux modèles d'annotation utilisés dans les différentes campagnes d'ANNODIS, et pourra probablement être utilisée à d'autres fins à l'issue de ce projet. Les quatre points suivants permettent d'en comprendre les principes :

- Nous disposons simultanément de deux « vues » sur le texte de travail. La vue principale [1] est la fenêtre de travail, dans laquelle les annotations sont saisies à la souris, tandis que la vue [2], appelée « ruban », présente le texte 'vu de haut'. Dans ces deux vues, les indices pré-définis sont associés à des styles. De tels styles sont également appliqués aux annotations effectuées. Ainsi, la vue "ruban" permet une appréhension globale du marquage, et la navigation rapide au sein des zones de concentration d'indices.

- Une boîte à outils [3] permet de choisir parmi les trois types d'objets d'annotation. Les « unités » sont des blocs (apparaissant sous forme de cadres colorés dans [1] et [2]). Les

---

[2] D'autres schémas structurels pourront être ultérieurement pris en compte.

« relations » permettent de relier deux unités (apparaissant sous forme de lignes transverses dans [1] et [2]), et les « schémas » permettent d'agglomérer unités et relations en une même entité. Par exemple, pour représenter une chaîne référentielle, on pourrait créer autant d'unités que de marques anaphoriques, puis mettre en relation les différentes marques, et enfin créer un schéma intégrant toutes ces marques et leurs relations.

- Pour une campagne donnée, ou pour l'une de ses sous-tâches, un modèle d'annotation peut être chargé afin de définir les catégories disponibles pour chacun des trois types d'objet (par exemple des unités de catégorie « Amorce », des relations de catégorie « Elaboration », des schémas de catégorie « Structure énumérative » etc.). Ces catégories apparaissent dans les trois colonnes de [5], qui permettent d'assigner la catégorie souhaitée à chaque annotation. Le modèle d'annotation prévoit aussi d'associer un jeu d'attributs/valeurs à chaque catégorie, lequel peut être vu et modifié dans la fenêtre [4]. Des styles (notamment des couleurs) sont associés aux différents types, et peuvent être modifiés grâce à l'interface [7].

- Le document traité peut être enrichi d'un marquage préalable, qui peut être issu d'une segmentation ou d'un repérage automatique d'indices ou de structures comme exposé en 3 ; ou résulter d'une précédente campagne d'annotation que l'on souhaite enrichir ou sur laquelle on souhaite s'appuyer pour étudier de nouvelles structures.

Enfin, différents outils de recherche et de navigation [6] peuvent être ajoutés ou retirés à la demande.

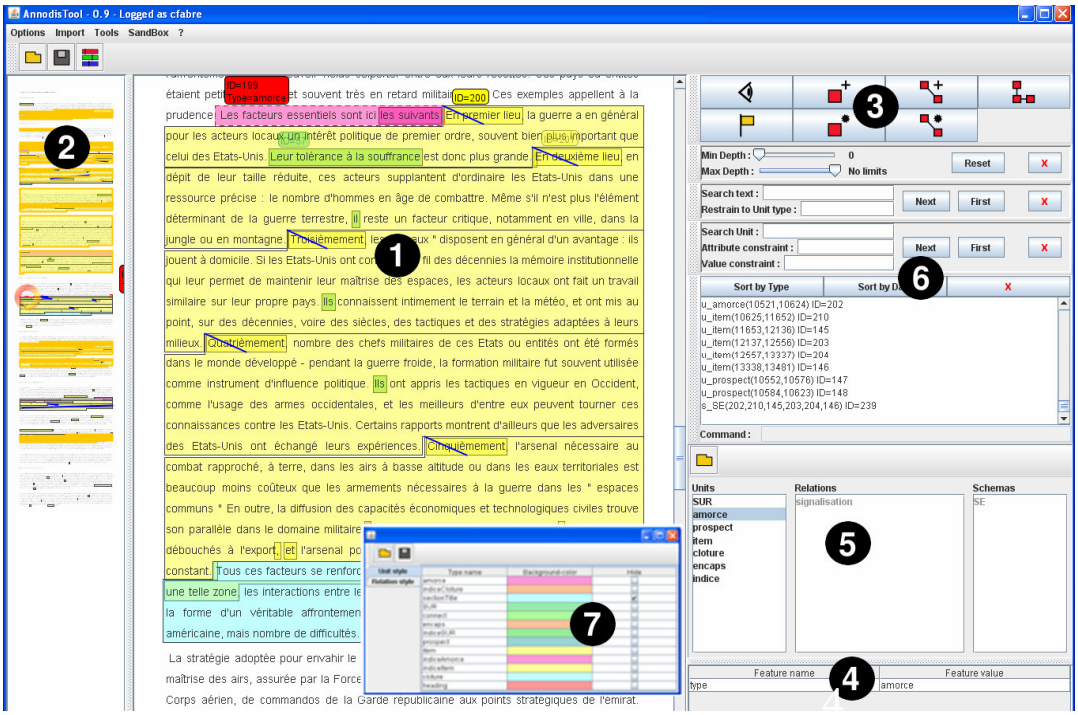

Figure 1 : La plateforme d'annotation Glozz

5. Enjeux et retombées

La constitution d'un corpus de textes diversifiés, dont beaucoup sont longs et complexes sur le plan de la structure de document, avec prise en compte de cette structure « logique » et

préservation de la mise en forme matérielle est en soi novateur. L'annotation systématique de relations et de structures discursives dans ce corpus diversifié fournira des données précieuses pour la description linguistique et la théorisation des fonctionnements discursifs. Déjà la rédaction des guides d'annotation a conduit à un travail considérable d'affinement et de précision des modèles. La construction d'outils logiciels satisfaisant aux exigences des deux approches ascendante et descendante a exigé une modélisation qui en garantit la robustesse et la généricité. Enfin, la confrontation/fusion des deux approches à travers cette double annotation est un enjeu majeur du projet.

Références

ASHER, N., LASCARIDES, A. (2003). *Logics of conversation.* Cambridge; New York: Cambridge University Press.

BESTGEN, Y., VONK, W. (2000). The role of temporal segmentation markers in discourse processing. *Disourse Processes* 19, 385-406.

BOURIGAULT D. (2007) *Un analyseur syntaxique opérationnel : SYNTEX.* Mémoire d'HDR en sciences du langage, Université de Toulouse 2, France.

CHAROLLES, M., & VIGIER, D. (2005). Les adverbiaux en position préverbale: portée cadrative et organisation des discours. *Langue Française* 148, 9-30.

FORBES, K., MILTSAKAKI, E., PRASAD, R., SARKAR, A., JOSHI, A., WEBBER, B. (2003). D-LTAG System: Discourse parsing with a lexicalized tree-adjoining grammar. *Journal of Logic, Language and Information*, 12(3), 261-279.

GOUTSOS, D. (1996). A model of sequential relations in expository text. *Text* 16(4), 501-533.

LUC, C., VIRBEL, J. (2001). Le modèle d'architecture textuelle : fondements et expérimentation. *VERBUM* 23(1), 103-123.

MANN, W., THOMPSON, S. (1987). Rhetorical Structure Theory: a theory of text organization. Information Science Institute.

PÉRY-WOODLEY, M.-P., SCOTT, D. (2006). Computational Approaches to Discourse and Document Processing. *TAL* 47(2), 7-19.

POLANYI, L., VAN DEN BERG, M., CULY, C., THIONE, G. L., AHN, D. (2004). A Rule Based Approach to Discourse Parsing. Actes de *SIGDIAL 2004*. Boston.

PRASAD, R., MILTSAKAKI, E., DINESH, N., LEE, A. JOSHI, A. (2006). *The Penn Discourse TreeBank 1.0 Annotation Manual*. http://www.seas.upenn.edu/~pdtb/papers/pdtb-1.0-annotation-manual.pdf

SCHNEDECKER, C. (1997). *Nom propre et chaînes de référence.* Paris: Klincksieck.

WOLF, F., & GIBSON, E. (2005). Representing Discourse Coherence: A Corpus Based Study. *Computational Linguistics*, 31(2), 249–287.