Annotation des structures discursives : l'expérience ANNODIS

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

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801286

Annotation des structures discursives : l'expérience ANNODIS

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

Université de Toulouse – UTM, CLLE-ERSS
{hodac et pery}@univ-tlse2.fr

1 Introduction

Cet article propose une présentation réflexive de l'expérience ANNODIS, qui a abouti à la diffusion d'un corpus diversifié de français écrit enrichi d'annotations concernant le niveau discursif, ainsi qu'à la création de l'interface Glozz spécialisée dans l'annotation discursive de documents. L'originalité du projet ANNODIS réside essentiellement dans la mutualisation de deux approches complémentaires qui permettent de poser un certain nombre de questions concernant l'annotation de structures discursives. La ressource ANNODIS[1] ayant fait ailleurs l'objet de présentations détaillées (Péry-Woodley *et al.,* 2011, Afantenos *et al.,* 2012), l'objectif ici n'est pas de la décrire mais de revenir sur ses enjeux, à la fois en tant que ressource pour la linguistique du discours, et en tant qu'expérience d'annotation. La section 2 contextualise et problématise les choix d'annotation. La section 3 fournit des éléments de réflexion sur la campagne d'annotation, en particulier sur certains biais découverts *a posteriori* et leur impact potentiel sur les annotations. En section 4, nous évoquons l'exploitation des données générées, notamment la nécessité – et la difficulté – d'en fournir des aperçus visualisables[2] à même de tirer un réel parti de leur richesse.

2 Annoter des structures discursives

2.1 Quand aucun choix d'annotation ne va de soi

L'annotation de structures discursives est un enjeu actuel particulièrement riche de promesses à en juger par l'essor récent de projets qui proposent ce type d'annotation. Cependant, force est de constater que les phénomènes discursifs annotés dans ces projets sont marqués par une grande hétérogénéité : relations rhétoriques, transitions temporelles, chaînes de référence, structures énumératives, zones argumentatives, segmentation thématique, etc. Dès la segmentation en unités élémentaires l'absence de consensus est flagrante, et on peut dire sans exagération qu'au niveau discursif aucun choix d'annotation ne va de soi. Dans une synthèse centrée sur l'analyse automatique (*Discourse Parsing*), Marcu attribue les difficultés de ce « jeune champ de recherche » au manque de maturité de la linguistique du texte et du discours (Marcu, 2006) : diversité et instabilité des modèles, manque de validation empirique, éclatement des travaux descriptifs... Ce manque de maturité, qui rend d'autant plus nécessaire la constitution de ressources pour l'analyse empirique des fonctionnements discursifs, est à son tour intimement lié à la complexité définitoire inhérente aux structures discursives. Pour la réflexion sur l'expérience ANNODIS dont nous saisissons ici l'occasion, nous nous référons à une autre synthèse récente qui a l'intérêt de tenter, à l'intention de chercheurs en ingénierie des langues, une présentation unifiée des principaux acquis en linguistique du discours (Webber *et al.,* 2012). Nous nous demanderons quels sont, parmi les types de structuration discursive qui y sont ainsi présentés « de haut », ceux que les annotations de la ressource ANNODIS recouvrent ou touchent. Nous nous proposons donc de partir de l'intéressante tentative de Webber *et al.* de dépasser l'éclatement des travaux sur le discours pour porter un autre regard, *a posteriori*, sur notre expérience d'annotation[3].

Article available at http://www.shs-conferences.org or http://dx.doi.org/10.1051/shsconf/20140801286

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801286

La définition consensuelle de « structure discursive » proposée par les auteurs comme traversant les études rassemblées dans leur état de l'art est celle-ci (Webber *et al.,* 2012 : 439) :

> Discourse structures are the *patterns* that one sees in multi-sentence (multi-clausal) texts. Recognizing these pattern(s) in terms of the elements that compose them is essential to correctly deriving and interpreting information in the text.

Les structures discursives sont ici définies *a minima* comme des motifs qui couvrent des portions de texte dépassant la phrase, et qui doivent être reconnus pour qu'il y ait compréhension du texte. Au-delà de cette définition très générale, les structures discursives peuvent recouvrir une large palette de phénomènes plus ou moins liés. Webber *et al.* (2012) distinguent quatre types de structuration : par topiques, fonctions, « éventualités », et relations de discours. Nous les parcourons en faisant le lien entre l'état de l'art proposé et d'autres travaux qui ont contribué à former le socle de l'expérience ANNODIS.

La structuration en topiques a trait à l'organisation thématique des textes telle qu'elle apparaît à travers les ensembles d'entités constituant des chaînes ; elle englobe des structures discursives où l'on peut discerner, en dépit de la diversité des dénominations et des paradigmes, une certaine cumulativité des résultats et des éléments de stabilisation : les progressions thématiques (Daneš, 1974), la théorie du centrage (Walker *et al.,* 1998), les chaînes de références (Charolles, 2002 *inter alia*).

La structuration en termes de fonctions rassemble en revanche des approches très diverses : organisation rhétorique plus ou moins conventionnelle selon les genres (cf. notions de *move* (Swales, 1981), de zone rhétorique (Teufel et Moens, 2002)), structure intentionnelle telle que définie par Grosz et Sidner (1996). Webber *et al.* évoquent la difficulté de modéliser les intentions humaines et leurs relations, et expliquent ainsi la rareté des travaux empiriques sur la structure intentionnelle. Dans le cadre de l'annotation des structures multi-échelles pour ANNODIS, en convergence avec le modèle d'architecture textuelle (Luc et Virbel, 2001), les intentions visées sont celles qui sont centrées sur l'organisation textuelle : pour qu'un motif soit compris, il est nécessaire que l'intention qu'il soit perçu (intention qui sous-tend chez le scripteur l'emploi de ce motif) soit reconnue par le lecteur. Il en va ainsi de l'identification des items d'une structure énumérative. De même, en interprétant un segment comme « prédictif » (une amorce d'énumération par exemple, cf. Tadros, 1994), ou une marque de discontinuité dans la séquentialité du discours (Goutsos, 1996), le lecteur reconnaît l'intention du scripteur de rendre perceptible cette prédiction ou cette discontinuité.

La structuration en « éventualités » est le domaine de l'organisation spatio-temporelle des événements et états relatés dans les textes. Ce type de structuration est associé par Webber *et al.* (2012) à la structuration par les relations de discours, que plusieurs modèles s'attachent à définir depuis les années 1980 : ainsi la *Rhetorical Structure Theory* (RST, Mann *et al.,* 1988), la *Segmented Discourse Representation Theory* (SDRT, Asher et Lascarides, 2003, base de l'annotation en relations rhétoriques dans ANNODIS) proposent une approche de la cohérence définie par la construction récursive (ou la décomposition pour la RST appliquée de façon descendante) de segments complexes reliés par des relations rhétoriques.

À cette complexité en quatre types de structuration s'ajoute le fait que ces types ne structurent pas de façon isolée. En effet, un motif peut être impliqué dans plusieurs structures conjointement, ce qui se matérialise à la surface du texte par une complexité de signalisation. Complexité encore intensifiée par le fait que les traits linguistiques impliqués dans la signalisation d'un motif, dont le rôle « instructionnel » est de signaler au lecteur plus ou moins explicitement une (dis)continuité avec le discours précédent (topicale, intentionnelle, événementielle) et/ou une relation rhétoriques entre des segments, nous semblent fonctionner en faisceaux, avec des effets de seuil, plutôt que comme des marques discrètes (voir section 4 et Ho-Dac *et al.,* 2012 ; 2013).

Face à cette complexité qui explique l'instabilité du domaine, le développement et la mise à disposition de ressources – corpus enrichis, outils de traitement et d'aide à l'annotation – sont des enjeux majeurs. Nous proposons au terme de l'expérience ANNODIS un corpus diversifié construit de manière raisonnée, mis en

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801286

forme selon les standards actuels, et enrichi de multiples annotations, avec l'idée de fournir une ressource partageable à différents niveaux d'enrichissement, et de contribuer ainsi à la constitution de communautés de chercheurs. Les objets d'annotation choisis, si on les envisage à la lumière de la synthèse de Webber *et al.* (2012), se répartissent sur les quatre types de structuration distingués : l'annotation en chaînes pour le premier type, l'annotation des structures énumératives pour le second[4], la segmentation manuelle en unités minimales de discours pour le troisième, et l'annotation des relations rhétoriques entre ces unités pour le dernier. Dans un domaine encore jeune comme le discours, les annotations constituent moins une référence définitive qu'une ressource partagée pour expérimenter des processus d'annotation, confronter les modèles et tester les hypothèses. Nous soulignons la valeur de ce type d'expérience pour l'élaboration théorique, à travers le travail d'explicitation et d'opérationnalisation des modèles d'abord, comme en témoignent les guides d'annotation (Muller *et al.,* 2012 ; Colléter *et al.,* 2012), par le passage à l'échelle dans la mise à l'épreuve de ces modèles ensuite. De plus, la systématicité des annotations permet d'appliquer ou plutôt d'expérimenter au niveau discursif des techniques de linguistique de corpus outillée et de TAL (Traitement Automatique des Langues) généralement élaborées pour des phénomènes moins complexes et plus locaux : corrélations statistiques, fouille de données et apprentissage automatique.

Le caractère expérimental d'ANNODIS s'est trouvé renforcé par la double approche qui, partant de démarches différentes initialement exprimées en termes d'objets à annoter – relations rhétoriques *versus* structures multi-échelles –, a conduit les participants à questionner l'implicite qui enveloppait le projet. Les objectifs d'ANNODIS sont apparus comme étant à la fois de stabiliser un certain nombre de définitions linguistiques de motifs discursifs ciblés (les structures multi-échelles) et de confronter aux données réelles une modélisation spécifique de la construction de la cohérence discursive (par relations rhétoriques). Les implications de ce double objectif en termes de décisions à chaque moment de la campagne d'annotation (choix de corpus, d'annotateurs, de définition du protocole…) sont à l'origine du mouvement réflexif dont nous tentons de présenter des éléments dans ce qui suit.

2.2 Annotation en structures multi-échelles : chaînes topicales et structures énumératives

L'annotation des structures multi-échelles 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 intentionnelle. Les textes sont envisagés dans leur dimension de document (cf. Péry-Woodley et Scott, 2006) avec une prise en compte spécifique de 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 discontinuité. La continuation étant la stratégie par défaut, les discontinuités 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 avertissent le lecteur d'un changement thématique ou rhétorique : adverbiaux détachés à l'initiale, redénominations. Cette approche s'articule avec l'approche en relations rhétoriques via l'hypothèse d'une influence de structures de haut niveau sur l'interprétation au niveau propositionnel et interpropositionnel.

Le modèle d'annotation en structures multi-échelles est centré sur deux stratégies discursives et deux motifs textuels susceptibles d'apparaître à de très hauts niveaux d'organisation : l'empaquetage, réalisé par les structures énumératives ; le chaînage, réalisé par les chaînes topicales. 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 : listes formatées, découpage en sections, en paragraphes,

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801286

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. À un niveau plus global, une structure énumérative constitue un segment caractérisé par une continuité au plan de la mise en texte comme au plan thématique. En ce qui concerne les chaînes topicales, elles sont définies de manière non restrictive comme un type particulier de chaîne de cohésion où les éléments sont des unités contenant un même référent (potentiellement) topical, et où le segment résultant est constitué de l'ensemble d'unités connectées.

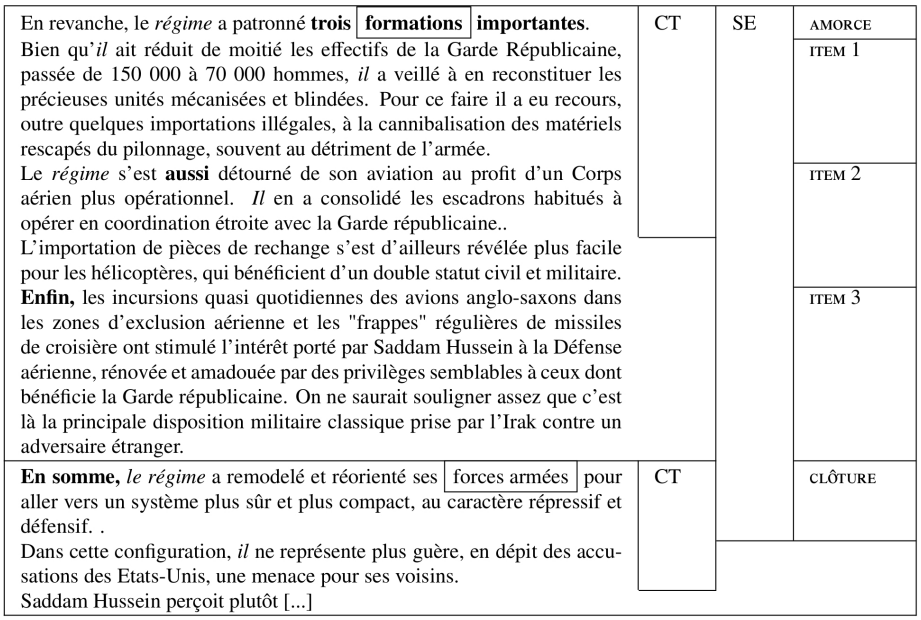


Figure 1 : Exemple de structures multi-échelles annotées. Les structures et leurs éléments sont indiqués dans les colonnes de droite. La colonne de gauche donne le texte correspondant, les sauts de paragraphes y sont représentés par des lignes. Les indices annotés apparaissent en italique pour les indices de chaîne topicale, et en gras pour les indices de structure énumérative. CT = chaîne topicale, SE = structure énumérative.

L'exemple de la figure 1 donne à voir un passage contenant deux chaînes topicales (CT) autour du *Régime de Saddam Hussein*, et une structure énumérative (SE) listant les trois formations armées dirigées par ce régime. Comme l'illustre l'exemple, les chaînes topicales correspondent à des segments mono-blocs alors que les structures énumératives présentent une organisation interne en trois types de composants : une amorce optionnelle mais généralement présente, au moins deux items, et une clôture optionnelle qui s'avère assez peu fréquente.

2.3 Annotation en relations rhétoriques

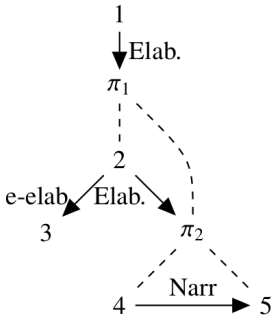
L'annotation en relations rhétoriques s'inscrit dans le cadre des théories du discours qui visent à construire une structure complète d'un discours vu comme un ensemble d'unités élémentaires reliées par des relations de cohérence. Comme on l'a évoqué plus haut, cette vision théorique du discours rassemble principalement les travaux autour de la RST (Mann *et al.,* 1988 ; Carlson *et al.,* 2003), du Penn Discourse Treebank (Prasad *et al.,* 2006) et de la SDRT (Asher et Lascarides, 2003). C'est la SDRT qui a servi de point de départ pour l'annotation en relations rhétoriques d'ANNODIS.

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801286

L'annotation en relations rhétoriques commence avec la segmentation exhaustive d'un texte en unités de discours élémentaires. Ces segments correspondent aux propositions, mais également à des syntagmes prépositionnels, adverbiaux détachés à gauche (par ex. les adverbiaux temporels et spatiaux) et à des incises. Sémantiquement, chaque segment correspond à la description d'une éventualité unique (événement ou état). Pour l'étape suivante, rattachement des unités et typage des relations, un ensemble restreint de 17 relations de base, à peu près communes à tous les modèles mentionnés ci-dessus, a été sélectionné. Ces relations, pour lesquelles le guide d'annotation en relations rhétoriques donne le détail des définitions et procédures d'annotation (Muller *et al.,* 2012), sont les suivantes :

- trois relations de causalité : explication, résultat, but ;
- trois relations « intentionnelles » : parallélisme, contraste, continuation ;
- deux relations logiques : alternative et condition ;
- une relation d'attribution (discours rapporté) ;
- cinq relations d'exposition/narration : arrière-plan, narration, *flashback*, encadrement, localisation temporelle ;
- deux relations d'élaboration: élaboration (d'événement), élaboration d'entité ;
- une relation de commentaire.

Selon ce modèle, une structure discursive est représentée par un graphe, dont les nœuds sont des unités discursives et dont les arcs étiquetés représentent les relations rhétoriques entre ces unités. Les unités discursives peuvent être simples, les EDU ou *Elementary Discourse Units*, ou complexes, les CDU ou *Complex Discourse Units*, qui regroupent plusieurs EDU en une structure discursive (un graphe). La figure 2 donne un exemple de représentation en graphe d'une brève journalistique.



[Milutinovic devant le TPI.]_1 [L'ancien président de Serbie Milan Milutinovic, [inculpé aux côtés de l'ex-chef d'Etat yougoslave Slobodan Milosevic pour crimes de guerre au Kosovo,]_3 s'est rendu volontairement hier au Tribunal pénal international pour l'ex-Yougoslavie de La Haye.]_2 [Arrivé aux Pays-Bas dans un avion du gouvernement yougoslave,]_4 [M. Milutinovic a été incarcéré au centre de détention du Tribunal pénal international en début d'après-midi.]_5

Figure 2 : Exemple de relations rhétoriques annotées. Les nœuds correspondent aux unités discursives: les EDU représentées par leur numérotation et les CDU par un nœud étiqueté πn. Les arêtes avec flèches représentent les relations rhétoriques, les arêtes en pointillé sans flèches représentent l'inclusion d'EDU dans un CDU. Elab. = Élaboration, e-elab = Élaboration d'entité, Narr. = Narration.

2.4 Corpus et annotations ANNODIS

Comme indiqué en introduction, les annotations ANNODIS ont été réalisées sur un corpus diversifié. Ce corpus diversifié a été conçu en vue de permettre des études comparatives et de mesurer la pertinence multi-genre des modèles d'annotation. Cependant, tous les textes sont de type expositif, plus propice à des organisations complexes et à des relations variées. La variation multi-genre est davantage liée à des questions de domaines, de format de diffusion et de visée discursive. La constitution de ce corpus a en premier lieu été déterminée par les besoins des différentes approches en termes de données à annoter.

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801286

Pour l'annotation en relations rhétoriques, seuls des textes courts pouvaient être envisagés vu la complexité de la tâche et l'exigence d'une annotation exhaustive de l'intégralité des textes. Pour les structures multi-échelles, il était impératif d'annoter des textes longs et structurés (structures de haut niveau, prise en compte de la structure de document) afin d'éprouver la capacité organisationnelle multi-échelle des structures sélectionnées. Par ailleurs, le caractère sporadique de l'annotation en structures multi-échelles requiert des textes assez longs (à l'inverse du « pavage » complet qu'implique l'annotation en relations rhétoriques) : la totalité du texte ne rentre pas nécessairement dans un schéma de type chaîne topicale ou structure énumérative. De ce fait le corpus propose à la fois des textes courts et des textes longs. Les textes courts ont été extraits du corpus court constitué dans le cadre du projet « PASSAGE (produire des annotations syntaxiques à grande échelle pour aller de l'avant), traitant de l'évaluation des analyseurs syntaxiques du français »[5]. Seules les brèves journalistiques issues du quotidien l'Est Républicain (NEWS) et les extraits d'articles Wikipedia (WIK1) ont été sélectionnés. Le corpus de textes longs a été spécifiquement constitué pour ANNODIS avec pour objectif la constitution d'un corpus de textes longs, entiers, libres de droits, normés selon la TEI-P5, permettant entre autres le renseignement des méta-données et de la structure du document. Ce corpus brut constitue une première partie de la ressource diffusée. Il est composé de trois types de textes étiquetés WIK2 (articles Wikipedia complets), LING (articles de linguistique issus du premier Congrès mondial de Linguistique Française – CMLF2008) et GEOP (rapports publiés par l'IFRI, *think tank* français, dans le domaine de la géopolitique).

Le choix des textes a donc également été déterminé par les possibilités de diffusion des textes annotés. Les textes issus de Wikipédia (WIK1 et WIK2) sont libres sous licence Creative Commons. Ceux issus de l'Est Républicain (NEWS) sont diffusables dans les mêmes conditions que dans le cadre du CNRTL. Pour les articles universitaires (LING) et les rapports (GEOP), des contrats émis par la cellule juridique de la délégation régionale 14 du CNRS ont été signés par les différents propriétaires intellectuels pour permettre la diffusion des corpus annotés sous licence Creative Commons. Le tableau 3 fait état du bilan global des annotations disponibles. Ces données ont été largement commentées dans les articles cités en introduction.

	Nombre de mots	**Nombre de textes**	**EDU**	**Relations rhétoriques**	**CDU**	**SE**	**CT**
NEWS	9 768	39	1159	1203	510		
WIK1	17 330	42	1949	2034	829		
WIK2	231 000	30	53	65	38	401	266
LING	169 000	25	12	14	9	297	88
GEOP	266 000	32	15	19	9	293	234
ANNODIS	**687 000**		**3188**	**3355**	**1395**	**991**	**588**

Tableau 1 : Relations rhétoriques et structures multi-échelles dans la ressource ANNODIS. EDU = Unité de discours élémentaire ; CDU = Unité de discours complexe. SE = Structure Énumérative ; CT = Chaîne Topicale

La ressource ANNODIS propose un volume d'annotations suffisant pour remplir les objectifs initiaux de chaque approche. Concernant l'annotation en relations rhétoriques, un certain nombre de postulats peuvent désormais être (in)validés empiriquement, comme par exemple la contrainte de la « frontière droite », hypothèse forte de la SDRT (Afantenos et Asher, 2010). De même, la segmentation systématique en EDU permet de décrire en corpus la nature des unités minimales du discours et le grand nombre de relations annotées associé au caractère multi-genre du corpus permet de dresser un inventaire quantifié des relations rhétoriques (Vergez-Couret, 2012).

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801286

Concernant l'annotation en structures multi-échelles, l'objectif initial était de fournir des données pour la définition d'objets discursifs aux contours pas encore entièrement stabilisés. L'annotation fournit un nombre important de chaînes topicales et de structures énumératives, ce qui indique que les définitions proposées dans les guides ciblent des phénomènes fréquents, avec une fréquence relativement stable à travers différents genres (cf. Ho-Dac *et al.,* 2010). La confiance dans ces définitions est renforcée par la qualité correcte de l'accord inter-annotateur des annotations en structures multi-échelles (F-mesure de 0,7, voir section suivante). Dans le but de faciliter l'accès à ces objets, un explorateur permet de naviguer de structure en structure et de visualiser ces structures en contexte.

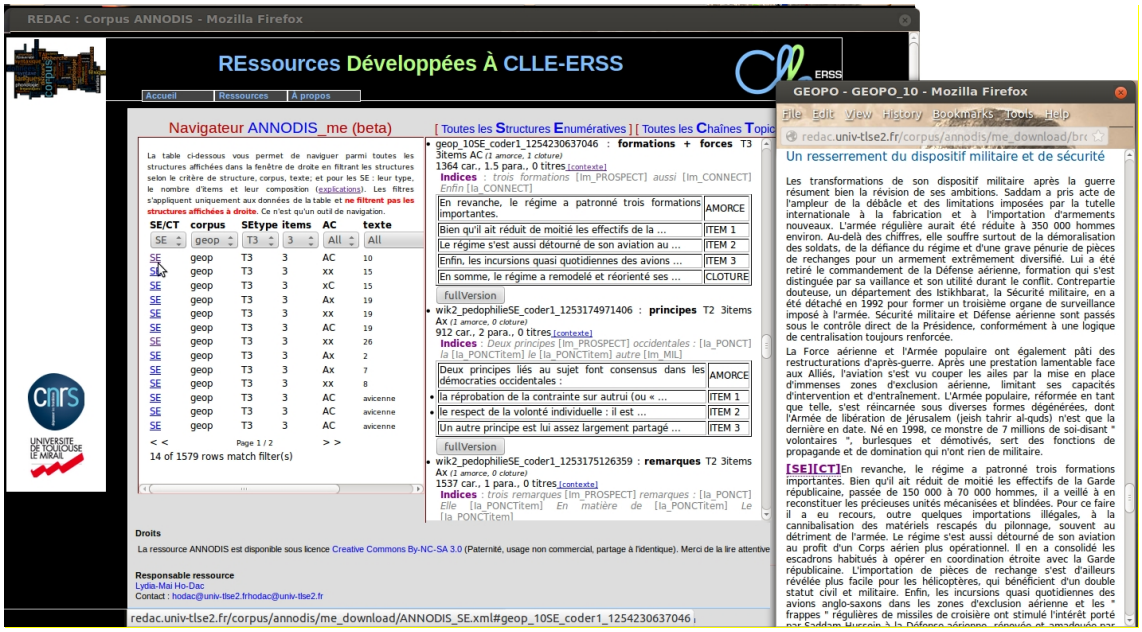


Figure 3 : Navigateur ANNODIS pour explorer les structures multi-échelles annotées (fenêtre centrale) et les visualiser en contexte (fenêtre de droite)

La figure 3 capture l'interface de navigation permettant de trouver la structure énumérative de l'exemple (1) « *En revanche...* » en sélectionnant un certain nombre de paramètres tels que le corpus, le texte, la composition de la SE (avec ou sans amorce, avec ou sans clôture, avec 2 items...) ou encore le type de SE (selon une typologie déduite présentée dans Ho-Dac *et al.,* 2010 et résumée en section 4). Un lien permet d'ouvrir une fenêtre pop-up localisant la structure dans l'article complet.

3 Premiers pas d'une jeune ressource

Les résultats de la campagne d'annotation ANNODIS tels que présentés dans le tableau 3 ci-dessus ne représentent en fin de compte que la naissance d'une ressource. Une fois la récolte des annotations réalisée, un ensemble de post-traitements et d'analyses préliminaires sont nécessaires pour présenter et mettre à disposition la ressource. Cette section parcourt différentes questions qui nous sont apparues lors de cette période de premiers pas, concernant en particulier les procédures d'annotation de chaque approche, le choix de l'annotation « naïve », et un certain nombre de biais qui nous semblent pour l'instant difficilement évitables.

Selon le type d'objet annoté, différentes procédures ont été mises en place afin de permettre une annotation valorisable. Ces procédures d'annotation sont largement détaillées dans Péry-Woodley *et al.* (2012) ainsi que dans les deux guides d'annotation publiés. Concernant l'annotation en relations

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801286

rhétoriques, deux étapes ont été distinguées donnant chacune lieu à une mesure de l'accord inter-annotateur. Les annotateurs ont dans un premier temps segmenté les textes en unités de discours élémentaires afin d'arriver à une version segmentée commune des textes. Cette version segmentée a été annotée en relations rhétoriques dans un deuxième temps. Ce deuxième temps consistait à relier chaque unité minimale à une autre unité et à catégoriser le lien établi. Ces étapes sont expliquées en détail dans le guide d'annotation en relations rhétoriques (Muller *et al.,* 2012).

Concernant les structures multi-échelles, leur annotation s'est déroulée en une seule étape qui consistait à délimiter les composants des structures (chaîne topicale, amorce, item et clôture), indiquer les indices de surface permettant la reconnaissance de ces composants, et enfin regrouper le tout dans une structure catégorisée CT ou SE. Cette annotation a été précédée d'un étiquetage automatique d'un certain nombre de traits, afin de permettre aux annotateurs de survoler le texte en zoomant sur des zones textuelles présentant des indices potentiels. L'annotation des structures multi-échelles est détaillée dans Colléter *et al.* (2012), où le manuel d'annotation est accompagné d'un certain nombre de « bonus » comme des témoignages d'annotateurs, des exemples d'arbitrage, le détail des postraitements réalisés, les mesures interannotateurs et une reformulation a posteriori des définitions du modèle d'annotation.

Pour les deux approches, le choix initial a été de faire appel à des annotateurs naïfs. Ce choix peut être remis en cause, selon que l'on conçoit l'annotation comme une expérimentation – une sorte d'enquête – ou comme un moyen de constituer une ressource « de référence », un *gold standard*. En effet, une annotation naïve va de pair avec un objectif d'évaluation de la réalité intuitive (voire cognitive) des phénomènes étudiés ; elle est sous-tendue par une hypothèse forte qu'un objet annotable (pour lequel on aboutit à une stabilisation des annotations, c'est-à-dire à des annotations montrant un accord correct entre annotateurs) est un objet *définissable* (dont on peut stabiliser la définition). Si l'objectif est d'emblée la constitution d'une ressource « de référence », l'annotation naïve peut donner lieu à des « erreurs » nécessitant une réannotation experte. C'est le cas des annotations en relations rhétoriques, qui sont diffusées à la fois dans une version « naïve » et dans une version « experte », cette dernière ayant pour objectif l'évaluation des postulats des modèles sous-jacents au manuel d'annotation. Les structures multi-échelles annotées sont quant à elles uniquement diffusées dans une version « naïve ». Pour ces structures, seuls les textes multi-annotés utilisés pour mesurer l'accord inter-annotateur ont fait l'objet d'un arbitrage afin de diffuser un jeu d'annotation par texte (des exemples d'arbitrage sont donnés dans Colléter *et al.* (2012)).

Lors de ces procédures d'annotation, un certain nombre de biais se sont avérés inévitables, biais qui sont généralement associés au guidage des annotateurs au cours de la campagne, que ce soit par les guides d'annotation ou les prétraitements. Dans la ressource ANNODIS, ces biais concernent principalement le rôle des indices de surface dans la définition des objets à annoter, et l'influence de l'affichage des textes à annoter. Du côté des indices de surface, il est évident que l'indication dans le guide d'annotation de « connecteurs typiques » pour certaines relations rhétoriques a grandement influencé les annotateurs, qui ont pu mettre en place un certain nombre de routines, aboutissant parfois à des « erreurs » d'annotation systématiques. Pour ce qui est de l'annotation en structures multi-échelles, c'est l'influence des indices prémarqués automatiquement qui est en cause, et pratiquement impossible à mesurer. Cette influence potentielle concerne à la fois les structures et les indices annotés. Cependant, sans ce prémarquage automatique, le caractère sporadique des structures aurait risqué de fortement handicaper leur identification.

Un autre biais difficilement évaluable a trait à la distance entre les données à annoter et les données primaires *i.e.* les données d'origine. Dans l'élaboration de cette ressource, une attention particulière a été apportée à cette question. L'interface d'annotation Glozz (Mathet et Widlöcher, 2009) a été spécifiquement développée pour permettre un affichage des textes à annoter au plus près de leur version originale. Cependant, certaines modifications de mise en forme se sont immiscées, nous amenant à prendre conscience de l'importance de mesurer la distance entre le matériel à annoter et le document d'origine. Pour illustrer ce problème rarement évoqué, nous citerons deux exemples : le premier concerne

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801286

l'annotation de textes courts en relations rhétoriques ; le second, l'annotation de textes longs en structures multi-échelles.

Dans le projet ANNODIS, les textes ont été présentés aux annotateurs via l'interface Glozz qui permet un affichage des textes avec leur mise en forme. Concernant l'annotation en relations rhétoriques, les seuls indices de mise en forme matérielle conservés par rapport aux originaux sont les sauts de paragraphes. Par conséquent, les titres de section ont parfois être transformés et intégrés dans le corps de texte. Ce type de transformation a pu sérieusement affecter l'interprétation des textes (et donc leur annotation) comme l'illustre l'exemple suivant dans lequel le titre de la brève journalistique « Une pluie d'étoiles » a été intégré au corps de l'article.

> [Une pluie d'étoiles]_1 [Non !]_2 [Il ne s'agit pas d'un phénomène météorologique accompagnant le solstice d'été.]_3 [Plus simplement,]_4 [les hasards du calendrier du Comité départemental d'action touristique a fait coïncider la promotion de l'Office de tourisme avec le nouveau classement de l'hôtel-restaurant Le Relais à Arc-et-Senans.]_5

Concernant les structures multi-échelles, malgré le souci de proposer aux annotateurs une version à annoter proche des documents originaux, certains choix et certaines erreurs de transformation ont également pu influencer l'annotation. Relevons d'abord les choix d'effacement de contenus « non textuels » tels que les tableaux et images insérés dans les documents. Cet effacement systématique enlève un contenu essentiel et souvent structurant, ce qu'illustre l'exemple de la figure 4, où le parallélisme entre items est renforcé par les schémas récurrents, schémas enlevés de la version à annoter.

Figure 4 : Influence des contenus graphiques dans l'interprétation d'un document

En second lieu, certaines influences de modifications d'affichage ont pu être observées, comme l'illustre l'exemple de la figure 5, où l'incohérence des puces et la modification de l'inscription spatiale des items a induit en erreur l'annotateur, qui n'a pas délimité correctement le deuxième item de l'énumération.

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801286

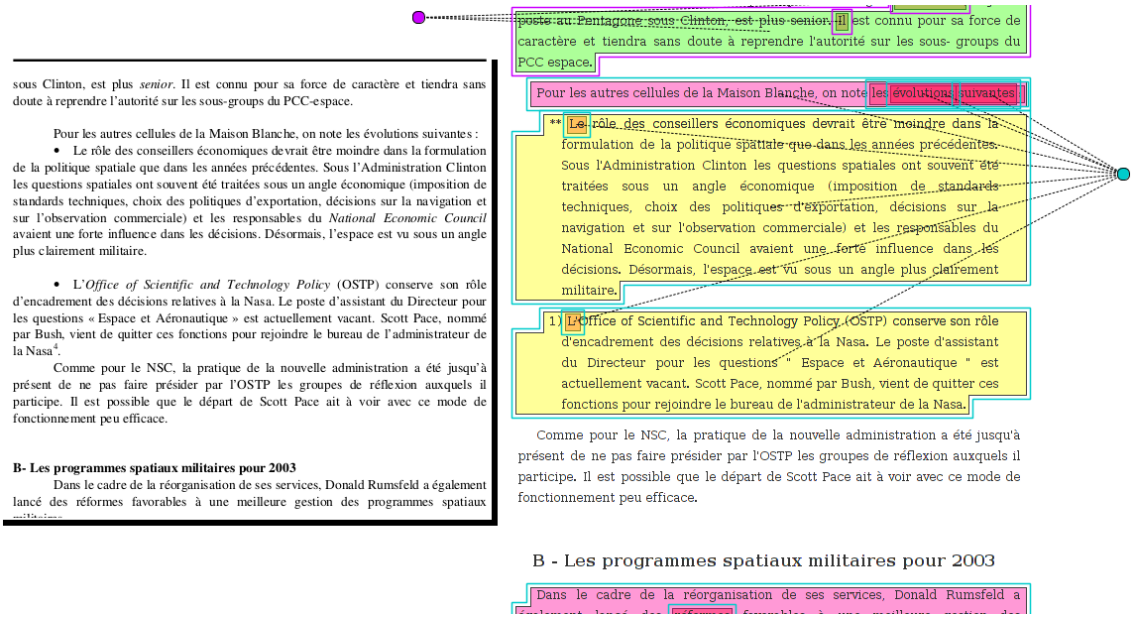


Figure 5 : Erreur d'annotation (2e item incomplet) due à une transformation malencontreuse de la mise en forme matérielle d'un document original (à gauche) en un document à annoter (à droite). Dans cette capture d'écran de l'interface d'annotation Glozz (à droite), les segments roses correspondent aux amorces, les jaunes aux items et les verts aux chaînes topicales. Le point bleu en marge droite indique une structure énumérative reliée aux différents segments colorés qui la composent (amorce, items, indices). Le point violet en marge gauche indique une chaîne topicale reliée aux différents segments colorés qui la composent (expressions coréférentielles)

4 Exploitation des annotations

Cette dernière section propose un petit inventaire des premières exploitations des différentes annotations, en mettant l'accent sur des méthodes visant à tirer le meilleur parti de données aussi volumineuses et complexes.

La première étape d'analyse concerne un peu inévitablement l'analyse quantitative des annotations obtenues. En effet, avant de pouvoir « plonger » qualitativement dans les données, un aperçu des données à disposition et de leur variété s'impose pour s'approprier l'objet obtenu.

L'accord inter-annotateur est généralement mesuré en cours d'annotation, afin d'ajuster les manuels pour assurer une campagne d'annotation efficace. Cependant, sa mesure constitue également un élément d'analyse des annotations récoltées. Par exemple, au niveau du typage des relations rhétoriques, les cas d'accords et de désaccords ont permis très tôt de distinguer différents types des relations, notamment en distinguant des relations « difficiles » à annoter ou encore des relations fréquemment confondues (cf. Atallah *et al.,* 2013). Ces différents cas permettent la mise en place d'analyses qualitatives sur des jeux de relations particulières (cf. Vergez-Couret, 2010), ce qui amène parfois à une réannotation de certains phénomènes ciblés[6].

En ce qui concerne les structures multi-échelles, l'analyse quantitative des annotations constitue une étape obligatoire pour appréhender ces motifs aux contours plutôt flous. Les annotations ont alors pour vocation première de rassembler les réalisations linguistiques d'un objet ayant fait l'objet d'une définition ouverte, de manière à préciser et stabiliser cette définition. Par exemple, la caractéristique multi-échelle et la

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801286

complexité de signalisation (qui sont deux facteurs étroitement liés) peuvent être appréhendés selon différents points de vue que la ressource permet de manipuler. L'insertion des structures dans la mise en forme matérielle du document peut ainsi être évaluée par des mesures de corrélations afin d'aboutir à la définition d'une typologie des structures par niveaux de grain, telle que présentée dans Ho-Dac *et al.* (2010). Cette typologie a ensuite été intégrée comme paramètre de filtre dans le navigateur ANNODIS (figure 2) qui distingue alors quatre types de structures énumératives : les sections titrées (type 1), les listes formatées (type 2), les structures multiparagraphiques (type 3) et les structures intraparagraphiques (type 4).

Les annotations peuvent également servir de base à la mise en œuvre de techniques de visualisation des données complexes. La figure 6 est une proposition de visualisation de la couverture textuelle des structures multi-échelles et de leur interaction, où chaque structure du texte est représentée par un segment horizontal correspondant à la zone de texte couverte, et où la disposition verticale traduit l'enchâssement des structures (Tanguy, 2012 : 127).

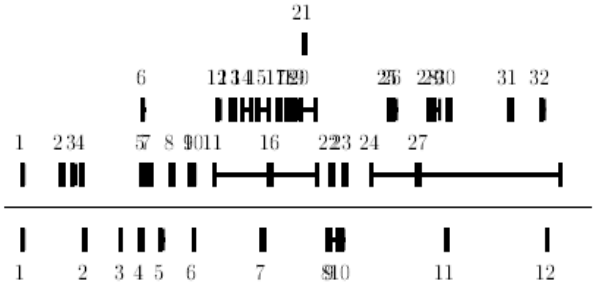


Figure 6 : Positions relatives des structures énumératives et des chaînes topicales dans un texte

Cette schématisation montre les zones couvertes par les structures énumératives (en haut) et les chaînes topicales (en bas), le long de l'axe du texte. « Lorsqu'il y a enchâssement, c'est-à-dire lorsqu'une structure débute à l'intérieur d'une autre, elle est positionnée à un niveau supérieur sur le graphique : le nombre d'étages ainsi obtenus permet de mesurer facilement le niveau d'enchâssement d'un segment de texte. » (Tanguy, 2012 : 126). Dans l'exemple de la figure 6, on peut ainsi observer que relativement peu de texte n'est pas inclus dans une structure énumérative (les structures énumératives couvrent 43 % de la surface textuelle totale du corpus ANNODIS) et que les enchâssements sont assez fréquents avec un niveau maximal de trois structures énumératives enchâssées pour la structure 21 (le niveau d'enchâssement maximal rencontré dans la ressource est de cinq).

L'autre enjeu majeur de ce type de ressource concerne l'étude de la signalisation des différentes structures. L'annotation des relations rhétoriques a ainsi permis de projeter et d'évaluer des lexiques de marqueurs discursifs (Vergez-Couret, 2012). En ce qui concerne la signalisation des chaînes topicales, la ressource ANNODIS permet de rassembler une grande variété de continuités topicales et de les confronter aux travaux théoriques sur les expressions référentielles, dont le rôle est souvent envisagé de façon isolée plutôt que dans la continuité topicale globale. Pour ce qui est des structures énumératives, la complexité de la signalisation est telle qu'il est difficile de l'appréhender dans sa globalité. Elle appelle des analyses qualitatives ciblées (e.g. Ho-Dac *et al.,* 2012 ; Rebeyrolle et Péry-Woodley, 2014 – ce volume) qu'il s'agira de mettre en relation, mais citons tout de même des tentatives de visualisation visant à défricher cette complexité, illustrées ici par la figure 7, qui représente une visualisation en treillis des indices associés par les annotateurs aux structures énumératives (Tanguy, 2012, Ho-Dac et Tanguy, 2013).

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801286

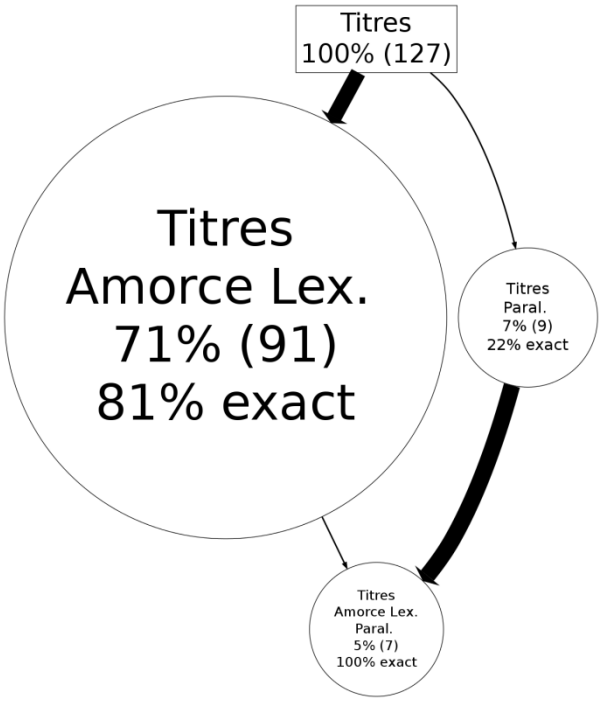


Figure 7 : « Treillis des types d'indices associés aux structures énumératives de type section titrée.

- Les nœuds correspondent à des cooccurrences d'indices, du plus générique au plus spécifique
- Le nombre et la proportion de SE sont indiquées pour chaque configuration
- Le second pourcentage indique le nombre de SE qui n'ont que ces types indices » (Tanguy 2012 : 120)

Le treillis de la Figure 7 indique les types d'indices associés aux énumérations de type 1. Ce type d'énumération est issu de la typologie des structures par niveaux de grain proposée dans Ho-Dac et al (2010). Il caractérise les énumérations qui correspondent à des sections titrées pour lesquelles chaque item est introduit par un titre de section. Ce caractère définitoire se retrouve ici dans l'élément racine du treillis (le rectangle supérieur) : 100 % des structures de type 1 (127 structures) sont associées à un titre de section en introduction d'item (Titres). La suite du treillis indique ensuite que sur ces 127 structures de type 1 :

- 91 (soit 71 %) sont associées à des titres de section accompagnés d'un indice lexical dans l'amorce (Amorce Lex.), par exemple *Il y a plusieurs types de X*.
- 9 (soit 7 %) sont associées à des titres de section accompagnés de parallélismes syntaxiques entre les items (Paral.).
- 7 (soit 5 %) sont associées à des titres de section accompagnés à la fois d'un indice lexical d'amorce et de parallélismes syntaxiques entre les items.

Cette visualisation permet ainsi d'observer que les titres de section sont généralement associés à un indice lexical dans l'amorce pour signaler la présence d'une structure énumérative.

Enfin, les annotations obtenues peuvent également servir de corpus d'apprentissage pour des traitements automatiques. Par exemple, la segmentation en EDUs produite par les annotateurs a permis de construire une implémentation par apprentissage automatique de la segmentation en EDU (Afantenos et al., 2010). L'annotation de segments discursifs de gros grain (chaînes topicales et éléments de structures

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801286

énumératives) peut également servir de référence à des techniques d'apprentissage de techniques de mesures de la cohésion lexicale des textes (cf. Adam, 2012).

5 Conclusion

Cet article présente un état des lieux d'une ressource issue d'une expérience d'annotation discursive qui s'ancre complètement dans les enjeux actuels du domaine de l'étude de l'organisation discursive. Ainsi que nous l'évoquions en citant Marcu (2006) en introduction, l'étude du discours est encore très éclatée, tant sur le plan de l'élaboration de modèles théoriques que sur celui des études empiriques. Le projet ANNODIS est autant une expérimentation à grande échelle que la construction d'une ressource « de référence ». Dans un domaine où ni les catégories ni les relations ne sont stabilisées, la construction d'un modèle d'annotation opérationnalisable, son explicitation dans un manuel d'annotation, la mise en œuvre de l'annotation avec des annotateurs extérieurs au projet sont des exercices éminemment difficiles, et dont on ne peut pas dire à ce stade qu'ils sont terminés. La diffusion d'une ressource annotée est envisagée ici comme la naissance de cette ressource plus que son aboutissement, comme le souligne la liste rapide des travaux relatifs au « devenir » des annotations. La diversité des objets annotés, l'empan des phénomènes discursifs couverts – mis en évidence ici par la confrontation avec la synthèse de Webber *et al.* (2012) –, sont, nous le souhaitons, des atouts pour la réutilisation de la ressource. C'est aussi par le biais de réannotations que des utilisateurs se l'approprieront, et que de nouvelles versions seront amenées à être proposées, touchant des parties spécifiques du corpus ou encore des phénomènes ciblés.

Références bibliographiques

Adam, C. (2012). *Voisinage lexical pour l'analyse du discours*. Thèse de Doctorat. Université de Toulouse.

Afantenos S. D., Asher N., Benamara F., Bras M., Fabre C., Ho-Dac L.-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 organization: the ANNODIS corpus. *LREC 2012*, Istanbul, Turkey, July 2012.

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

Atallah, C., Vergez-Couret, M. et Savreux, F. (2013). Du versant empirique au versant théorique : quand l'analyse des données enrichit la SDRT. Conférence Corpus et Outils en Linguistique, Langues et Parole : Statuts, Usages et Mésusages, Strasbourg, 3-5 juillet 2013.

Biber, D., Connor, U., et Upton, T. (2007). *Discourse on the move: Using corpus analysis to describe discourse structure*. Studies I, corpus Linguistics, 28. John Benjamins Publishing Company : Amsterdam/Philadelphia.

Carlson, L., Marcu, D., et Okurowski, M. E. (2003). Building a discourse-tagged corpus in the framework of Rhetorical Structure Theory. In J. van Kuppevelt et R. Smith (Eds.), *Current Directions in Discourse and Dialogue*. Dordrecht : Kluwer Academic Publishers, pp. 85-109.

Charolles, M. (2002). *La référence et les expressions référentielles en français*. Paris : Orphys.

Colléter, M., Fabre, C., Ho-Dac, L.-M., Péry-Woodley, M.-P., Rebeyrolle, J. et Tanguy, L. (2012). La ressource ANNODIS multi-échelle : guide d'annotation et bonus. *Carnets de grammaire* 20, CLLE-ERSS.

Daneš, F. (1974). Functional sentence perspective and the organisation of text: Different types of thematic progression. In F. Daneš (Ed.), *Papers on functional sentence perspective*. La Hague : Mouton de Gruyter, pp. 106-128.

Grosz, B. et Sidner, C.L. (1986). Attention, intentions, and the structure of discourse. *Computational Linguistics*, 12(3):175–204.

Habert, B. (2005). *Instruments et ressources électroniques pour le français*. Paris : Orphys.

Halliday, M. A. K. (1985). *An Introduction to Functional Grammar*. London : Edward Arnold.

Ho-Dac, L.-M., Péry-Woodley, M.-P. et Tanguy, L. (2010). Anatomie des structures énumératives. *Actes de TALN 2010*. Université de Montréal, Montréal.

Ho-Dac, L.-M., Fabre, C., Péry-Woodley, M.-P., Rebeyrolle, J., et Tanguy, L. (2012). An empirical approach to the signalling of enumerative structures. *Discours* 10.

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801286

Ho-Dac, L-M. et Tanguy, L. (2013) Identification des marqueurs complexes des structures multi-échelles, *Journée d'étude "Les structures énumératives dans le discours"*, CLLE-ERSS, Toulouse, 8 novembre 2012.

Leech, G. (1997). Introducing corpus annotation. In R. Garside, G. Leech et T. McEnery (Eds.), *Corpus annotation: Linguistic information from computer text corpora*. London : Addison Wesley, pp. 1–18.

Luc, C. et Virbel, J. (2001). Le modèle d'architecture textuelle : fondements et expérimentation. *Verbum* 23(1), 103-123.

Mann, W. C., et Thompson, S. A. (1988). Rhetorical Structure Theory: a theory of text organization. In L. Polanyi (Ed.), *The Structure of Discourse*: Norwood, N.J. : Ablex.

Marcu, D. (2006). Automatic Discourse Parsing. *Encyclopedia of Language and Linguistics*. Oxford : Elsevier , pp. 649—654.

Mathet, Y. et Widlöcher, A. (2009). La plate–forme GLOZZ : environnement d'annotation et d'exploration de corpus. *Actes TALN 2009*, Senlis.

Muller, P., Vergez, M., Prevot, L., Asher, N., Benamara, F., Bras, M., Le Draoulec, A. et Vieu, L. (2012). Manuel d'annotation en relations de discours du projet ANNODIS. *Carnets de grammaire* 21, CLLE-ERSS.

Péry-Woodley, M.-P. et Scott, D. (2006). Computational approaches to discourse and document processing. *TAL* 47(2), 7—19.

Péry-Woodley, M.-P., Afantenos, S. D., Ho-Dac, L.-M., Asher, N. (2011). La ressource ANNODIS, un corpus enrichi d'annotations discursives. *TAL 52(3)*, 71-101.

Prasad, R., Miltsakaki, E., Dinesh, N., Lee, A., et Joshi, A. (2006). *Penn Discourse TreeBank 1.0 Annotation Manual*. University of Pennsylvania Institute of Research in Cognitive Science Technical Report Series 3-29-2006.

Rebeyrolle, J. et Péry-Woodley, M-P. (2014). Énumération et structuration discursive. *Actes CMLF 2014*, Berlin.

Swales, J. (1981). *Aspects of Article Introductions*. Aston ESP research reports. Language Studies Unit, University of Aston in Birmingham.

Tadros, A. (1994). Predictive Categories in Expository Texts. In M. Coulthard (Ed.), *Advances in Written Text Analysis*. London-New York : Routledge, pp. 69-82.

Teufel, S. et Moens, M. (2002). Summarizing scientific articles – experiments with relevance and rhetorical status. *Computational Linguistics*, 28(4):409–445.

Vergez-Couret, M. (2010). *Étude en corpus des réalisations linguistiques de la relation d'Élaboration*. Thèse de Doctorat. Université de Toulouse.

Vergez-Couret M. (2012). Relations et marqueurs du discours dans ANNODIS. *Journées d'étude Relations de discours marquées vs. non marquées*, ICAR, Lyon, 29-30 octobre 2012.

Walker, M., Joshi, A., et Prince, E. (Eds.) (1998). *Centering Theory in Discourse* . Oxford : Clarendon Press.

Webber, B., Egg, M., and Kordoni, V. (2012). Discourse structure and language technology. *Natural Language Engineering*, 18(4):437–490.

---

[1] Le corpus brut et la ressource ANNODIS sont diffusés sur le site REDAC : http://redac.univ-tlse2.fr/.

[2] Bien que nous n'ayons pas été directement impliquées dans l'élaboration et l'exploitation des annotations en relations rhétoriques, il nous a semblé intéressant d'envisager ici l'ensemble de l'expérience ANNODIS. Seules les parties concernant les structures multi-échelles comportent des éléments non-publiés par ailleurs, pour les éléments de réflexion concernant les relations rhétoriques, nous faisons référence à des travaux déjà publiés.

[3] Précisons que D. Marcu et B. Webber ont tous deux joué un rôle majeur dans l'élaboration de ressources annotées discursivement, ainsi que dans la réflexion sur ce type d'expérience : voir RST Discourse Treebank (http://www.isi.edu/~marcu/discourse/Corpora.html) et Penn Discourse Treebank (http://www.seas.upenn.edu/~pdtb/).

[4] Notons cependant que la mise en parallèle est approximative dans la mesure où la notion de motif de structuration textuelle est absente de la conception présentée par Webber *et al.* : pour emprunter à la terminologie systémique-fonctionnelle (Halliday, 1985), les intentions envisagées par ces auteurs (*convaincre, remplacer une croyance*…) relèvent principalement de la métafonction interpersonnelle, alors qu'énumérer relève fortement de la métafonction textuelle.

[5] http://atoll.inria.fr/passage

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801286

[6] La thèse de Caroline Attalah (CLLE-ERSS), en cours de finalisation, se base sur l'annotation des relations causales et propose une réannotation de celles-ci dans la ressource ANNODIS.