La ressource ANNODIS, un corpus enrichi d'annotations discursives

Marie-Paule Péry-Woodley, Stergos Afantenos, Lydia-Mai Ho-Dac, Nicholas Asher

La ressource ANNODIS, un corpus enrichi d'annotations discursives

Marie-Paule Péry-Woodley* – Stergos D. Afantenos – Lydia-Mai Ho-Dac* – Nicholas Asher****

** CLLE-ERSS (UMR 5263 et Université de Toulouse) Université Toulouse 2-Le Mirail 5, allées Antonio-Machado – 31058 Toulouse Cedex 9, France.*

pery,hodac@univ-tlse2.fr

*** IRIT (UMR 5505 et Université de Toulouse) Université Paul Sabatier 118, route de Narbonne – 31062 Toulouse Cedex 9, France.*

stergos.afantenos,Nicholas.Asher@irit.fr

RÉSUMÉ. *Cet article décrit la ressource ANNODIS, issue d'un projet financé par l'ANR, corpus de français écrit enrichi à différents niveaux, dont un niveau d'annotation manuelle de structures discursives. Une originalité de la ressource est de proposer un corpus diversifié (plusieurs types de textes sont représentés) et deux annotations fondées sur des approches distinctes de la structuration des discours. La description de la ressource – objets annotés, textes composant le corpus – s'accompagne de la présentation des ancrages théoriques sous-jacents aux modèles d'annotation, et des choix méthodologiques qui ont guidé les diverses phases de préparation et d'annotation du corpus. Nous formulons les enjeux d'une telle ressource pour la linguistique et le TAL, et présentons les premières exploitations.*

ABSTRACT. *This paper describes the ANNODIS ressource, a corpus of written French enriched with several markups, including a manual annotation of discourse structures. The resource is original in that it offers a diversified corpus representing several text types, and two annotations based on different approaches to discourse organisation. As well as a description of the ressource – annotated objects, composition of the corpus – the paper presents the theoretical underpinnings of the annotation models and the methodological choices underlying corpus preparation and annotation. It also sketches the potential contribution of such a resource for linguistics and NLP, and describes initial results of its exploitation.*

MOTS-CLÉS : *ressources linguistiques, annotation de corpus, structures de discours.*

KEYWORDS: *language resources, corpus annotation, discourse structures.*

1. Introduction

Nous décrivons ici la ressource ANNODIS [1], un corpus diversifié de français écrit enrichi à différents niveaux, dont un niveau d'annotation manuelle de structures discursives. Cette ressource, issue d'un travail expérimental d'annotation, a d'emblée été conçue pour être mise à la disposition des chercheurs en linguistique du discours et en traitement automatique des langues.

En comparaison avec les corpus existants annotés discursivement (en particulier le Penn Discourse TreeBank pour l'anglais), le corpus ANNODIS présente deux caractéristiques propres : 1) il est composé de textes diversifiés (en genre, longueur, et type d'organisation discursive) ; 2) il est le résultat de deux types d'annotation manuelle, correspondant à deux approches de la structuration des discours. La première vise à construire la structure complète d'un discours, dans une démarche ascendante, à partir d'unités élémentaires reliées par des relations rhétoriques. La seconde envisage les textes dans leur dimension de document et, s'intéressant spécifiquement à leur mise en texte (y compris dans ses aspects visuels), vise l'annotation sélective de structures discursives multi-échelles. Les deux approches s'articulent *via* l'hypothèse d'une influence de la perception de structures de haut niveau sur l'interprétation des relations entre unités élémentaires (processus descendant).

Les prétraitements appliqués à une large partie du corpus constituent une double innovation, d'une part à travers l'encodage XML (TEI-P5) non seulement des métadonnées mais également de la structure visuelle des textes, d'autre part grâce au prémarquage automatique de traits associés à la signalisation du discours.

La section 2 fournit une description synthétique de la ressource : objets annotés, corpus, comparaison avec l'existant. Les sections 3 et 4 sont consacrées à la présentation détaillée des fondements théoriques, des procédures d'annotation, et des premières exploitations pour les deux annotations : relations rhétoriques et structures multi-échelles. Nous évoquons ensuite la partie de la ressource et les travaux à l'intersection des deux approches (5), avant de poser un premier bilan (6).

2. Description de la ressource

La ressource ANNODIS est le résultat d'une double annotation discursive, fondée sur deux approches distinctes de la structuration et de l'interprétation des discours. Cette dualité théorique et méthodologique oblige salutairement à rendre explicites des choix – objets à annoter, textes du corpus, prétraitements – qui au sein d'une même

1. Projet financé par l'ANR (appel Corpus 2007). Partenaires : CLLE-ERSS (Toulouse), IRIT (Toulouse), GREYC (Caen). Voir `http://w3.erss.univ-tlse2.fr/annodis` et (Péry-Woodley *et al.*, 2009) pour une présentation des différentes facettes du projet. Seuls les corpus construits et annotés dans le cadre du projet sont décrits ici. Outre les quatre auteurs, ont contribué à cette entreprise : F. Benamara, M. Bras, C. Fabre, A. Le Draoulec, P. Muller, L. Prévot, J. Rebeyrolle, L. Tanguy, M. Vergez-Couret, L. Vieu.

communauté paraîtraient aller de soi. Elle est par ailleurs le reflet de l'état des connaissances en linguistique du discours. En effet, si nous reprenons à notre compte la définition de l'annotation comme « consist[ant] à ajouter de l'information (une interprétation stabilisée) aux sons, aux caractères, aux gestes » (Habert, 2005, p. 148), force est de constater que dans le domaine du discours aussi bien les catégories que les associations entre formes et catégories sont peu stabilisées. L'annotation systématique de corpus implique à la fois une stabilisation forcée pour la rédaction du guide d'annotation, et une mise à l'épreuve par la confrontation aux textes, mise à l'épreuve qui fera évoluer les modèles. L'annotation selon deux modèles participe fortement du caractère expérimental de l'entreprise. Dans cette section, nous présentons brièvement les objets annotés (2.1) et la constitution du corpus (2.2) avant de fournir une vue d'ensemble de la ressource (2.3). La dernière sous-section situe la ressource ANNODIS parmi les ressources similaires et en présente les enjeux pour le TAL et la linguistique du discours.

2.1. *Objets annotés*

2.1.1. *Relations rhétoriques*

L'annotation des relations rhétoriques 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 discursives reliées par des relations de cohérence (dites rhétoriques). Une structure discursive peut donc être représentée par un graphe, où les nœuds sont des unités discursives et où les arcs étiquetés représentent les relations rhétoriques entre ces unités.

[Principes de la sélection naturelle.]_1 [La théorie de la sélection naturelle [telle qu'elle a été initialement décrite par Charles Darwin,]_2 repose sur trois principes :]_3 [1. le principe de variation]_4 [2. le principe d'adaptation]_5 [3. le principe d'hérédité]_6

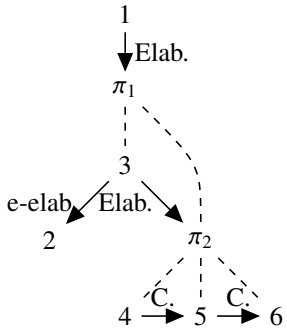

Figure 1. *Exemple de graphe discursif. Les nœuds correspondent aux unités discursives : les EDU représentées par leur numérotation et les CDU par un nœud étiqueté π_n. Les arêtes avec flèches représentent les relations rhétoriques, les arêtes en pointillé sans flèches représentent l'inclusion d'EDU dans un CDU. Elab. = Élaboration, e-elab = Élaboration d'entité, C. = Continuation.*

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 1 est un exemple de structure discursive comportant des EDU et des CDU. Les nœuds 1 à 6 sont des EDU, tandis que les nœuds π_1 et π_2 représentent des CDU. Les arcs en pointillé joignent les CDU à leurs sous-constituants. Les chiffres dénotant les EDU sont des étiquettes pour des portions de texte considérées par l'annotation comme étant des atomes de sens dans la construction de la structure discursive.

2.1.2. *Structures multi-échelles*

Deux structures multi-échelles ont été annotées : les structures énumératives et les chaînes topicales. Leur annotation (détail en 4.2) implique à la fois la délimitation de segments d'au moins deux phrases et l'indication des traits de surface signalant ces structures (les traits annotés acquièrent alors le statut d'« **indice** »). Une structure énumérative (**SE** en abrégé) est un segment de texte caractérisé par une structuration interne mettant en jeu les sous-segments suivants :

– une amorce (optionnelle) : segment qui introduit l'énumération ;

– plusieurs items : segments qui constituent l'énumération (au moins deux items doivent être identifiés pour qu'une structure soit annotée) ;

– une clôture (optionnelle) : segment qui résume ou clôt l'énumération.

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 préverbale (*i.e.* potentiellement topicale) dans plusieurs des phrases du segment.

La figure 2 montre une portion de texte dans laquelle on trouve deux chaînes topicales autour du référent *le régime* situées au début (CT1) et à la fin (CT2) d'une structure énumérative (SE). Cette structure énumérative comporte une amorce (*En revanche, le régime a patronné trois formations importantes.*), trois items et une clôture. Les lexèmes encadrés (*formations* dans l'amorce et *forces armées* dans la clôture) spécifient le critère de coénumérabilité ou principe d'énumération, c'est-à-dire qu'ils précisent que les items 1 à 3 énumèrent trois types de *formations* ou *forces armées*. Ils réalisent ce que nous appelons l'**énumérathème**. Ces deux structures sont alignées sur un même début de paragraphe.
Étant donné que les annotateurs ont annoté à la fois les structures et les traits qui les signalent, l'annotation fournit des indices de continuité topicale (*le régime*, *il*, *il*, *il*, *Le régime*, *Il* ; en italique dans les CT de l'exemple 2), et, dans le cas des structures énumératives, des indices plus diversifiés (en gras dans l'exemple), liés aux phénomènes de prospection en amorce (*trois formations*), d'encapsulation en clôture (*En somme,*) et de discontinuité dans les différents items (*aussi* et *Enfin,*).
Une propriété essentielle des structures énumératives, et dans une moindre mesure, des chaînes topicales, réside dans leur caractère « multi-échelle », c'est-à-dire le fait que ces structures peuvent apparaître à des niveaux de granularité divers, y compris, dans

En revanche, *le régime* a patronné **trois** formations importantes.	CT1	SE	AMORCE
Bien qu'*il* ait réduit de moitié les effectifs de la Garde républicaine, passée de 150 000 à 70 000 hommes, *il* a veillé à en reconstituer les précieuses unités mécanisées et blindées. Pour ce faire *il* a eu recours, outre quelques importations illégales, à la cannibalisation des matériels rescapés du pilonnage, souvent au détriment de l'armée.			ITEM 1
Le régime s'est **aussi** détourné de son aviation au profit d'un Corps aérien plus opérationnel. *Il* en a consolidé les escadrons habitués à opérer en coordination étroite avec la Garde républicaine.			ITEM 2
L'importation de pièces de rechange s'est d'ailleurs révélée plus facile pour les hélicoptères, qui bénéficient d'un double statut civil et militaire.			
Enfin, les incursions quasi quotidiennes des avions anglo-saxons dans les zones d'exclusion aérienne et les « frappes » régulières de missiles de croisière ont stimulé l'intérêt porté par Saddam Hussein à la Défense aérienne, rénovée et amadouée par des privilèges semblables à ceux dont bénéficie la Garde républicaine. On ne saurait souligner assez que c'est là la principale disposition militaire classique prise par l'Irak contre un adversaire étranger.			ITEM 3
En somme, *le régime* a remodelé et réorienté ses forces armées pour aller vers un système plus sûr et plus compact, au caractère répressif et défensif.	CT2		CLÔTURE
Dans cette configuration, *il* ne représente plus guère, en dépit des accusations des Etats-Unis, une menace pour ses voisins.			

Figure 2. *Portion de texte contenant deux chaînes topicales et une structure énumérative. Les lignes pleines délimitent les structures annotées ainsi que les sauts de paragraphes (dans la colonne de gauche). Les indices annotés apparaissent en italique pour les indices de chaîne topicale, et en gras pour les indices de structure énumérative. Les deux lexèmes encadrés correspondent à des énumérathèmes. CT = chaîne topicale, SE = structure énumérative.*

le cas des structures énumératives, à de très hauts niveaux (plusieurs sous-sections). Notre deuxième exemple (figure 3) illustre cette propriété, ainsi que l'aptitude des structures énumératives à s'enchâsser : on y voit (fortement coupées pour raison de place) une SE enchâssante de haut niveau (SE1) qui se déploie sur l'ensemble d'une section, chaque item constituant une sous-section. Son amorce est le titre de section ; deux énumérathèmes ont été identifiés qui précisent que les items 1 à 4 sont énumérés en tant que *fondements* ou *approches* d'une même question. L'item 4 est « déplié » pour dévoiler une SE enchâssée (SE2) dont les items, introduits par une amorce com-

portant l'énumérathème *aspects* [2], sont quant à eux signalés par des puces. Trois types de signalisation sont donc à l'œuvre dans nos exemples à des niveaux de granularité différents : en figure 3, découpage en sections et sous-sections pour la structure SE1 et liste formatée pour la structure enchâssée SE2 (signalée par la mise en forme particulière : puces, espaces verticaux et horizontaux) ; en figure 2, marques lexicales dans une structure intraparagraphe.

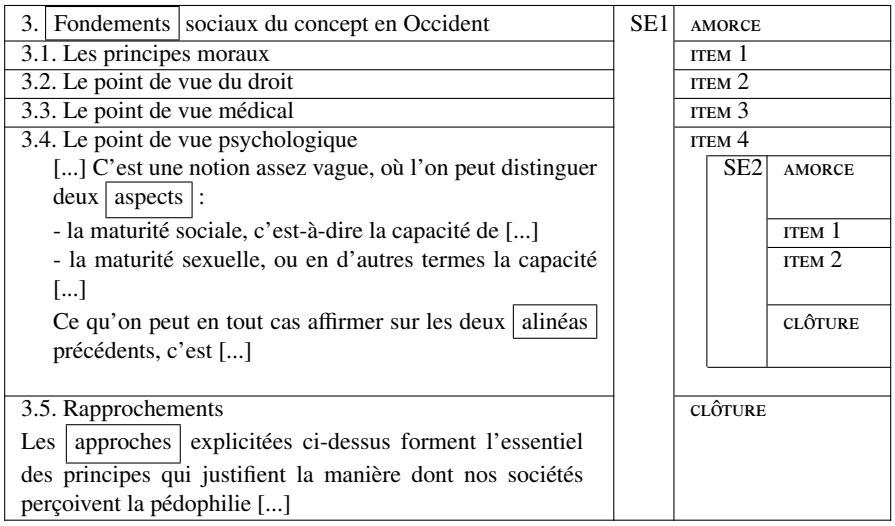

Figure 3. *Structures énumératives enchâssées. Les lignes horizontales de la colonne de gauche représentent les sauts de section. SE1 = structure énumérative enchâssante ; SE2 = structure énumérative enchâssée.*

2.2. *Corpus*

La ressource ANNODIS est composée de brèves et d'articles à visée informative représentant une certaine diversité en terme de genre textuel, de type dominant et de structure de document [3]. Notre objectif a été d'intégrer d'entrée de jeu l'hypothèse de la variation dans les réalisations discursives en fonction de variations extralinguistiques, ce qui la distingue des corpus annotés en relations et structures de discours existants (pour l'anglais) composés uniquement de textes journalistiques brefs

2. Un deuxième énumérathème est exprimé en clôture : *alinéas*, exemple de basculement d'une perspective « conceptuelle » à une perspective « textuelle » dans la même SE (cf. 4.3.2).

3. *« Informally, document structure describes the organization of a document into graphical constituents like sections, paragraphs, sentences, bulleted lists, and figures ; it also covers some features within sentences, including quotation and emphasis. »* (Power *et al.*, 2003, p. 214)

ou de dépêches [4]. La diversification des textes du corpus n'a donc pas été envisagée dans l'optique de fournir un corpus de référence des genres écrits du français, mais dans l'idée de constituer des données autorisant des comparaisons intergenres. Typiquement, on observe un faible niveau de structuration pour les brèves issues de la presse écrite, et plus généralement les récits, et un fort niveau de structuration pour les articles de recherche ou ceux de Wikipédia, et plus généralement les textes expositifs longs.
Le tableau suivant dresse l'inventaire des sous-corpus de la ressource ANNODIS.

Nom	Source (S), genre (G), type dominant (T)	Structure de document	Nombre de	
NEWS	S = *Est Républicain* G = brèves T = narratif	faible	articles mots mots par texte	39 10 000 250
WIK1	S = *extraits Wikipédia* G = art. encyclopédique T = expositif	faible	extraits mots mots par texte	30 11 000 412
WIK2	S = *articles Wikipédia entiers* G = art. encyclopédique T = expositif	forte	articles mots mots par texte	30 231 000 7 700
LING	S = *CMLF* [5] *: colloque de linguistique* G = articles de recherche T = expositif	moyenne	articles mots mots par texte	25 169 000 6 760
GEOP	S = *IFRI* [6]*, institut de géopolitique* G = rapports et articles T = argumentatif	moyenne	articles mots mots par texte	32 266 000 8 325
TOTAL : 156 textes ou extraits, 687 000 mots				

Tableau 1. *Constitution du corpus ANNODIS. S = Source des textes, G = Genre textuel, T = Type dominant, niveau de structure de document, nombre de textes, nombre de mots et longueur moyenne des textes.*

Outre les critères exposés plus haut, le choix des textes du corpus a également été déterminé par les possibilités de diffusion des corpus annotés. Les textes issus de Wikipédia sont libres sous licence Creative Commons. Ceux issus de l'*Est Républicain* sont diffusables dans les mêmes conditions que dans le cadre du CNRTL. Pour les articles de recherche et les rapports, des conventions de distribution ont été mises en place afin de diffuser les corpus annotés sous licence Creative Commons.

4. Le Penn Treebank, sur lequel se basent à la fois le Penn Discourse Treebank et le RST Corpus, est exclusivement composé d'articles courts issus du *Wall Street Journal* (longueur moyenne 458 mots par texte pour la sélection du RST corpus).
5. Les articles sélectionnés proviennent du premier Congrès Mondial de Linguistique Française (CMLF-08)
6. L'IFRI (Institut français des relations internationales) est un *think-tank* ou « laboratoire d'idées » français consacré à l'analyse des questions internationales (`http://www.ifri.org`)

2.3. *Vue d'ensemble de la ressource*

La ressource ANNODIS, déposée sur le site REDAC (Ressources Développées à CLLE-ERSS `http://redac.univ-tlse2.fr/`)[7] sous licence « Creative Commons » BY-NC-SA 3.0 (paternité, pas d'utilisation commerciale, partage des conditions initiales à l'identique), regroupera les cinq sous-corpus dans leur format brut (avant annotation) et sous divers formats d'annotation débarquée (détaillés ci-dessous), ainsi que les guides d'annotation. Ces guides, dont la rédaction a été l'occasion d'un important travail d'explicitation des objets à annoter, fournissent le détail des modèles d'annotation et des balises délimitant les annotations produites ainsi que les procédures d'annotation réalisées *via* l'interface GLOZZ[8].
Le tableau 2 indique le nombre d'objets annotés dans chacun des sous-corpus présentés dans la section précédente.

Corpus	Objets annotés						
	EDU	Rel.	CDU	SE	indices(SE)	CT	indices(CT)
NEWS	1 159	1 203	510				
WIK1	1 949	2 034	829				
WIK2	53	65	38	401	2 221	265	1 837
LING	12	14	9	295	1 185	89	450
GEOP	15	19	9	249	1 071	218	1 034
ANNODIS	3 188	3 355	1 395	945	4 477	572	3 321

Tableau 2. *Relations rhétoriques et structures multi-échelles dans ANNODIS. EDU = Unité de discours élémentaire ; Rel. = Relation rhétoriques ; CDU = Unité de discours complexe ; SE = Structure Énumérative ; CT = Chaîne Topicale ; indices = traits de surface signalant une structure.*

Les disparités qui apparaissent entre les sous-corpus quant au type et au nombre d'objets annotés sont le reflet de la dualité théorique et méthodologique qui sous-tend le projet ANNODIS (double annotation discursive, fondée sur deux approches distinctes de la structuration et de l'interprétation des discours). De fait, les deux approches n'ont pas pu être appliquées à tous les textes du corpus, chaque approche définissant des besoins et des limites parfois incompatibles dans le cadre d'une campagne d'annotation. Ainsi l'annotation de structures multi-échelles est pertinente pour des textes longs non narratifs où différents modes de structuration sont susceptibles d'être sollicités et signalés, mais elle n'a pas grand intérêt dans des brèves de quelques phrases. Parallèlement, l'annotation complète des relations rhétoriques dans des textes longs n'était pas envisageable à l'échelle du projet (à moins de se limiter à quelques textes, ce qui poserait un problème de couverture). Par ailleurs le choix de textes très courts

7. La ressource sera également accessible à partir du site du CNRTL grâce à un lien renvoyant sur REDAC.
8. La plate-forme d'annotation GLOZZ (`www.glozz.org/`, de Mathet et Widlöcher (2009)) est le résultat du versant logiciel du projet ANNODIS.

issus de la presse pour l'annotation des relations rhétoriques autorise des comparaisons avec des ressources similaires pour d'autres langues (*e.g.* RST Tree Bank).
Les objets annotés se répartissent donc dans les sous-corpus de la manière suivante :

– les brèves (NEWS) et les extraits de Wikipédia (WIK1) sont entièrement et uniquement annotés au niveau des relations rhétoriques ;

– les articles entiers (WIK2, LING et GEOP) sont entièrement annotés au niveau des structures multi-échelles et partiellement au niveau des relations rhétoriques (voir tableau 3).

Sous-corpus	Nombre d'extraits annotés en structures multi-échelles et relations rhétoriques
LING	3 extraits de 3 articles de linguistique
WIK2	8 extraits de 8 articles de Wikipédia 4 extraits d'un même article de Wikipédia
GEOP	3 extraits d'un même rapport de géopolitique
ANNODIS	18 extraits issus de 13 articles différents

Tableau 3. *Intersection des annotations dans la ressource ANNODIS*

Par ailleurs, de nombreux textes ou extraits ont été multi-annotés par des **annotateurs naïfs et/ou experts** : plus de la moitié des textes annotés en relations rhétoriques sont annotés par deux annotateurs naïfs [9] puis arbitrés par un expert [10] (voir section 3.2). Du côté des structures multi-échelles, une dizaine de textes sont annotés par trois naïfs puis arbitrés par un expert, les autres textes étant mono-annotés (voir section 4.2.2). Toutes ces versions sont mises à disposition.
Concernant les **formats de diffusion**, plusieurs formats sont disponibles :

– toutes les annotations sont diffusées sous un format XML défini par l'interface d'annotation Glozz (Mathet et Widlöcher, 2009), ce format distingue le texte brut d'une part (fichier texte) et les annotations débarquées d'autre part (fichier XML) et permet de visualiser les annotations dans les textes, *via* Glozz (alors utilisée en tant qu'interface d'exploitation) ;

– tous les textes collectés pour la campagne sont disponibles dans un format brut, précédant l'annotation : PDF ou HTML d'origine et XML normé selon la TEIP5 (méta-données et mise en forme matérielle) ;

– les annotations en relations de discours sont également disponibles au format texte (tel que présenté dans les figures 4 et 5) et sous un format XML concaténant les deux fichiers Glozz (le texte et les annotations débarquées).

Finalement, une interface d'interrogation (en cours de construction) permettra d'explorer les annotations.

9. Étudiants de linguistique de niveau licence 3, sans connaissances préalables des fondements théoriques du modèle d'annotation.
10. Chercheur participant au projet ANNODIS.

2.4. *Contexte et enjeux d'une annotation discursive*

Dans sa synthèse sur l'analyse discursive automatique pour l'*Encyclopedia of Language and Linguistics*, Marcu décrit ce domaine comme un jeune champ de recherche qui se trouve dans une position un peu difficile, en dépit du besoin très réel de traitements de niveau discursif dans diverses applications du traitement automatique des langues (Marcu, 2006). Cette difficulté est en grande partie attribuable, selon lui, au manque de maturité de la linguistique du texte et du discours : diversité des modèles, éclatement des travaux descriptifs. Du côté des travaux en linguistique de discours, Péry-Woodley (2005), citant en particulier Biber (1988) et Bestgen *et al.* (2003), propose le constat suivant : « Pour résumer la situation telle que la décrivent les auteurs cités dans cette section, les études sur le discours se caractérisent actuellement par une approche qualitative, sur des données de faible volume, avec des méthodes manuelles et donc subjectives ("*analyst-dependent*"), ce qui fait obstacle à leur reproductibilité – et partant à leur validation –, et à la généralisation de leurs résultats » (Péry-Woodley, 2005, p. 185). Dans ce contexte, le développement et la mise à disposition de corpus annotés sur le plan discursif constituent un enjeu majeur. Si des corpus de textes annotés au niveau discursif existent pour l'anglais – Penn Discourse Treebank (PDTB) [11], RST Discourse Treebank [12], GraphBank (Wolf et Gibson, 2006) – et qu'il s'en développe pour d'autres langues [13], il n'en existait pas pour le français. La mise à disposition d'un corpus diversifié encodé selon les standards actuels et enrichi de multiples annotations devrait favoriser une réflexion collective plus focalisée, une meilleure prise en compte de l'exigence de reproductibilité, et surtout l'acheminement vers un processus cumulatif, en d'autres termes la constitution d'une communauté de chercheurs. En amont de la mise à disposition du corpus annoté, un programme d'annotation systématique comme celui d'ANNODIS est en lui-même une forme d'expérimentation en linguistique. Les enjeux de cette entreprise pour l'élaboration théorique sont considérables : tout d'abord à travers le travail d'opérationnalisation des modèles qu'il exige pour la rédaction des guides et la préparation du protocole d'annotation, ensuite par la mise à l'épreuve à la fois pointilleuse et à grande échelle des modèles choisis. Enfin, le choix que nous avons fait d'une approche duelle place d'emblée la confrontation des modèles au cœur de notre travail [14].

Pour le traitement automatique des langues et ses applications, les enjeux d'une meilleure compréhension et modélisation des fonctionnements discursifs sont formulés par Marcu : « *From a natural language engineering perspective, the need for*

11. `http://www.seas.upenn.edu/~pdtb/`

12. Corpus annoté selon la Rhetorical Structure Theory : `http://www.isi.edu/~marcu/discourse/Corpora.html`

13. Voir par exemple l'annotation des relations de discours et de la coréférence dans le Prague Dependency Treebank : `http://ufal.mff.cuni.cz/pdt2.0/`

14. Un troisième programme d'annotation, portant sur l'organisation thématique et rhétorique des textes, a également été réalisé durant le projet. Ce programme ne s'appliquant pas au corpus ANNODIS mais à un ensemble d'articles du journal *Le Monde*, nous ne l'incluons pas dans le présent article consacré à la ressource ANNODIS. Une présentation en est fournie dans (Labadié *et al.*, 2010).

text-level processing systems is uncontroversial : because sentence-level processing modules (syntactic and semantic parsers, named-entity recognizers, language translators and generators, etc.) operate at sentential level, they are not able to make text-level inferences and/or produce outputs that are text-level coherent/consistent ». (Marcu, 2006). Le développement du PDTB a d'emblée été conçu dans la perspective des applications : « *An increasing interest in moving human language technology beyond the level of the sentence in text summarization, question answering, and natural language generation (NLG) inter alia has recently led to the development of several resources that are richly annotated at the discourse level* » (Prasad *et al.*, 2006).
La ressource ANNODIS se situe elle aussi sur ces différents plans, tout en se démarquant des corpus existants annotés sur le plan discursif d'abord par sa focalisation sur les structures, ensuite par son annotation selon deux modèles, que nous allons maintenant présenter dans le détail.

3. Annotation des relations rhétoriques

3.1. *Fondements théoriques*

Notre annotation de la structure rhétorique se base sur la *Segmented Discourse Representation Theory* (Asher, 1993 ; Asher et Lascarides, 2003). Cette théorie se fonde sur la sémantique dynamique (Kamp et Reyle, 1993 ; Groenendijk et Stokhof, 1991) et fournit pour un texte une représentation (*Segmented Discourse Representation Structure* ou SDRS) en forme de modèle du premier ordre $\langle D, F \rangle$, où D est un ensemble de constituants discursifs et F associe à chaque constituant discursif une formule logique. Aux unités élémentaires, les EDU, sont associées des formules de sémantique dynamique standard, tandis qu'aux constituants d'ordre supérieur, ou CDU, sont associées des formules qui relient les constituants dans S avec une ou plusieurs relations rhétoriques. On peut aussi décrire chaque SDRS comme un graphe étiqueté où (1) chaque nœud signifie un constituant (CDU ou EDU) du texte ; (2) les arcs dirigés et étiquetés expriment des relations rhétoriques entre les constituants, l'étiquette servant à identifier la relation, *e.g. Explication, Narration* ; (3) les arcs dirigés et non étiquetés joignent un constituant complexe à ses sous-constituants.
Il y a deux sortes d'arcs étiquetés, des arcs *coordonnants* et des arcs *subordonnants*, ce qui donne aux graphes une structure bidimensionnelle. Les arcs non étiquetés introduisent une récursivité dans la structure, ce qui permet à la SDRT de décrire des structures discursives à toutes les échelles. En fait, une CDU est une SDRS à elle seule. La sémantique d'une SDRS est donnée par ses conditions de vérité, ou plutôt, pour la sémantique dynamique, par sa capacité de déterminer des continuations possibles. Pour les EDU, la sémantique est donnée par les clauses de la sémantique dynamique, tandis que la sémantique d'une CDU repose sur la sémantique de ses sous-constituants. Pour chaque relation rhétorique, la SDRT se sert des outils en sémantique utilisés pour décrire soit la modalité soit la structure événementielle d'un texte. Par exemple, pour les relations Elaboration ou Frame, nous exploitons la contrainte que si nous avons Elab(a, b) ou Frame(a, b), alors l'éventualité décrite par a inclut l'éventualité décrite

par *b* et que le contenu de *b* fait partie du contenu de *a* (Asher, 1993).
La SDRT précise la portée des relations rhétoriques et permet des attachements à longue distance. Mais la théorie incorpore une contrainte sur l'attachement, la **contrainte de la frontière droite (*right frontier constraint*)**, qui restreint les points d'attachement possibles à longue distance :

Definition 1 (*right frontier*) Soit S un SDRS. Nous définissons d'abord $\prec_S$ sur les nœuds de S. Soient α and δ deux nœuds de S ; $\alpha \prec_S \delta$ ssi : (1) $\alpha \in \delta$; (2) il y un arc de δ à α dans S étiqueté relation **subordonnante** R. La frontière droite $RF(S)$ de S contient last_S, l'EDU qui a été ajoutée en dernier à S, et tout nœud δ tel que $\text{last}_S \prec_S^* \delta$. [15].

3.2. *Procédure d'annotation des relations rhétoriques*

L'annotation des relations rhétoriques se divise en deux phases. [16] Au cours de la première phase, trois annotateurs naïfs ont été entraînés pendant une semaine, après quoi nous leur avons fourni le guide d'annotation. Chaque texte de notre corpus a été annoté par deux annotateurs naïfs. Pendant la deuxième phase, les chercheurs impliqués dans le projet ont repris ces annotations et les ont corrigées de façon à constituer un corpus « expert ».
Afin d'annoter les relations rhétoriques il nous a fallu d'abord annoter les unités discursives qui servent comme points d'attachement pour ces relations. Nous avons donc procédé ainsi :

– segmentation exhaustive en EDU (*Elementary Discourse Units*) : ceci constitue un pavage complet du texte à annoter ;

– regroupement d'EDU en CDU (*Complex Discourse Units*) ;

– attachements entre DU (*Discourse Units*, c'est-à-dire à la fois les EDU et les CDU) et typage de la ou des relations rhétoriques qui relient une DU à d'autres DU.

3.2.1. *Segmentation en EDU*

Le guide d'annotation définit une EDU de la façon suivante : « La première étape dans l'annotation des relations de discours est de segmenter un texte en segments, appelés EDU – *Elementary Discourse Units*, qui doivent correspondre aux unités à mettre en relation avec des relations rhétoriques. L'EDU prototypique est une proposition indépendante. En général, une EDU correspond à la description d'un événement ou d'un état unique. ». Le texte a ainsi été segmenté en EDU sur la base de cette définition et d'un ensemble de règles de délimitation précises (cas des relatives, des incises,

15. $\prec_S^*$ dénote la clôture transitive de $\prec_S$
16. À ces deux phases propres au projet ANNODIS, nous pourrions également ajouter une autre phase préalable dont le but était la construction du guide qui a accompagné les annotateurs de la première phase. Pendant cette phase préalable deux annotateurs de niveau master 2 ont annoté conjointement 47 textes, en discutant des difficultés rencontrées pour chaque relation avec les chercheurs impliqués dans le projet. Ces discussions ont contribué à la construction du guide.

des éléments cadratifs, etc.). Selon le guide, les EDU ne peuvent pas se chevaucher ; elles peuvent en revanche s'emboîter, mais seulement au milieu d'un segment ; dans tous les autres cas les EDU doivent être juxtaposées.
Ainsi que nous l'avons indiqué ci-dessus, chaque texte a été annoté simultanément par deux annotateurs naïfs. Avant de commencer l'annotation des relations, les deux annotateurs ont segmenté ensemble chaque texte dont ils partageaient l'annotation afin de stabiliser une segmentation commune. La figure 4 montre un exemple de segmentation stabilisée issue d'un extrait de Wikipédia – WIK2 [17]. [18]

[Principes de la sélection naturelle.]_1 [La théorie de la sélection naturelle [telle qu'elle a été initialement décrite par Charles Darwin,]_2 repose sur trois principes :]_3 [1. le principe de variation]_4 [2. le principe d'adaptation]_5 [3. le principe d'hérédité]_6
[Principe 1 :]_7 [Les individus diffèrent les uns des autres.]_8 [En général, dans une population d'individus d'une même espèce, il existe des différences plus ou moins importantes entre ces individus.]_9 [En biologie,]_10 [on appelle caractère, tout ce qui est visible et peut varier d'un individu à l'autre.]_11 [On dit qu'il existe plusieurs traits pour un même caractère.]_12 [Par exemple, [chez l'être humain,]_13 la couleur de la peau, la couleur des yeux sont des caractères pour lesquels il existe de multiples variations ou traits.]_14 [La variation d'un caractère chez un individu donné constitue son phénotype.]_15 [C'est là, la première condition pour qu'il y ait sélection naturelle :]_16 [au sein d'une population,]_17 [certains caractères doivent présenter des variations,]_18 [c'est le principe de variation.]_19
[Principe 2 :]_20 ...[xxxx]_57
[Principe 3 :]_58 ... [xxxx]_70

Figure 4. *Exemple de segmentation en EDU – Elementary Discourse Units. Le crochet ouvrant représente le début d'une EDU, le crochet fermant la fin d'une EDU et les numéros qui les accompagnent représentent la numérotation des EDU.*

3.2.2. *Construction des CDU*

Pour cette étape d'annotation les annotateurs naïfs avaient pour consigne de ne construire des CDU que si celles-ci semblaient refléter une partie de la structure discursive. Le guide décrit les CDU comme des segments qui constituent une unité avec une cohérence forte que l'on peut mettre en relation de façon globale avec un autre segment, en précisant qu'une CDU peut être de n'importe quelle taille, et peut comporter des segments éloignés dans le texte. Ce dernier point permet la construction de composants pas nécessairement reliés syntaxiquement. De plus, les CDU peuvent être liées entre elles ou avec d'autres EDU.
Les nombreuses CDU construites par les annotateurs naïfs et, en majorité, reprises pas les annotateurs experts (voir le tableau 5) sont le signe de principes d'organisation à la fois riches et compliqués. La figure 5 montre les relations annotées et les CDU construites dans l'exemple de la figure 4. Par exemple, elaboration(9/[10-17]) indique

17. Article sur la sélection naturelle, version du 10 août 2009, url : http ://fr.wikipedia.org/w/index.php ?title=S %C3 %A9lection_naturelle&oldid=43733592
18. Il faut noter que les annotateurs pouvaient corriger d'éventuelles erreurs de segmentation ultérieurement en utilisant la pseudo-relation « fusion » qui fusionne deux segments.

une relation d'élaboration entre l'EDU 9 et la CDU qui s'étend sur les EDU 10 à 17. En revanche, frame([7,19]/[8-17]) signifie que la CDU composée par les EDU 7 et 19 sont en relation de « cadre » avec la CDU s'étendant sur les EDU 8 à 17.

3.2.3. *Rattachement et typage*

Pour la troisième étape comprenant les tâches conjointes de rattachement et de typage, 17 relations rhétoriques ont été utilisées pour l'attachement. Une description sommaire de ces relations est donnée dans le tableau 4. Les annotateurs ont été avertis que les relations peuvent être explicitement indiquées par un marqueur du discours, mais elles peuvent également être implicites, dans le sens où un lien entre deux segment existe bien qu'aucun marqueur ne soit présent.
Concernant l'attachement, nous avons incité les annotateurs à considérer cette procédure comme une tâche incrémentale en considérant les segments élémentaires les uns après les autres dans l'ordre de lecture du texte. Nous avons explicitement mentionné qu'un segment peut s'attacher au sein de la même phrase, du même paragraphe ou dans un contexte plus large. Nous n'avons donc posé aucune restriction concernant la distance d'attachement. [19] Nous avons mentionné que lorsqu'il s'agit d'une proposition subordonnée ou en apposition, elle s'attache systématiquement à la principale, indépendamment du fait que celle-ci se trouve vers l'avant ou l'arrière dans le texte.
Un exemple d'attachement et typage se trouve dans la figure 5 et le graphe correspondant est fourni dans la figure 6.

elaboration(1/[2,3,4-80])	frame([7,19]/[8-17])	continuation(14/15)
e-elab(3/2)	elaboration(8/[9-17])	continuation(15/16)
elaboration(3/[7-70])	elaboration(9/[10-17])	elaboration(16/17)
elaboration([4-6]/[7-70])	frame(10/[11,12])	frame(17/18)
elaboration(3/[4-6])	continuation(11/12)	continuation(7/19)
continuation(4/5)	elaboration(12/13)	continuation([7-19]/20)
continuation(5/6)	frame(13/14)	

Figure 5. *Exemple de rattachement et typage des relations rhétoriques. Le graphe correspondant est montré dans la figure 6*

3.2.4. *Bilan des textes annotés en relations rhétoriques*

Au total, 87 textes ou extraits ont été annotés en relations rhétoriques. 47 ont été annotés simultanément par deux annotateurs naïfs, puis ont fait l'objet d'un arbitrage par des annotateurs experts. Les 40 textes ou extraits restants ont été mono-annotés par un expert participant au projet. Nous obtenons ainsi deux modalités d'annotation à disposition :

– (47×2) textes ou extraits avec une annotation par deux annotateurs naïfs ;

– 87 textes ou extraits avec une annotation experte : 47 avec une annotation naïve arbitrée et 40 avec une annotation experte.

19. Ce qui n'est le cas ni en RST ni pour le PDTB où les segments attachés doivent être contigus.

Nom	Description
Explication(a, b)	le segment b explique le segment a
But(a, b)	le segment b est un but du segment a
Résultat(a, b)	le segment b exprime un résultat produit par le segment a
Parallèle(a, b)	une construction parallèle relie les segments a et b
Contraste(a, b)	le segment b fait contraste avec le segment a
Continuation(a, b)	les segments a et b continuent une relation discursive (sur la CDU contenant a et b)
Conditionnel(a, b)	si segment a alors segment b
Alternation(a, b)	relation de disjonction entre deux segments : segment a ou segment b
Attribution(a, b)	le segment b est une spécification de ce qu'a dit son producteur (le segment a contient le fait que le producteur a dit quelque chose)
Arrière-plan(a, b)	le segment b fournit l'arrière-plan du segment a
Narration(a, b)	le segment b introduit un événement qui fait une suite narrative avec l'éventualité (événement ou état) introduit par le segment a
Flashback(a, b)	une narration relie a et b dans l'ordre temporel inverse
Encadrement(a, b)	le segment a (un adverbial ou groupe prépositionnel détaché en tête de phrase) sert de cadre pour b
Loc. temp.(a, b)	le segment b exprime une localisation temporelle du segment a, sans autre fonction discursive claire
Élaboration(a, b)	le segment b décrit un sous-événement de l'éventualité décrite par a
Élab. d'entité(a, b)	le segment b décrit une entité présente dans le segment a
Commentaire(a, b)	le segment b commente ou donne un avis sur le segment a
Méta-relations	avec une relation comme méta-explanation(a, b) le segment b décrit pourquoi le locuteur a prononcé le segment a

Relation	Exemple
explication(1,2)	*[L'équipe a perdu lamentablement hier.]_1 [Elle avait trop de blessés.]_2*
but(1,2)	*[Les chercheurs ont fait grève]_1 [pour montrer leur mécontentement.]_2*
résultat(1,2)	*[Nicholas avait bu trop de vin]_1 [et a donc dû rentrer chez lui en métro.]_2*
élaboration(1,[2-4])	*Cette année-là vit de nombreux changements dans la vie de nos héros.]_1 [Jean épousa Adèle,]_2 [Marie s'acheta une maison à la campagne,]_3 [et Paul partit pour le Brésil.]_4*

Tableau 4. *Relations rhétoriques annotées dans le corpus ANNODIS. La partie inférieure du tableau fournit des exemples de quelques unes de ces relations.*

Le tableau 5 donne une idée générale des annotations expertes en relations rhétoriques présentes dans le corpus ANNODIS. En moyenne, chaque texte contient 37 EDU, 18,5 CDU et 39 relations. Ce tableau permet également de distinguer les sous-corpus NEWS (brèves issues du quotidien l'*Est Républicain*) et WIK1/WIK2 (extraits d'articles Wikipédia, dont 4 extraits issus du même article) qui représentent la quasi-totalité des données annotées (seuls 3 extraits issus de 3 articles de linguis-

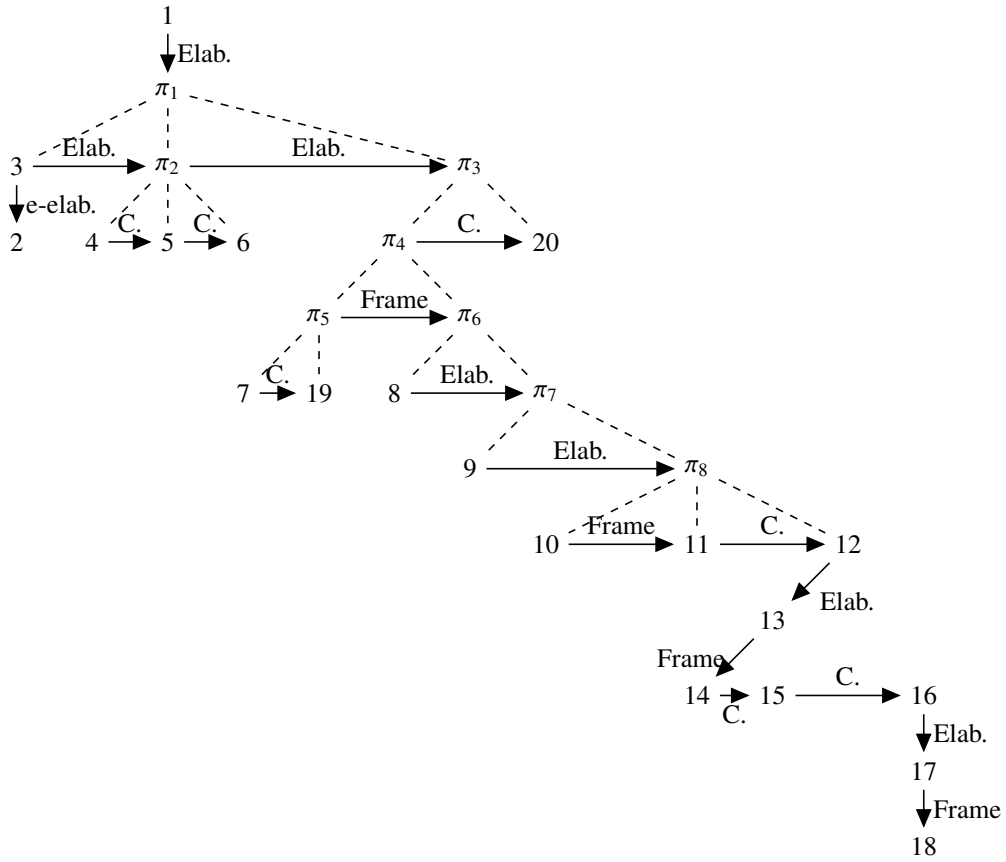

Figure 6. *Exemple de rattachement et typage des relations rhétoriques dans une forme graphique, basé sur les relations de la figure 5. Les nœuds correspondent aux unités discursives ; les EDU sont représentées par un nœud numéroté n et les CDU par un nœud numéroté π_n. Les arêtes pointillées sans flèches représentent l'inclusion dans une CDU, alors que les arêtes avec flèches représentent les relations rhétoriques. Elab. = Élaboration, e-elab = Élaboration d'Entité et C. = Continuation. Cette figure ne respecte pas la convention utilisée dans (Asher et Lascarides, 2003) où les relations subordonnantes sont verticales.*

tique et 3 extraits issus d'un même rapport en géopolitique ont été annotés à la fois au niveau des relations rhétoriques et des structures multi-échelles, voir tableau 5).

L'avantage d'une annotation par deux annotateurs naïfs est de permettre une analyse fine des difficultés que soulève la confrontation d'un modèle théorique du discours à la réalité des données et des capacités cognitives d'annotateurs humains. Ainsi, une première série d'observations montre une confusion fréquente entre certaines rela-

Relation rhétorique	Total (Nb)	(%)	NEWS (%)	WIK1 (%)
Alternation	18	0,5	0,3	0,6
Attribution	75	2,2	3,0	1,7
Arrière-plan	155	4,6	5,2	4,8
Commentaire	78	2,3	3,6	1,3
Continuation	681	20,3	20,1	21,1
Contraste	144	4,3	3,7	4,6
Élaboration d'entité	527	15,7	14,1	16,4
Élaboration	625	18,6	16,3	19,4
Explication	130	3,9	4,4	3,3
Flashback	27	0,8	1,4	0,6
Encadrement	211	6,3	6,2	5,7
But	95	2,8	3,1	2,4
Narration	349	10,4	11,1	10,4
Parallèle	59	1,8	2,2	1,8
Résultat	163	4,9	4,7	5,4
Localisation temporelle	18	0,5	0,5	0,5
Total	3 355	100	1 203	2 034

Tableau 5. *Données issues de l'annotation experte des relations rhétoriques*

tions de discours (*e.g.* relations d'explication et de résultat) et une association parfois biaisée entre marqueurs et relations.

L'annotation en double a également permis d'estimer un degré d'accord entre annotateurs et de mesurer la difficulté des différentes tâches. Concernant le rattachement entre EDU et CDU, le Kappa de Cohen tourne autour de 0,6 pour les attachements de segments, et est légèrement inférieur à 0,6 pour les relations étant donné un accord sur le point d'attachement. Ce score est considéré comme acceptable dans la littérature. Pour des tâches similaires comme la construction de graphes temporels dans le corpus TimeBank, les taux d'accord estimés sont également assez bas (55 % au total sur les relations, avec un kappa de 0,7 sur le type des relations dans le cas où la paire d'événements à relier est commune). De plus, nous avons intentionnellement évité de donner beaucoup de consignes sur les points d'attachement, pour ne pas introduire de biais sur les prédictions des théories du discours comme la RST ou la SDRT que nous voulions tester. Par ailleurs, la mesure d'accord inter-annotateur était assez brutale, ignorant largement des équivalences structurelles et entre relations, très délicates à mesurer.

3.3. *Premières exploitations*

3.3.1. *Segmentation automatique en EDU*

La segmentation produite par les annotateurs nous a permis de construire une implémentation par apprentissage automatique de la segmentation en EDU, dont une description a été publiée dans (Afantenos *et al.*, 2010). Contrairement aux autres théories du discours (Mann et Thompson, 1987 ; Webber et Joshi, 1998 ; Wolf et Gibson, 2005, entre autres), en SDRT ces unités peuvent être enchâssées l'une dans l'autre, ce qui complexifie nettement cette tâche.
Grâce à l'apprentissage automatique réalisé uniquement sur les EDU délimitées et après validation croisée, notre reconnaissance des bornes des segments est estimée à environ 88 % de F-score. Cette approche produisant éventuellement des segments mal formés, elle est complétée par un post-traitement symbolique ; la chaîne complète obtient alors un F-score de 73,3 % sur les EDU entières. La détection des bornes est proche des résultats existants pour la segmentation sans enchâssement en RST (90 %), même si la détection d'EDU entières est un peu en dessous (85 %) (Fisher et Roark, 2007).

3.3.2. *Détermination des points d'attachement et frontière droite*

Un des nos buts était de confronter les théories existantes avec des données issues de la phase d'annotation. Depuis les travaux de Wolf et Gibson (2005), les présupposés de la *Rhetorical Structure Theory* (RST) sur les points d'attachement possibles dans une structure discursive sont considérés comme faux ou du moins incomplets, particulièrement concernant l'attachement des segments non contigus. Wolf et Gibson (2005 ; 2006) soutiennent l'hypothèse qu'une analyse discursive doit permettre des dépendances croisées entre les segments. Cette hypothèse va à l'encontre de la contrainte de la frontière droite stipulée par la SDRT (Asher et Lascarides, 2003) : un nouveau segment peut être attaché à une DU non contiguë, mais il doit être attaché à une DU qui se trouve dans la frontière droite (définie de manière précise dans (Asher et Lascarides, 2003)).
Afin de tester la validité de la contrainte de la frontière droite en SDRT nous avons utilisé le corpus issu de la phase dite « naïve » où les annotateurs n'avaient aucune connaissance de cette contrainte (le guide d'annotation ne la mentionne pas). La contrainte de la frontière droite n'est violée que dans 5 % des attachements effectués dans le corpus (Afantenos et Asher, 2010). De plus, une analyse des erreurs effectuées sur ces 5 % montre que celles-ci sont principalement le fait des biais de l'interface Glozz sur le processus d'attachement. Nous avons également calculé qu'environ 20 % d'attachements dans le corpus naïf ne peuvent pas être formulés en RST parce qu'ils impliquent des segments non contigus.

3.3.3. *Détermination des relations rhétoriques*

Ce troisième aspect, portant sur le typage des liens rhétoriques est celui qui a le plus alimenté la réflexion théorique. Les aspects pratiques (détection automatique

par exemple) sont en cours d'étude. Pendant la phase de préparation du guide d'annotation, nous avons été confrontés à des confusions possibles entre relations et à des cas où aucune de nos relations existantes n'était satisfaisante. Ces problèmes ont donné lieu à plusieurs clarifications de la sémantique de certaines relations et indices qui signalent leur présence, concernant notamment la sémantique de la relation d'élaboration – relation la plus fréquemment annotée – (Vergez-Couret, 2010), les indices lexicaux et grammaticaux de l'élaboration (Adam et Vergez-Couret, 2010), le rôle de marqueurs potentiels de relations subordonnantes (Bras *et al.*, 2008 ; Bras et Schnedecker, 2009), la définition de nouvelles relations, comme l'élaboration d'entité (Prévot *et al.*, 2009).

4. Annotation des structures multi-échelles

4.1. *Structures énumératives et chaînes topicales : ancrages théoriques et travaux liés*

L'annotation des structures multi-échelles vise à constituer des données pour l'étude de l'organisation discursive à différents niveaux de granularité, en privilégiant la structuration de haut niveau en lien avec la « structure de document » (Power *et al.*, 2003), (Péry-Woodley et Scott, 2006). Elle trouve son ancrage premier dans la notion d'unité texte en linguistique systémique fonctionnelle, pour laquelle « *text is the unit of the semantic process* » (Halliday, 1977, p. 63), et qui incite à formuler des hypothèses sur le rôle de la perception de structures globales dans le calcul du sens à un niveau plus local [20]. Cette approche s'inspire aussi des recherches sur le rôle, dans la signalisation de la continuité ou des discontinuités, de marques associées à différents modes d'organisation discursive. De nombreux travaux se sont par exemple centrés sur les adverbiaux détachés à l'initiale ou sur les redénominations, qui sont susceptibles d'avertir le lecteur d'un changement thématique ou rhétorique. Au-delà de ces orientations générales, le choix des structures et la définition des modèles d'annotation s'appuient sur des travaux distincts pour chaque structure, travaux que nous évoquons dans ce qui suit.

La décision d'annoter les **structures énumératives** et l'intérêt pour leurs propriétés multi-échelles s'ancrent dans les travaux s'intéressant à la structure de document et à son marquage visuel. Power *et al.* (2003), se plaçant dans la perspective de la génération de texte, insistent sur ce qu'ils nomment le « composant graphique » de tout texte écrit. À la suite de la « *text grammar* » proposée par Nunberg (1990), ils définissent une « structure de document abstraite », indépendante des réalisations spécifiques, vue comme un niveau de description distinct de la structure rhétorique. Leurs préoccupations les rapprochent des travaux inspirés par le modèle d'architecture textuelle de Virbel (Luc et Virbel, 2001 ; Luc *et al.*, 1999 ; Luc *et al.*, 2000),

20. « *A text, as we are interpreting it, is a semantic unit, which is not composed of sentences but realized in sentences* » (Halliday, 1977, p. 46)

qui fonde notre conception des structures énumératives. Selon ce modèle, les structures énumératives sont des objets résultant d'un acte textuel par lequel « [l']identité de statut des constituants au sein de l'énumération exprime l'identité de statut des entités recensées dans le monde (cette identité étant, dans les deux cas, la coénumérabilité). » (Luc *et al.*, 2000, p. 25). Loin d'être un simple formatage, cette structure générique est à même d'organiser des segments de texte aux fonctions variées – argumentation, procédure, chronologie. Les unités coénumérées apparaissent dans une relation d'égalité vis-à-vis du critère de coénumérabilité exprimé ou non. La fonction discursive spécifique d'une structure énumérative est liée à la nature de ce critère d'interprétation, dont la réalisation linguistique (notre énumérathème) a pu être abordée dans les travaux sur les noms sous-spécifiés (Legallois, 2006) ou encore « *abstract* » ou « *shell nouns* » (Francis, 1989 ; Schmid, 2000 ; Mahlberg, 2005). La plupart des travaux directement concernés par ce que nous appelons structures énumératives se sont centrés sur les marques d'items, appelées diversement « marqueurs d'intégration linéaire » (Turco et Coltier, 1988 ; Jackiewicz, 2005), ou « *sequencers* » (Hempel et Degand, 2008). Tout en s'appuyant sur ces études, en particulier pour le prémarquage (cf. 4.2.1), notre démarche s'en distingue dans la mesure où nous n'envisageons l'étude des marqueurs que dans un deuxième temps : le fait d'avoir un ensemble de structures énumératives annotées manuellement fournit un référentiel pour l'étude des corrélats linguistiques des stratégies discursives qu'elles réalisent. Il devient notamment possible de tester notre hypothèse concernant la nature complexe des marqueurs discursifs, *i.e.* de fouiller nos données à la recherche de configurations récurrentes d'indices (voir section 4.3).

Quant aux **chaînes topicales**, nous les définissons pour cette annotation comme un type particulier de chaîne de cohésion dont les éléments contiennent des unités coréférentielles (potentiellement) topicales. Depuis l'ouvrage pionnier de Halliday et Hasan (1976) sur la cohésion textuelle, le lien entre anaphore et organisation discursive a suscité de nombreux travaux, dans le cadre des linguistiques fonctionnelles bien sûr (Cornish, 1998 ; Cornish, 1999), mais aussi au-delà en linguistique cognitive à travers la théorie de l'accessibilité (Ariel, 2001 ; Gundel *et al.*, 1993) ; dans la théorie du centrage (Grosz *et al.*, 1995 ; Gundel, 1998 ; Passonneau, 1998) ; en TAL pour la résolution d'anaphores ou le choix d'expressions référentielles en génération automatique. Plus généralement, la notion de cohésion est à l'origine de diverses approches de la « segmentation thématique », qui ont rencontré un intérêt considérable en TAL (cf. 4.3). Comme pour les structures énumératives, nous avons choisi de mettre l'accent sur les segments résultants, *i.e.* les segments de textes perçus par les annotateurs comme étant dominés par une chaîne référentielle majeure. Nous cherchions de cette façon à utiliser la campagne d'annotation pour sonder les intuitions des annotateurs sur l'existence de segments définissables en termes de référent « topical ». La procédure d'annotation pour ces chaînes topicales et pour les structures énumératives est présentée dans la section suivante.

4.2. *Procédure d'annotation des structures multi-échelles*

Nous commençons par préciser le modèle d'annotation pour les deux structures, avant de présenter les prétraitements, les trois phases d'annotation et enfin les post-traitements.

Le modèle d'annotation pour les **structures énumératives** contient les objets (segments de texte) introduits en section 2.1.2 : l'amorce, les items, la clôture et l'énumérathème. Sont également annotés et traités comme objets dans le modèle (unités pour GLOZZ) : les indices et la structure énumérative elle-même (le segment englobant). La première tâche consiste à détecter les structures et à en annoter les constituants (délimitation et typage) ; vient ensuite l'annotation des indices, c'est-à-dire des traits de surface considérés par les annotateurs comme ayant contribué à la détection des objets annotés dans la première tâche. L'annotation des indices peut passer par la validation de traits prémarqués (voir *infra*) ou par l'annotation de « nouveaux » traits.

Le modèle d'annotation pour les **chaînes topicales** repose uniquement sur la délimitation d'un segment appelé « unité référentielle » et des indices de continuité topicale associés (traits prémarqués validés ou nouveaux traits comme ci-dessus pour les structures énumératives). Il n'est pas impératif que ce segment soit uniquement composé de propositions portant sur le référent qui fait l'unité du segment : des commentaires, illustrations, par exemple, peuvent se trouver inscrits dans une chaîne topicale. Cette définition nous éloigne quelque peu des chaînes topicales telles qu'elles ont pu être définies dans la littérature, les annotations recueillies relevant davantage de ce que l'on pourrait nommer des « segments thématiques ». L'objectif de l'annotation n'était cependant pas de réaliser un pavage complet du texte (comme dans la segmentation thématique automatique) mais bien de repérer des segments multi-échelles (*i.e.* susceptibles de couvrir de larges portions de texte) construits, dans le cas des chaînes topicales, autour de continuités référentielles.
Afin d'encourager l'annotateur à porter un regard « global » sur le texte, les consignes d'annotation comportent une contrainte sur la taille, spécifiant que toute structure annotée doit couvrir plus d'une phrase ponctuationnelle, et une incitation explicite à chercher des structures de haut niveau. Elles suggèrent aux annotateurs de procéder à l'annotation en plusieurs passes, en se concentrant dans un premier temps sur les titres de sections et sur le surlignage des indices en position initiale de phrase (*via* les différents jeux de styles que permet l'interface GLOZZ).

4.2.1. *Prétraitements : encodage et prémarquage*

Préalablement à l'annotation, nous avons mis en œuvre un ensemble de prétraitements spécifiques faisant appel à des techniques de TAL. Les textes entiers du corpus ont tout d'abord été XMLisés et adaptés à la norme TEI-P5, norme internationale permettant une description complète des métadonnées à associer à tout document numérique. Ces métadonnées assurent la traçabilité des textes : de leur source

publiée en dehors du projet jusqu'à leur format numérisé et encodé selon la TEI-P5, en notant tous les traitements effectués ainsi que leur responsable technique. Si l'encodage des métadonnées est une étape désormais classique de la constitution de corpus, une originalité de cette phase de préparation a été l'encodage des propriétés visuelles des textes, alors que dans les corpus existants, rien n'est préservé de la mise en forme des textes d'origine (espaces verticaux et horizontaux, mise en valeur typographique, taille des caractères, etc.). Cet encodage novateur a été élaboré pour les besoins de l'annotation de structures multi-échelles, qui nécessitait que les textes apparaissent avec la mise en forme d'un « document réel » [21]. Les textes ont ensuite été prémarqués automatiquement. La figure 7 reprend l'extrait précédemment utilisé pour illustrer l'annotation en relations rhétoriques, visualisé ici dans l'interface d'annotation dans sa mise en forme « document » et avec le prémarquage.

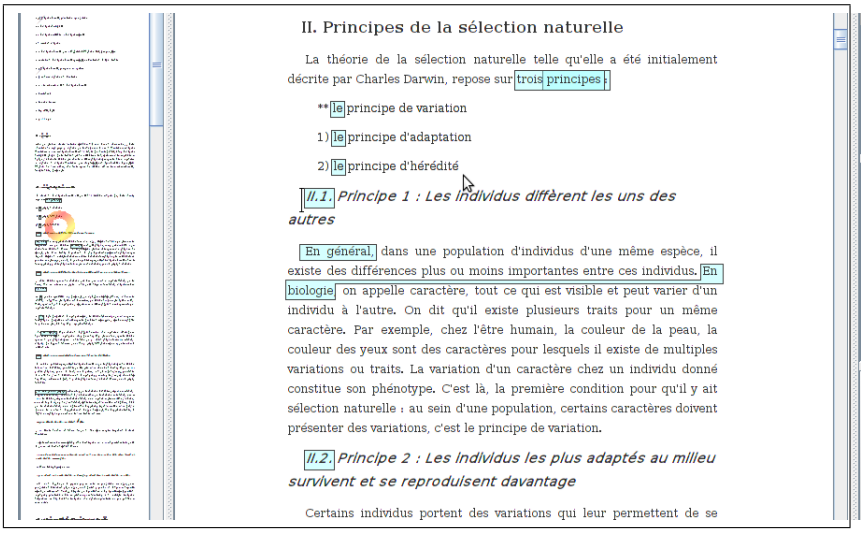

Figure 7. *Visualisation* GLOZZ *du prémarquage (éléments encadrés) pour l'annotation assistée des structures multi-échelle. Le ruban de gauche offre une vision synoptique du texte simultanément à la partie droite qui en donne une vision lisible.*

Ce prémarquage automatique, conçu dans une démarche inspirée des travaux de Biber (1988) et de Marcu (2000), concerne un ensemble de traits de surface, détaillés dans le tableau 6. La majorité de ces traits ont été définis dans la littérature comme des marques discursives participant à la signalisation des deux structures à l'étude. La figure 7 fournit des exemples d'éléments susceptibles de signaler des amorces (*trois principes* en fin de phrase et « *:* » en fin de paragraphe), des items (titres de sous-

21. Un des atouts de l'interface GLOZZ est la possibilité de visualiser et d'annoter des « documents ».

sections, puces et *En biologie*) et des clôtures (*En général*) [22]. Ce prémarquage fait

Traits participant à la structuration du discours	
Éléments lexico-syntaxiques	adverbiaux circonstanciels (spatiaux, temporels, notionnels), connecteurs...
Typographie et disposition	titres de sections, sauts de paragraphes...
Position textuelle	éléments détachés en initiale, position sujet...
Traits participant à la signalisation des structures énumératives	
Listes formatées	puces, indentations...
Patrons ponctuationnels	Fin de paragraphe ponctuée de « *:* » et motifs « *: [...] ; [...] et/ou [...]* »
Séquenceurs (MIL)	détachés en initiale (*Premièrement*, *Parallèlement*, *Enfin*, etc.) ou en position sujet (*Un second X*, *une autre*)
Prospections	SN pluriels + *selon*, *suivant(s)*, etc. ou SN dont la tête est un nom générique (*points*, *éléments* , *faits*)
Encapsulations	SN démonstratifs dont la tête est un nom générique et/ou avec adjectif numéral (*ces trois scénarios*)
Traits participant à la signalisation des chaînes topicales	
Expressions coréférentielles	pronoms, SN possessifs, réitérations
Appositions	détachées en initiale

Tableau 6. *Liste des traits prémarqués pour l'annotation assistée des structures multi-échelles*

appel à différentes procédures de repérage automatique qui s'appuient sur la structure de documents, ainsi que sur les sorties d'un étiquetage morphosyntaxique (TreeTagger) et d'une analyse syntaxique réalisée par l'analyseur SYNTEX (Bourigault, 2007). Un premier objectif visé était d'assister et de guider les annotateurs en rendant possible des procédures de « *text scannings* » (l'interface GLOZZ permettant de colorer les traits prémarqués selon différentes feuilles de style). Dans un deuxième temps, les traits prémarqués peuvent être associés aux annotations pour l'exploitation du corpus, que ce soit dans une optique d'analyse linguistique « classique » ou de mise en œuvre de techniques de fouilles. La figure 8 donne un exemple d'annotation réalisée sur la portion de texte prémarquée de la figure 7.

4.2.2. *Annotation et post-traitements*

L'annotation a été organisée en trois étapes afin de mesurer, au fil de la campagne, la faisabilité des tâches et le caractère « annotable » des objets (tableau 7).

22. Pour faciliter la visualisation « de loin » d'éléments très ténus, le premier mot voisin a été prémarqué : premier à droite pour les éléments susceptibles d'apparaître en début de segment *e.g.* « *puces + xxx* » ; premier à gauche pour les éléments susceptibles d'apparaître en fin de segment *e.g.* « *xxx + :* » en fin de paragraphe. Par ailleurs, pour éviter une trop grande surface de texte prémarqué, seul le début des traits a été prémarqué dans le cas d'expressions longues : numérotation pour les titres de section, trois premiers mots pour les adverbiaux circonstanciels, séquenceurs, prospections, encapsulations, expressions coréférentielles, appositions.

Phase A	3 textes (1 par sous-corpus) annotés chacun par les 3 annotateurs, qui pouvaient échanger entre eux et avec des experts, de manière à régler les problèmes de compréhension du guide d'annotation ou de la tâche
Phase B	6 textes (2 par sous-corpus) annotés chacun par les 3 annotateurs
Phase C	73 textes (20 WIKI, 16 CMLF et 20 GEOPO) annotés par 1 annotateur chacun

Tableau 7. *Trois phrases d'annotation pour les structures multi-échelles*

Les phases A et B ont permis d'évaluer l'accord inter-annotateur sur un ensemble de 9 textes multi-annotés (18 textes annotés). La F-measure obtenue est de 0,7 en comparant les frontières des structures et de leurs composants par paire d'annotation. Sur la base de ce résultat assez satisfaisant, la dernière phase en mono-annotation a été entamée. Au terme de cette phase, nous avons appliqué les tests de variance et

II. Principes de la sélection naturelle La théorie de la sélection naturelle telle qu'elle a été initialement décrite par Charles Darwin, repose sur **trois** **principes** : - le principe de variation - le principe d'adaptation - le principe d'hérédité	SE1	AMORCE SE2 AMORCE ITEM 1 ITEM 2 ITEM 3
II.1. Principe 1 : Les individus diffèrent les uns des autres En général, dans une population d'individus d'une même espèce, il existe des différences plus ou moins importantes entre ces individus. [...]. C'est là, **la première condition** pour qu'il y ait sélection naturelle : au sein d'une population, certains caractères doivent présenter des variations, **c'est le principe de variation**.		ITEM 1
II.2. Principe 2 : Les individus les plus adaptés au milieu survivent et se reproduisent davantage Certains individus portent des variations qui leur permettent de se reproduire davantage que les autres, [...]		ITEM 2
II.3. Principe 3 : Les caractéristiques avantageuses doivent être héréditaires **La troisième condition** pour qu'il y ait sélection naturelle est que les caractéristiques des individus doivent être héréditaires, [...]		ITEM 3
Ces trois premiers **principes** entraînent donc que [...]		CLOTURE

Figure 8. *Portion de texte annotée en structures multi-échelles. Cet extrait est celui annoté en relations rhétoriques (Fig. 6) et prémarqué (Fig 7). Dans la colonne de gauche, les lignes horizontales représentent les sauts de sections. Les éléments apparaissant en gras sont des indices annotés.*

Khi2 sur les textes annotés de la phase C : aucune corrélation entre un annotateur et la composition des structures n'a été observée.
Certaines opérations de nettoyage et d'harmonisation ont été mises en œuvre pour préparer la mise à disposition des corpus : élimination de doublons, mise en cohérence des noms d'indices ajoutés par les annotateurs. Ces post-traitements visent à améliorer la fiabilité des extractions et des analyses à partir du corpus.
Nous avons également réalisé une annotation experte sur un sous-ensemble des textes annotés de manière à fournir un indice de fiabilité des annotations naïves. De plus, dans le cas de textes triplement annotés (phases A et B), une annotation arbitrée a été réalisée afin de ne fournir qu'un jeu d'annotations par texte.

4.2.3. *Bilan des textes annotés en structures multi-échelles*

Au total, 82 textes entiers ont été annotés en structures multi-échelles : 829 structures énumératives et 487 chaînes topicales (répartition par sous-corpus dans le tableau 8).

Corpus	SE	Item	Amorce	Clôture	Énumérathème	CT
WIK2	332	1 639	296	34	167	232
LING	263	838	224	46	151	68
GEOP	234	716	180	43	120	187
ANNODIS	829	3 193	700	123	438	487

Corpus	Indices ajoutés	Indices prémarqués validés
WIK2	1 677	2 428
LING	937	708
GEOP	1 130	993
ANNODIS	3 744	4 129

Tableau 8. *Structures multi-échelles : effectifs d'objets annotés par sous-corpus*

4.3. *Enjeux et premières analyses*

4.3.1. *Réalité discursive des structures multi-échelles*

Les résultats de l'annotation des structures multi-échelles confirment les hypothèses concernant « l'annotabilité » des deux structures (structures facilement identifiables), leur fréquence (structures très utilisées dans les différents corpus), et l'intérêt d'une approche permettant d'observer des motifs textuels multi-échelles susceptibles d'apparaître à de très hauts niveaux d'organisation. La ressource ANNODIS comptabilise entre 5 et 12 chaînes topicales et entre 11 et 18 structures énumératives pour 10 000 mots. Ces structures couvrent en moyenne 15 % de la « surface textuelle » pour les chaînes topicales avec une taille moyenne de 187 mots, et 43 % pour les structures énumératives avec une taille moyenne de 388 mots.
Le caractère multi-échelle est davantage représenté par les structures énumératives qui

apparaissent à tous les niveaux de granularité de la structure du document (sections, sous-sections, paragraphes). Le tableau 9 indique la répartition de ces structures selon ces différents niveaux de granularité, qui constituent la base d'une première typologie de structures énumératives (Ho-Dac *et al.*, 2010). Par ailleurs, les structures énumératives sont souvent enchâssées, allant jusqu'à présenter dans notre corpus cinq niveaux d'enchâssement (corpus LING). La présence des structures énumératives à différents niveaux de granularité associée à cette capacité d'enchâssement en font un objet particulièrement intéressant pour aborder la complexité de l'organisation discursive.

Corpus	Niveaux de granularité			
	sections titrées	liste formatée	multiparagraphe	intraparagraphe
WIK1	19,3 %	39,1 %	20,8 %	20,8 %
LING	9,1 %	23,2 %	26,6 %	41,1 %
GEOP	6,8 %	10,3 %	20,9 %	62 %

Tableau 9. *Structures énumératives et structure de document*

Pour l'étude de la signalisation des structures énumératives (nous n'avons pas encore exploité les annotations concernant les chaînes topicales), les annotations fournissent des données susceptibles d'être « fouillées » pour identifier des configurations de traits – lexico-syntaxiques, typographiques, positionnels – qui constituent ce que nous appelons des « marqueurs discursifs complexes ». Ces configurations peuvent être corrélées aux types évoqués plus haut, elles devraient par ailleurs fournir la base d'analyseurs permettant leur repérage automatique. À travers la notion de « marqueur discursif complexe », nous examinons les relations entre différents modes de signalisation des structures discursives : structure de document, marqueurs lexico-syntaxiques, cohésion lexicale (Ho-Dac *et al.*, 2012).

4.3.2. *Sémantique des structures énumératives*

Le fait que les énumérathèmes (voir définition en section 4.1) soient annotés permet d'accéder directement au critère de coénumérabilité qui justifie une organisation sous forme de structure énumérative. Une première classification a permis d'observer la distribution générale des structures énumératives selon trois types d'énumérathèmes : les énumérathèmes désignant un « concept » (*la théorie repose sur trois* *principes*), une « entité » (*les individus se répartissent dans différents* *départements*) ou un « objet textuel » (*cet article de divise en trois* *sections*). Selon cette catégorisation, la grande majorité des structures énumératives (plus de 80 %) semblent être gouvernées par un concept contre seulement 9 % d'entités et 7,5 % d'objets textuels. Clairement, la classe majoritaire « concept » demande à être affinée, ce résultat préliminaire suggérant que la SE est de manière prédominante un procédé de création de catégories par le discours plutôt que d'expression de catégories préexistantes.

4.3.3. *Variations entre sous-corpus*

La diversité représentée dans le corpus ANNODIS permet à la fois d'évaluer la réalité d'un phénomène discursif à travers différents genres et de mesurer la variabilité de ses réalisations (Ho-Dac *et al.*, 2011). Les structures annotées sont significativement plus couvrantes dans les articles issus de Wikipédia (WIK2) que dans les deux autres sous-corpus (Khi2, $p < 0{,}01$) : les structures énumératives couvrent 54 % des textes de WIKI contre 29 % pour GEOP. Le caractère multi-échelle permet également de contraster les différents sous-corpus (sur la base des données présentées dans le tableau 9), avec des structures énumératives très formatées dans WIK2, des structures énumératives multiparagraphes (*i.e.* couvrant plusieurs paragraphes) dans LING et intraparagraphes dans GEOP (Khi2, $p < 0{,}05$). De même, le niveau d'enchâssement des structures distingue :

– les articles de Wikipédia (WIK2), caractérisés par des structures énumératives souvent enchâssées ;

– les articles scientifiques (LING) dont les structures énumératives, très profondes, peuvent présenter jusqu'à cinq niveaux d'enchâssement ;

– les rapports (GEOP) dont les structures énumératives sont au contraire plus plates.

5. Intersections : structures énumératives et « structures élaboratives »

L'interfaçage des annotations de type relations rhétoriques et structures multi-échelles n'a pas encore fait l'objet d'études approfondies, la proportion de la ressource ANNODIS proposant une annotation complète étant actuellement très faible (7 364 mots, voir tableau 2). La mise à disposition de la ressource encouragera, nous l'espérons, de nouvelles campagnes d'annotation qui combleront cette faiblesse.
Quelques pistes ont cependant déjà été suivies pour explorer cette intersection. Une manifestation de la récursivité dans les structures enumératives a ainsi fait l'objet d'une analyse et d'une formalisation dans Vergez-Couret *et al.* (2008). Un examen approfondi d'un possible isomorphisme entre structures énumératives et « structures élaboratives » (*i.e.* construites par un ensemble de relations d'élaboration) a été mené par Vergez-Couret (2010).
Par ailleurs, le rôle des adverbiaux temporels en tant que déclencheurs de la relation Frame (Topic) et l'interaction avec la relation Background (Vieu *et al.*, 2010) sont des terrains actuellement à l'étude qui nécessite cette double approche caractéristique de la ressource ANNODIS.

6. Bilan et perspectives

La ressource ANNODIS est le premier corpus écrit de français systématiquement enrichi d'annotations de structures discursives. Notre objectif dans cet article a été d'en fournir une description complète et accessible, depuis les soubassements

théoriques des deux modèles d'annotation (relations rhétoriques et structures multi-échelles) jusqu'aux premières exploitations, en passant par les procédures d'annotation et les post-traitements. L'annotation de structures discursives, ancrée dans un domaine de recherche encore peu stabilisé, constitue un défi considérable. La construction de modèles d'annotation opérationnalisables, leur explicitation dans des guides d'annotation, la définition et l'implémentation des prétraitements et la mise en œuvre de l'annotation manuelle ont constitué une expérimentation à grande échelle qui a soulevé de nombreux questionnements, parmi lesquels [23] :

– le choix des annotateurs : naïfs ou non ? Peut-on parler d'annotateurs naïfs s'ils doivent faire l'objet d'une formation et se référer à un guide complexe ? Peut-on alors encore considérer qu'on accède à leurs intuitions de locuteurs du français ? Les lecteurs naïfs peuvent-ils formuler des intuitions dans des tâches aussi complexes que l'interprétation des discours ?

– le calcul de l'accord inter-annotateur : quelle mesure ? Quelle marge autoriser sur la définition des segments et des relations ? Comment calculer l'accord quand la procédure autorise plusieurs catégories, façon légitime de gérer l'incertitude pour une annotation de niveau discursif ?

– le corpus mis en ligne doit-il refléter les aléas des campagnes d'annotation (priorité à la dimension expérimentale) ou doit-il être nettoyé (priorité à la notion de corpus de référence) ? Y a-t-il lieu de mettre en ligne différentes annotations pour un même texte, même si certaines nous paraissent erronées ?

– les guides d'annotation ont évolué au cours des campagnes et depuis, reflétant les avancées dans la compréhension des phénomènes annotés. Quelle version mettre en ligne ?

La mise en ligne de la ressource ANNODIS marque la fin d'une première phase, celle prévue dans le cadre du projet qui en a permis le développement initial. Les équipes à l'origine de la ressource entendent poursuivre leurs objectifs de recherche par son exploitation, dont les premiers éléments ont été décrits ici. Au-delà, ce qui est visé par la diffusion d'une ressource comme celle-ci, c'est que d'autres chercheurs se l'approprient, la complètent et éventuellement la détournent selon leurs objectifs propres.

Remerciements

Nous remercions les relecteurs anonymes pour leurs remarques et suggestions sur les premières versions de cet article.

23. L'école thématique du CNRS sur l'annotation de données langagières (Biarritz 10-16 septembre 2011) a été l'occasion d'une réflexion sur ces questions ; voir support « Aspects méthodologiques (écrit) : Multi-annotation discursive de corpus écrit » disponible sur le site http ://annotationlinguistique.fr.

7. Bibliographie

Adam C., Vergez-Couret M., « Signalling Elaboration : Combining Gerund Clauses with Lexical Cues », *Proceedings of Signalling Text Organisation (Multidisciplinary Approaches to Discourse 10)*, 2010.

Afantenos S. D., Asher N., « Testing SDRT's Right Frontier », *Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010)*, p. 1-9, 2010.

Afantenos S. D., Denis P., Muller P., Danlos L., « Learning Recursive Segments for Discourse Parsing », *Proceedings of LREC 2010*, 2010.

Ariel M., *Accessibility Theory : an overview*, John Benjamins : Amsterdam/Philadelphia, 2001.

Asher N., *Reference to Abstract Objects in Discourse*, Kluwer, 1993.

Asher N., Lascarides A., *Logics of Conversation*, Studies in Natural Language Processing, Cambridge University Press, Cambridge, UK, 2003.

Bestgen Y., Degand L., Spooren W., « On the use of automatic techniques to determine the semantics of connectives in large newspaper corpora : an exploratory study », *in* L. Lagerwerf, W. Spooren, L. Degand (eds), *Determination of Information and Tenor in Texts : Multidisciplinary Approaches to Discourse*, Stichting Neerlandistiek & Nodus Publikationen, Amsterdam/Münster, p. 189-202, 2003.

Biber D., *Variation Across Speech and Writing*, Cambridge : Cambridge University Press, 1988.

Bourigault D., « Un analyseur syntaxique opérationnel : SYNTEX », 2007, Mémoire d'HDR, Université de Toulouse.

Bras M., Prévot L., Vergez-Couret M., « Quelle(s) relation(s) de discours pour les structures énumératives ? », *Actes du Congrès Mondial de Linguistique Française*, Paris, p. 1945-1964, 2008.

Bras M., Schnedecker C., « Dans un (premier+second+nième) temps et les relations de discours : de l'élaboration à la contre-argumentation », *Linguistic and Psycholinguistic Approaches to Text Structuring*, 2009.

Cornish F., « Les chaînes topicales : leur rôle dans la gestion et la structuration du discours », *Cahiers de Grammaire*, vol. 23, p. 19-40, 1998.

Cornish F., *Anaphora, Discourse and Understanding. Evidence from English and French.*, Calendron Press : Oxford, 1999.

Fisher S., Roark B., « The utility of parse-derived features for automatic discourse segmentation », *ACL*, 2007.

Francis G., « Thematic selection and distribution in written discourse », *Word*, vol. 40, p. 201-221, 1989.

Groenendijk J., Stokhof M., « Dynamic Predicate Logic », *Linguistics and Philosophy*, vol. 14, p. 39-100, 1991.

Grosz B., Joshi A., Weinstein S., « Centering : A framework for modelling the local coherence of dicourse », *Computational Linguistics*, vol. 21, n° 2, p. 203-225, 1995.

Gundel J., « Centering Theory and the Givenness Hierarchy : Towards a Synthesis », *in* A. J. M. Walker, E. Prince (eds), *Centering Theory in Discourse*, Calendron Press : Oxford, p. 183-198, 1998.

Gundel J. K., Hedberg N., Zacharski R., « Cognitive Status and the Form of Referring Expressions in Discourse », *Language*, vol. 69, p. 274-307, 1993.

Habert B., *Instruments et ressources électroniques pour le français*, Ophrys, Gap/Paris, 2005.

Halliday M., « Text as semantic choice in social contexts », *in* T. van Dijk, J. Petofi (eds), *Grammars and Descriptions*, Walter de Gruyter : Berlin, p. 176-226, 1977.

Halliday M., Hasan R., *Cohesion in English*, Longman : London, 1976.

Hempel S., Degand L., « Sequencers in Different Text Genres : Academic Writing, Journalese and Fiction », *Journal of Pragmatics*, vol. 40, p. 676-693, 2008.

Ho-Dac L.-M., Fabre C., Péry-Woodley M.-P., Rebeyrolle J., Tanguy L., « High-level discourse structures : Topical Chains and Enumerative Structures in a diversified annotated corpus », *Corpus Linguistics*, Birmingham, 2011.

Ho-Dac L.-M., Fabre C., Péry-Woodley M.-P., Rebeyrolle J., Tanguy L., « On the signalling of multi-level discourse structures », *Discours*, 2012.

Ho-Dac L.-M., Péry-Woodley M.-P., Tanguy L., « Anatomie des structures énumératives », *TALN 2010*, ATALA, Université de Montréal, Montréal, July, 2010.

Jackiewicz A., « Les séries linéaires dans le discours », *Langue française*, vol. 148, p. 95-110, 2005.

Kamp H., Reyle U., *From Discourse to the Lexicon : Introduction to Modeltheoretic Semantics of Natural Language, Formal Logic and Discourse Representation Theory*, Kluwer Academic Publishers, 1993.

Labadié A., Enjalbert P., Mathet Y., Widlöcher A., « Discourse structure annotation : Creating reference corpora », *LREC 2010 Proceedings*, Malte, May, 2010.

Legallois D., « Quand le texte signale sa structure : la fonction textuelle des noms sous-spécifiés », *CORELA - Organisation des textes et cohérence des discours (numéro thématique)*, 2006.

Luc C., Mojahid M., Péry-Woodley M.-P., Virbel J., « Les énumérations : structures visuelles, syntaxiques et rhétoriques », *Actes de CIDE 2000 (Colloque International sur le Document Électronique)*, p. 21-40, 2000.

Luc C., Mojahid M., Virbel J., Garcia-Debanc C., Péry-Woodley M.-P., « A linguistic approach to some parameters of layout : A study of enumerations », *AAAI 1999 Fall Symposia "Using Layout for the Generation, Understanding or Retrieval of Documents*, North Falmouth, Massachussetts, p. 20-29, 1999.

Luc C., Virbel J., « Le modèle d'architecture textuelle : fondements et expérimentation », *Verbum*, vol. 1, n° 23, p. 103-123, 2001.

Mahlberg M., *English General Nouns : A corpus theoretical approach*, Amsterdam, Philadelphia : John Benjamins, 2005.

Mann W., Thompson S., Rhetorical Structure Theory : a theory of text organization., Technical report, Information Science Institute, 1987.

Marcu D., « The Rhetorical Parsing of Unrestricted Texts : A Surface-based Approach », *Computational Linguistics*, vol. 26, n° 3, p. 395-448, 2000.

Marcu D., « Automatic Discourse Parsing », *Encyclopedia of Language and Linguistics*, 2nd edn, Elsevier, Oxford, p. 649-654, 2006.

Mathet Y., Widlöcher A., « La plate–forme GLOZZ : environnement d'annotation et d'exploration de corpus », *TALN 2009*, ATALA, LIPN, Senlis, June, 2009.

Nunberg G., « The Linguistics of Punctuation », *csli, Lecture Notes, University of Chicago Press*, 1990.

Passonneau R. J., « Interaction of Discourse Structure with Explicitness of Discourse Anaphoric Noun Phrases », *in* M. Walker, A. Joshi, E. Prince (eds), *Centering Theory in Discourse*, Calendron Press : Oxford, p. 327-358, 1998.

Péry-Woodley M.-P., « Discours, corpus, traitements automatiques », *in* A. Condamines (ed.), *Sémantique et corpus*, Paris : Hermès, p. 177-210, 2005.

Péry-Woodley M.-P., Asher N., Enjalbert P., Benamara F., Bras M., Fabre C., Ferrari S., Ho-Dac L.-M., Le Draoulec A., Mathet Y., Muller P., Prévot L., Rebeyrolle J., Tanguy L., Vergez-Couret M., Vieu L., Widlöcher A. *et al.*, « ANNODIS : une approche outillée de l'annotation de structures discursives », *TALN 2009*, ATALA, LIPN, Senlis, June, 2009.

Péry-Woodley M.-P., Scott D., « Computational Approaches to Discourse and Document Processing », *TAL*, vol. 2, n° 47, p. 7-19, 2006.

Power R., Scott D., Bouayad-Agha N., « Document Structure », *Computational Linguistics*, vol. 2, n° 29, p. 211-260, 2003.

Prasad A., Miltsakaki R., Dinesh E., Lee N., Joshi A., Webber B., *Penn Discourse TreeBank 1.0 Annotation Manual*. 2006, `www.seas.upenn.edu/~pdtb/`.

Prévot L., Asher N., Vieu L., « Une formalisation plus précise pour une annotation moins confuse : la relation d'Élaboration d'entité », *Journal of French Language Studies*, vol. 19, n° 2, p. 207-228, 2009.

Schmid H.-J., *English abstract nouns as conceptual shells : from corpus to cognition*, Berlin, New York : Mouton de Gruyter, 2000.

Turco G., Coltier D., « Des agents doubles de l'organisation textuelle, les marqueurs d'intégration linéaire », *Pratiques*, vol. 57, p. 57-79, 1988.

Vergez-Couret M., Étude en corpus des réalisations linguistiques de la relation d'Élaboration, PhD thesis, Université de Toulouse II, Le Mirail, 2010.

Vergez-Couret M., Prévot L., Bras M., « Interleaved discourse : the case of two step enumerative structures », *Proceedings of Contraints In Discourse III*, 2008.

Vieu L., Bras M., Prévot L., « On the compositionality of temporal locating adverbial modification », *Actes des 8es Journées Sémantique et Modélisation (JSM '10)*, 2010.

Webber B., Joshi A., « Anchoring a lexicalized tree-adjoining grammar for discourse », *Coling/ACL Workshop on Discourse Relations and Discourse Markers*, Montreal, Canada, p. 86-92, 1998. http ://www.cis.upenn.edu/ bonnie/dwork98-col.ps.gz.

Wolf F., Gibson E., « Representing Discourse Coherence : A Corpus Based Study », *Computational Linguistics*, vol. 31, n° 2, p. 249-287, 2005.

Wolf F., Gibson E., *Coherence in Natural Language : Data Stuctures and Applications*, The MIT Press, 2006.