Vers l'annotation discursive de textes d'élèves

Claudine Garcia-Debanc, Lydia-Mai Ho-Dac, Myriam Bras, Josette Rebeyrolle

Vers l'annotation discursive de textes d'élèves

Claudine GARCIA-DEBANC, Lydia-Mai HO-DAC,
Myriam BRAS, Josette REBEYROLLE
Université de Toulouse, Laboratoire CLLE-ERSS, UMR 5263,
CNRS & UT2 Jean-Jaurès, SFR AEF ESPE de l'Académie de
Toulouse

Les « grands corpus de textes scolaires » peuvent apparaître de taille très modeste par rapport aux corpus actuellement traités en linguistique de corpus, qui rassemblent plusieurs millions de mots. Leur constitution apparait toutefois indispensable par rapport à la tradition des travaux en didactique du français langue première sur l'analyse des textes scolaires, qui s'attachent généralement à une analyse fine d'un échantillon très limité de textes (Fabre-Cols, 2000, Cappeau, Roubaud, 2005, Boré, 2011, Masseron, 2005). Or, comme l'a déjà souligné Elalouf (2011), la constitution de tels corpus correspond à des enjeux à la fois scientifiques et sociétaux. D'un point de vue scientifique, leur analyse permettrait de faire progresser les connaissances sur la mise en place des compétences rédactionnelles tout au long du cursus scolaire (école, collège, lycée) par l'étude systématique d'un nombre important de productions écrites, afin d'élaborer une cartographie d'indicateurs de réussite pour les différents niveaux de fonctionnement de la langue (lexical, syntaxique, textuel, discursif), à la manière du travail déjà engagé sur les compétences orthographiques (Geoffre, 2014). D'un point de vue sociétal, de tels corpus peuvent constituer des ressources pour les chercheurs mais aussi pour les formateurs d'enseignants et les étudiants se préparant aux métiers de l'enseignement.

La présente contribution se propose de mettre en évidence les problèmes épistémologiques et méthodologiques posés par l'annotation discursive de productions écrites d'élèves et les choix qu'implique sa mise en œuvre. Elle prend appui sur les acquis des travaux d'annotation discursive réalisés dans le

cadre du projet ANNODIS[1], dans lequel ont été impliquées trois des auteurs de cette contribution.

Après avoir indiqué les enjeux de la constitution de « grands » corpus scolaires (section 1), nous rappellerons les problèmes posés par l'annotation discursive de corpus (section 2). Nous interrogerons la transférabilité de ces méthodes aux textes d'élèves et évoquerons quelques-uns des problèmes méthodologiques rencontrés. Nous nous attacherons plus particulièrement à deux questions précédemment travaillées dans le projet ANNODIS, d'une part celle des Relations de Discours (section 3), d'autre part celle de l'annotation de chaînes référentielles (section 4).

1. De « grands » corpus scolaires de textes en réponse à une consigne identique

La constitution de corpus de textes scolaires peut répondre à des enjeux différents. Si le projet est de décrire l'enseignement de l'écriture dans les classes, tel qu'il est réalisé aujourd'hui, il est nécessaire de collecter des travaux d'écriture de façon écologique au plus près des pratiques des enseignants. Le projet de notre équipe est différent, dans la mesure où nous nous proposons d'élaborer une cartographie des compétences lexicales, textuelles et discursives en cours de construction. Notre choix est donc de collecter des textes en réponse à des consignes d'écriture identiques pour pouvoir comparer les marques linguistiques utilisées préférentiellement par des rédacteurs d'âges et de niveaux de compétences différents.

Deux corpus sont actuellement en cours de constitution dans notre équipe. Le premier corpus, dit « Microloup », a été conçu dans le cadre d'une recherche collaborative innovante portant sur l'enseignement des verbes de déplacement (Garcia-Debanc, Gangneux, 2015 ; Aurnague, Garcia-Debanc, 2016). De même que dans les recherches psycholinguistiques, pour éviter l'amorçage lexical, a été utilisé un inducteur sous la forme d'une

1 Projet financé par l'ANR (appel Corpus, 2007-2010). Partenaires : CLLE-ERSS (Toulouse), IRIT (Toulouse), GREYC (Caen).

image. La tâche d'écriture s'appuie sur un dessin d'animation, *Micro-loup*, de Richard Mac Guire, d'après un scénario de Grégoire Solotareff et Jean-Luc Fromental, WB France télévisions Distribution (support accessible en ligne : http://vimeo.com/984305). Après double visionnement du film, les élèves sont invités à rédiger individuellement un récit rendant compte de l'histoire pour quelqu'un qui n'a pas vu le film, en réponse à la consigne suivante :

> Raconte les aventures de Micro Loup pour un camarade qui n'a pas vu le film.

Ont été collectés environ 350 textes d'élèves du CE2 (3° primaire) à la sixième (1° année de l'enseignement secondaire). Sont archivés les textes rédigés avant et après la séquence d'enseignement sur les verbes de déplacement. Les textes ont été rédigés en classe, sans matériel particulier et sont donc manuscrits. L'enjeu premier de la constitution de ce corpus était de comparer les verbes de déplacement utilisés dans les deux productions. Les données peuvent également être utilisées pour analyser d'autres faits de langue relatifs à la syntaxe ou à l'organisation textuelle et discursive.

La présente contribution portera plus particulièrement sur le deuxième corpus en cours de construction, dit corpus Charolles, constitué en vue de travailler sur la cohésion textuelle. La consigne d'écriture proposée est une tâche-problème imposant aux élèves la résolution de problèmes d'anaphores (Garcia-Debanc, Bonnemaison, 2014 ; Garcia-Debanc, Bras, 2016). Cette consigne d'écriture est très nettement inspirée de celle de Charolles (1988). Toutefois, alors que Charolles (1988), dans la tradition des travaux psycholinguistiques, choisit comme support inducteur un ensemble d'images en bande dessinée, pour éviter un amorçage linguistique, nous avons préféré utiliser un ensemble de phrases, dans la mesure où nous projetions d'analyser la prise en compte de certains indices linguistiques par les élèves dans la résolution de cette tâche. La tâche pourrait être qualifiée d'épilinguistique, dans la mesure où elle attire l'attention des élèves sur une dimension du fonctionnement de la

langue, sans pour autant utiliser les termes du langage technique. La consigne proposée est la suivante :

> Racontez une histoire dans laquelle vous insèrerez séparément et dans l'ordre donné les trois phrases suivantes :
> Elle habitait dans cette maison depuis longtemps. (P_a)
> Il se retourna en entendant ce grand bruit. (P_b)
> Depuis cette aventure, les enfants ne sortent plus la nuit. (P_c)

Pour éviter que des modifications soient apportées aux phrases imposées, celles-ci sont inscrites sur des bandelettes de papier que les élèves collent là où ils le souhaitent dans le texte en cours d'écriture. Ce dispositif a été choisi pour éviter d'induire la rédaction d'une longueur donnée de texte si ces phrases avaient été séparées par des pointillés.

La tâche proposée conduit les élèves à gérer des marques de cohésion textuelle en résolvant plusieurs anaphores référentielles : pronoms personnels (*elle, il*), syntagmes nominaux comportant un déterminant démonstratif anaphorique (*cette maison, ce grand bruit, cette aventure*), syntagme nominal comportant un article défini pluriel (*les enfants*). En revanche, la consigne ne donne aucune indication sur les personnages. De même, le choix du genre du récit est tout à fait ouvert. Il est guidé par la phrase fermoir, avec les mots *nuit* et *aventure* et le caractère de moralité qu'elle semble revêtir, qui suggère un danger et une peur. Ces éléments linguistiques constituent donc à la fois des contraintes linguistiques à prendre en compte et des lanceurs d'écriture, tremplins de l'imaginaire. Ainsi la dernière phrase joue un rôle important pour la planification textuelle (Hayes, Flower, 1980), dans la mesure où elle contraint le cadre temporel de l'épisode-clé du récit, la nuit, et oriente l'atmosphère générale : si les enfants ne sortent plus la nuit, c'est qu'un ou plusieurs d'entre eux ont vécu une ou plusieurs aventures terrifiantes ou dangereuses la nuit. Les rédacteurs doivent également décider si les enfants mentionnés dans la dernière phrase sont un terme générique reprenant les personnages désignés précédemment par *il* et *elle* ou bien s'il s'agit d'autres personnages ou encore si cette moralité est une généralisation à partir de l'aventure arrivée à un seul enfant. Par ailleurs, l'emploi

des temps verbaux, imparfait dans la première phrase, passé simple dans la seconde et présent de vérité générale dans la troisième, induit une narration dans le système imparfait-passé simple, le présent de la phrase finale se justifiant par le caractère générique de l'affirmation. Le rédacteur doit également prendre en compte les indications spatiales et temporelles présentes dans la consigne : *dans cette maison, depuis longtemps, la nuit.* Malgré ces diverses contraintes, la consigne est ouverte, dans la mesure où elle permet une grande diversité de solutions du point de vue du genre du récit et des caractéristiques des personnages. Pour chacune des occurrences à insérer, la résolution du problème d'absence de référence explicite est à réaliser en insérant des phrases supplémentaires avant l'occurrence. Toutefois les distances entre une reprise anaphorique et un terme source peuvent être diverses. La résolution peut être effectuée selon des enchaînements locaux entre phrases successives ou mettre en jeu des chaînes anaphoriques.

Une analyse de la tâche permet aussi de faire apparaître la nécessité d'une planification à rebours (Schneuwly, 1988) : la phrase finale induit la construction d'un récit qui doit mettre en scène des enfants, apparaître comme une « aventure » et conduire à l'abandon des sorties nocturnes, probablement du fait d'une scène effrayante.

L'unicité de la consigne permet de comparer les productions en vue d'élaborer une cartographie des moyens linguistiques choisis par les élèves et de mettre en évidence des indicateurs de progressivité.

Le corpus est constitué des textes d'élèves scannés et de leur transcription, assortie de métadonnées indiquant le niveau scolaire, l'identité du rédacteur, l'école et la date de collecte des données. La version originale du texte scannée permet en effet de revenir au texte original, dans sa mise en page et avec ses ratures.

Le corpus enrichi est constitué des documents complémentaires correspondant au traitement didactique des écrits produits, notamment des réécritures des textes individuelles ou en binômes, ainsi que des documents fournis par

les enseignants sur la mise en œuvre de l'activité dans leur classe et sa contextualisation.

Le corpus complet est constitué de 400 textes d'élèves de 9 à 16 ans (3° P à 6° P et première et dernière année de collège) et de 40 textes d'étudiants de Master (Master Professeurs des Écoles, Master Métiers de l'écriture et Master Sciences du Langage). Ces textes sont manuscrits et nécessitent une transcription.

2. Questions sur l'annotation discursive de corpus : les leçons d'ANNODIS

Afin de discuter les questions que pose l'annotation discursive d'un corpus de textes d'élève, nous proposons dans cette section un aperçu des enseignements tirés de l'expérience d'annotation discursive qui a abouti à la création de la ressource ANNODIS[2] (Péry-Woodley et alii, 2011), corpus de textes en français écrit standard annoté discursivement. Les textes constitutifs sont des articles de presses issus du journal *L'Est Républicain*, des articles encyclopédiques prélevés dans l'encyclopédie Wikipédia, des écrits académiques figurant dans les actes du *Congrès Mondial de Linguistique Française* (CMLF 2008) et des rapports publiés par l'Institut Français des Relations Internationales (IFRI). Bien que le type de textes composant la ressource ANNODIS diffère grandement des textes d'élèves, un certain nombre de questions communes nous semblent se poser, quel que soit le type des textes (expositifs, narratifs, etc.) et le format des textes (longs ou courts, structurés ou non).

Le projet ANNODIS avait pour objectif la construction, l'analyse et la diffusion d'un corpus diversifié construit de manière raisonnée, mis en forme selon les standards actuels, et enrichi de multiples annotations, dans le but de mettre à disposition une ressource permettant d'expérimenter au niveau discursif des techniques de linguistique de corpus outillée et de TAL (Traitement Automatique des Langues). Quatre

2 La ressource Annodis est diffusée librement sur le site REDAC (Ressources Développées à CLLE-ERSS) : http://redac.univ-tlse2.fr/

phénomènes discursifs ont été annotés : les chaînes de référence, les structures énumératives, les Unités de Discours Élémentaires (désormais UDE) et les Relations de Discours entre ces unités. Les chaînes de référence et les structures énumératives sont définies comme des structures multi-échelles susceptibles d'organiser des petites comme de très grandes portions de texte (cf. Halliday, 1977). Ces phénomènes relèvent d'une annotation ciblée étant donné que les objets annotés ne couvrent pas la totalité du texte. Annoter de tels objets vise essentiellement à rassembler des occurrences d'un phénomène pour l'étudier dans une approche exploratoire (Ho-Dac et alii, 2010, Rebeyrolle, Péry-Woodley, 2014). En revanche, l'analyse en UDE et en relations rhétoriques s'inscrit dans le cadre des théories du discours visant à modéliser la structure complète d'un discours vu comme un ensemble d'unités élémentaires reliées par des Relations de Discours (e.g. *Rhetorical Structure Theory*, Mann, Thompson, 1987; *Segmented Discourse Representation Theory*, Asher, 1993). Au contraire des structures multi-échelles, l'annotation de ces phénomènes implique donc une couverture complète des textes dans la mesure où le texte entier est segmenté en UDE et que toute UDE est reliée à une autre UDE par une Relation de Discours. Annoter de tels objets vise à la fois à évaluer un modèle théorique (Afantenos, Asher, 2010) et à permettre une approche onomasiologique des relations des discours (e.g. Vergez-Couret (2010), sur la relation d'élaboration et Atallah (2014), sur les relations causales).

Les différentes perspectives réunies dans le projet ANNODIS – annotation ciblée et couverture complète, approche exploratoire et évaluation d'un modèle – ont permis une confrontation réelle à une variété de problèmes (cf. Ho-Dac, Péry-Woodley, 2014) face auxquels certaines recommandations semblent pertinentes pour la mise en place d'une campagne d'annotation discursive des textes d'élèves. Deux questions sont discutées dans les deux sections suivantes : quelle version de texte annoter et quels objets annoter et comment ?

2.1. Quelle version de texte annoter ? Mesurer la distance entre données primaires et texte à annoter

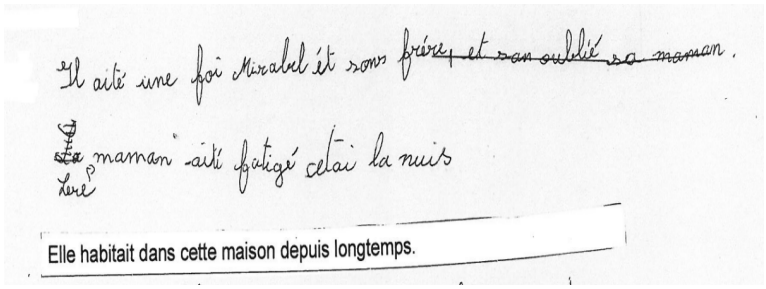
Dans le projet ANNODIS, les textes ont été présentés aux annotateurs via l'interface GLOZZ (Mathet et Widlöcher, 2009), spécifiquement développée pour permettre l'annotation de textes visualisés sous une forme la plus proche possible de leur version d'origine. Permettre ce type d'affichage passe nécessairement par l'encodage des propriétés visuelles du document à annoter, encodage qui consiste à identifier et définir les « objets textuels » (Virbel et Luc 2001) présents dans le document : niveaux et titres de section, paragraphes, citations, listes formatées, etc. Cet encodage des propriétés typo-dispositionnelles du texte ne répond pas à un souci esthétique mais à une intégration, dans le modèle de discours, des indices de mise en forme (cf. Virbel et Luc 2001). Selon ces modèles, les sauts de paragraphes, les espaces verticaux, les mises en gras, etc. constituent des traces du processus d'écriture auxquelles correspondent des « contreparties discursives » qu'il faut prendre en compte dans l'analyse du discours.

Une solution pour encoder ces objets est de suivre la norme TEI-P5, norme internationale et pérenne mise en place par le consortium Text Encoding Initiative (http://www.tei-c.org) composé d'universitaires et éditeurs dans le but d'homogénéiser le codage des textes utilisés en sciences humaines. Cette norme concerne à la fois les métadonnées associées à un document, l'annotation structurelle (chapitre, paragraphe, section, phrase, etc.) et l'annotation de contenu (tableaux, dessins, formules). Les métadonnées assurent la traçabilité des textes : de leur contexte de production jusqu'à leur format et licence de diffusion. Il s'agit d'indiquer par exemple dans l'élément appelé « teiHeader » : le contexte de la récolte des données, les consignes d'écriture et toutes les étapes qui ont permis la numérisation du texte (transcription, normalisation, encodage, annotation). Le manuel de la TEI décrit tous les types de métadonnées qui peuvent être associées à un document, comme par exemple l'élément « correction » qui « établit comment et dans quelles circonstances des corrections ont été apportées au texte » ou encore l'élément « revisionDesc » qui permet de lister toutes les révisions d'un document. Le renseignement dans les métadonnées de chaque

étape de numérisation permet de procéder en plusieurs passes et ainsi, de constituer au fur et à mesure la ressource.

Concernant l'encodage de la structure du document, la norme TEI-P5 propose une grande variété d'objets textuels dont certains dédiés à l'encodage des documents manuscrits[3] qui peuvent directement être appliqués aux textes d'élèves : les ratures (<del>), les ajouts (<add>) ou encore les éléments <zone> et <surface> qui peuvent être utilisés pour indiquer le collage de bouts de texte. Cet encodage qui repose sur le format XML facilite un grand nombre de traitements automatiques (calcul des spécificités lexicales, lemmatisation, etc.). L'exemple (1) illustre l'encodage selon la TEI-P5 d'un texte d'élève de CE2 issu du corpus « Charolles ». La transcription proposée ici ne vise que la digitalisation des données visibles. La normalisation de l'orthographe et de la syntaxe ne sont pas prises en compte ici : le texte d'élève est reproduit à l'identique.

(1)




```
Il aité une foi Mirabel ét sons frére, et san oublié sa maman.
La maman aité fatigé celai la nuis
Lere
Elle habitait dans cette maison depuis longtemps.
```

```
Il aité une foi Mirabel ét sons frére <del rend="strikethrough">, et san oublié sa maman</del>.<lb/><del rend="hatched">La</del><add place="below">Lere</add> maman aité fatigé cetai la nuis<lb/>
<surface attachment="glued">
<zone>Elle habitait dans cette maison depuis longtemps. </zone>
</surface>
```

Un dernier point concernant la présentation des textes à annoter préconise la mise à disposition systématique de la version

3 Voir la section 11 des *TEI Guidelines :* « Representation of Primary Sources »

originale du document (document pdf, copie d'élève, etc.). En effet, malgré le souci de proposer aux annotateurs une version à annoter proche des documents originaux, certains objets visuels comme les images et les tableaux ne peuvent être affichés dans un outil d'annotation. Cet effacement enlève un contenu essentiel et souvent structurant, ce qui peut biaiser l'annotation. Dans l'exemple précédent, la forme de la flèche et l'horizontalité des lignes ne peut être transcrite. La seule solution à ce jour est d'indiquer dans le guide d'annotation la nécessité de prendre connaissance de la version originale avant d'annoter.

2.2. Quels objets annoter et comment ? Définition du modèle d'annotation et importance des guides d'annotation

La définition du modèle d'annotation est une étape essentielle qui couvre à la fois la définition du type d'objet à annoter et des procédures envisagées pour les annoter. Afin de clarifier le modèle et de tester les procédures d'annotation, une étape d'annotation exploratoire est indispensable pour produire un guide utilisable. Pour ANNODIS, la phase exploratoire a consisté à faire appliquer par les membres du projet une version préliminaire du guide sur trois textes. Les guides d'annotation stabilisés à la suite de cette phase exploratoire donnent une description explicite des données annotées et des procédures à appliquer. La publication du guide est essentielle pour la diffusion d'une ressource annotée et pour permettre une reproductibilité de la campagne d'annotation.

La suite de cette section présente les différents types d'objets qu'il nous semble important de distinguer pour une annotation discursive, les procédures envisageables pour ces différents objets et un retour d'expérience sur la question du niveau d'expertise des annotateurs.

Différents types d'objets à annoter peuvent être distingués : des unités, des relations et des structures complexes que nous appellerons « schémas » en référence au terme utilisé dans GLOZZ. L'annotation des unités consiste à (1) délimiter des segments de texte et (2) associer à ces segments une catégorie (e.g. UDE). Les unités peuvent être de tailles très variables (e.g.

expression coréférentielle, item de structure énumérative). L'annotation des relations consiste à (1) tisser des liens entre des unités préalablement annotées (manuellement ou automatiquement) et (2) associer une catégorie à ce lien (e.g. relation temporelle). L'annotation des « schémas » consiste à regrouper des unités et/ou des relations dans une même structure. Ce niveau d'annotation est nécessaire pour représenter des structures complexes comme les structures énumératives qui sont composées d'unités de statuts différents (amorces, items, indices d'énumération).

L'annotation ANNODIS a impliqué deux procédures différentes. L'annotation des structures multi-échelles a consisté en une annotation ciblée lors de laquelle les annotateurs devaient (1) délimiter des unités définies dans le modèle et (2) regrouper ces unités dans des schémas. L'annotation en Relations de Discours a procédé, elle, en trois étapes : (1) segmentation des textes en UDE (annotation complète en UDE), (2) concertation entre annotateurs pour aboutir à une version segmentée stable des textes permettant en (3) l'annotation complète des Relations de Discours entre toutes les UDE ou groupes d'UDE. Deux sous-ensembles du corpus ont été ainsi annotés : annotation des structures multi-échelles dans les 87 textes longs (textes de plus de 1000 mots issus de Wikipédia, du CMLF 2008 et de l'IFRI) et annotation en Relations de Discours de 86 textes courts (textes d'environ 300 mots issus de l'*Est Républicain* et extraits de Wikipédia, du CMLF 2008 et de l'IFRI).

Le choix des procédures d'annotation est évidemment influencé par l'utilisation ou pas d'une interface d'annotation. En effet, l'annotation de relations et de schémas est difficilement envisageable sans un outil d'annotation spécifique. Cependant, toute annotation n'exige pas l'utilisation d'une interface spécifique, la production d'annotations par le biais d'outils de bureautique peut alors tout à fait être envisagée. D'ailleurs, l'annotation en UDE et en relations rhétoriques a pendant longtemps été réalisée par le biais d'un éditeur de texte (cf. les exemples donnés dans la section 3).

D'un point de vue plus technique, deux conseils ressortent de l'expérience ANNODIS. Premièrement, il apparaît préférable de réaliser des annotations dites « déportées », c'est-à-dire enregistrées dans un fichier différent du texte à annoter (par exemple en consignant les annotations dans un fichier texte à part ou une feuille de calcul), ce qui permet à la fois de ne pas toucher au texte en lui-même et d'envisager plusieurs couches d'annotation qui pourront par la suite être croisées. Deuxièmement, il est primordial de conserver une trace de la localisation dans le texte des annotations réalisées afin de permettre à la fois une visualisation du texte annoté et une analyse des indices linguistiques apparaissant dans le contexte des annotations. Une façon de localiser une annotation déportée est d'indiquer sa position *offset*, c'est-à-dire la position des caractères de début et de fin dans le texte. On comprend alors l'importance de ne pas « toucher au texte » pour ne pas « décaler » les annotations. L'intérêt des interfaces comme GLOZZ ou ANALEC (Victorri 2012) est qu'elles permettent à la fois de définir un modèle d'annotation et de réaliser des annotations déportées avec *offset* renseigné.

Pour finir, il est nécessaire de bien mesurer les différences en termes de procédures et résultats d'annotation selon que l'on a recours à des annotateurs « naïfs » ou « experts ». Là où les premiers fournissent un regard sur le caractère « intuitif » du modèle d'annotation, les seconds permettent une annotation plus fine et souvent plus riche. Si l'objectif de l'annotation est davantage de rassembler des occurrences d'un phénomène au contour encore flou pour permettre dans un second temps une analyse fine, l'annotation naïve permet quant à elle de récolter des données rapidement sans devoir au préalable définir formellement le phénomène à l'étude. Si l'objectif est d'éprouver un modèle préétabli, alors une annotation experte semble plus appropriée. Cependant, l'annotation naïve peut aussi être utilisée pour tester l'intuition des locuteurs sur le modèle et peut alors mettre en avant certaines difficultés du modèle notamment pas l'observation des désaccords inter-annotateurs.

Dans les sections suivantes, deux « chantiers d'annotation » sont présentés qui tirent tous les deux partie, à leur manière, de l'expérience ANNODIS : l'annotation, dans les textes d'élèves, des Relations de Discours et des chaînes de référence.

3. Vers une annotation des Relations de Discours

Nous décrivons dans cette section trois expériences d'annotation ou d'analyse exploratoires du corpus « Charolles». L'objectif est d'une part d'évaluer l'applicabilité du guide d'annotation en Relation de Discours développé pour ANNODIS à ce corpus et d'autre part d'établir une première cartographie des marques de cohésion impliquées dans l'expression de la cohérence temporelle. Ces expériences ont été menées sur un petit corpus de 12 textes d'élèves de primaire et de 6ème et de 6 textes d'étudiants de M2 de Sciences du Langage suivant un séminaire sur la sémantique et la pragmatique du discours. Nous distinguons ci-dessous les deux étapes de l'annotation : la segmentation en Unités de Discours Élémentaires (3.1) et la caractérisation des Relations de Discours (3.2), puis nous présentons en 3.2 un début d'analyse de l'utilisation des marques de cohésion dans les textes d'élèves.

3.1 Segmentation en Unités de Discours Élémentaires

Les 18 textes ont été segmentés sur la base de critères ponctuationnels, syntaxiques et sémantico-référentiels suivant les principes de segmentation décrits dans le guide d'annotation en Relation de Discours du projet ANNODIS (Muller et al., 2012) : un segment minimal (UDE) décrit un événement ou un état de fait, que celui-ci soit décrit par une proposition finie ou non, indépendante ou non, elliptique ou non. Certains éléments, comme les introducteurs de cadres (Charolles, 1997), sont segmentés parce qu'ils ont une autonomie discursive (Vieu et al. 2005). La segmentation a été réalisée par des annotateurs experts ou semi-experts (étudiants de M2). Nous donnons ci-dessous deux textes segmentés suivant ces principes. Les phrases à introduire figurent en italique, et sont segmentées en deux UDE pour deux d'entre elles.

(2) **Texte de Lucie (5° P)**

[La fillette et le bandit]$_1$

[Il était une fois une fillette de 11 ans qui vivait avec sa mère dans une petite maison à la lisière d'une forêt.]$_2$ [Elle habitait dans cette maison depuis longtemps.]$_3$

[La fillette aimait beaucoup aller se balader dans la forêt.]$_4$ [Mais sa mère avait très peur de cette forêt]$_5$ [car on la surnomait « La forêt du bandit ».]$_6$

[Une nuit,]$_7$ [comme sa mère refusait toujours aussi catégoriquement de la laisser aller dans la forêt,]$_8$ [la petite fille sortit pour aller dans la forêt.]$_9$ [Elle se baladait tranquillement]$_{10}$ [quand elle entendit un fin bruit de pas.]$_{11}$ [Elle regarda derrière elle]$_{12}$ [mais ne vit rien.]$_{13}$ [Le bruit s'était arrêter.]$_{14}$ [Soudain,]$_{15}$ [un jeune homme vêtu de noir et armé d'un pistolet apparu devant elle.]$_{16}$ [Elle se mit à courir à toute vitesse.]$_{17}$ [L'homme s'arrêta]$_{18}$ [car la fillette avait disparu de son champ de vision.]$_{19}$ [La petite fille jeta un gros caillou dans la rivière.]$_{20}$

[Il se retourna]$_{21}$ [en entendant ce grand bruit.]$_{22}$ [Le temps que l'homme sorte de sa torpeur,]$_{23}$ [la petite fille courut chez elle]$_{24}$ [et se couchat rapidement.]$_{25}$

[Depuis cette aventure,]$_{26}$ [les enfants ne sortent plus la nuit.]$_{27}$

(3) **Texte de Yassim (5° P)**

[Il était une fois une ~~petite~~ fille qui ~~s'appeller~~ appellait Laura]$_1$ [qui voulez se rendre ~~allez~~ à The Voice pour chanter.]$_2$

[Sur le chemin,]$_3$ [elle voit son ancinne maison. ~~ou elle habitait~~.]$_4$

[Elle habitait dans cette maison depuis longtemps.]$_5$

[Elle entandit un bruit]$_6$ [elle se retourna ~~tou~~]$_7$ [c'été tout simplement un voiture qui venait de faire un accident.]$_8$ [Son papa qui l'accompagnier eu très peur.]$_9$ [Elle se retourna]$_{10}$ [en entendant ce grand bruit.]$_{11}$

[Puis arriver à The Voice pour chanter devant le public ;]$_{12}$ [Auretour]$_{13}$ [son père lui fait vraiment peur.]$_{14}$ [Elle lui pardonneras plus jamais.]$_{15}$ [Sur le chemain du retour,]$_{16}$ [elle lui parle pas]$_{17}$ [dès est arriver chez ~~elle pour sortir~~]$_{18}$ [Elle lui demande aux enfants de sortir.]$_{19}$

[Le ~~coucher du soleil~~ soleil se couche]$_{20}$ [et ils sont au plein milieu d'une forêt]$_{21}$ [et un loup-garou arrive.]$_{22}$ [Les enfants et Laura sont arriver à la maison vivant.]$_{23}$

[Depuis cette aventure,]$_{24}$ [les enfants ne sortent plus la nuit.]$_{25}$

Cette première expérience de segmentation de textes du corpus « Charolles » montre qu'on convoque préférentiellement le critère sémantico-référentiel pour segmenter les écrits des élèves de primaire, alors que pour les textes d'étudiants, le recours aux trois critères – ponctuationnel, syntaxique et sémantico-référentiel – est plus équilibré.

La raison en est qu'en segmentant, l'annotateur expert rétablit la ponctuation et les structures syntaxiques normées, invalidant de ce fait les critères concernés au profit du critère sémantico-référentiel. Cette attitude peut revenir, dans certains cas, à ne pas tenir compte de la segmentation ponctuationnelle ou syntaxique mise en œuvre par les élèves. Le fait de ne pas les prendre en compte est un élément de « normalisation » dans le traitement. Cependant, grâce à la méthode d'annotation déportée choisie (cf. section 2), nous n'opérons aucune modification des marques de surface, et nous rendons possible la confrontation de la segmentation en UDE avec une segmentation selon d'autres unités (phrase ponctuationnelle, phrase syntaxique, etc.) pour d'autres chercheurs s'intéressant aux autres dimensions.

Concernant les textes d'étudiants, il faut noter qu'ils révèlent parfois des conflits entre les trois critères, ce qui rend la tâche de segmentation moins facile, mais que globalement, sur l'ensemble des 18 textes du corpus, la tâche de segmentation a été parfaitement réalisable en suivant le guide.

3.2 Analyse de la cohérence temporelle

Dans (Garcia-Debanc et Bras 2016), nous avons mené une analyse exploratoire d'un premier échantillon de 6 textes ainsi segmentés. Nous avons montré que l'inventaire des marqueurs de cohésion et l'analyse de leur variété ainsi que de leur nature permet de faire une première évaluation du degré de maitrise de la cohérence temporelle. Les deux textes donnés sous (2) et (3) permettent d'illustrer cette corrélation.

Le texte de Lucie, sous (2), montre une très bonne maitrise des temps du récit (passé simple, imparfait, plus-que-parfait) et du maniement des outils de connexion. Le seul

introducteur de cadre utilisé, *une nuit* au segment 7, l'est à bon escient puisqu'il introduit l'étape de la complication du récit, et confère ainsi à l'ensemble la structuration décor/complication. On note aussi l'emploi des trois connecteurs *mais, car, soudain*, argumentatif, causal et narratif, ainsi que de trois subordonnées causales ou temporelles (notamment avec un quand inverse en 11).

Le texte de Yassim, en (3), en revanche, présente un emploi non maitrisé des temps verbaux : enchainement imparfait, présent, imparfait en 1-5 ; séquence narrative incomplète au présent/futur en 12-23 terminée par une phrase au passé composé. On note également l'absence de subordonnées temporelles, la présence de connecteurs narratifs *puis et et* et l'emploi de trois cadratifs spatio-temporels *sur le chemin, au retour, sur le chemin du retour* qui ne structurent pas véritablement le texte. Enfin, ce texte donne à voir une belle illustration du fait que la présence de marqueurs de cohésion ne suffit pas à garantir pas la cohérence (Charolles 1995) : à première vue, la résolution des anaphores *elle* et *cette maison* du segment 5 est réalisable, puisque des référents de discours de type « personnage féminin » et « maison » sont saillants dans le segment 4. Au plan temporel, le segment 5 (la première phrase à introduire) ne constitue pas une continuation de la description du décor, comme on a pu le voir dans le texte de Lucie, mais une sorte de mise en arrière-plan par rapport à l'événement décrit par le segment 4, ce qui est parfaitement possible. Mais cette mise en relation bute sur une incohérence sémantique entre les contenus propositionnels des segments 4 et 5 : l'emploi de l'adjectif *ancienne* dans *son ancienne maison* implique que la maison désignée n'est plus la maison de la protagoniste, ce qui entre en contradiction avec le sens du syntagme verbal *habiter dans cette maison depuis longtemps* en 5, qui implique que la protagoniste habite encore cette maison au moment de l'événement qui sert de repère (celui qui est décrit en 4).

L'analyse du nombre de marqueurs présents dans l'échantillon de textes et de leur type permet de montrer que l'emploi de subordonnées temporelles est directement corrélé

avec une bonne maitrise de la cohérence temporelle (cas du texte de Lucie). Il permet également de confirmer l'observation de Fayol (1986) sur les connecteurs dans les récits : l'emploi de connecteurs explicites de relations causales ou contrastives (*car, donc, mais*) reflète une meilleure maitrise de la cohérence – cf. texte de Lucie, que l'emploi de connecteurs marquant une simple énumération ou succession temporelle (*et, et puis, puis*) – cf. texte de Yassim.

3.3 Annotation des Relations de Discours

Une première expérience d'annotation du corpus « Charolles » en Relations de Discours a été menée dans le cadre d'un séminaire de M2 consacré à la sémantique et à la pragmatique du discours (Bras 2015-2016). Chaque étudiant a été invité à produire un texte selon la consigne puis à analyser un corpus de textes d'élèves. En fin de séminaire, un travail de groupe a abouti à la segmentation et à l'annotation en Relations de Discours d'un corpus mixte comportant deux textes d'élèves et deux textes d'étudiants. Outre l'intérêt que représente l'analyse de textes dont la cohérence n'est pas optimale pour comprendre les ressorts de la cohérence, cette expérience a permis une triple annotation des textes par des annotateurs semi-experts suivie d'une phase de recherche d'accord pour aboutir à une annotation de référence, permettant aux étudiants de mieux comprendre le sens et le rôle des Relations de Discours et en même temps de faire des hypothèses sur les intentions de communication des scripteurs des textes. Nous ne pouvons donner ici une description détaillée des résultats de cette expérience mais nous pouvons déjà indiquer le succès de l'entreprise :

Sur le plan empirique, d'une part, le guide d'annotation en Relations de Discours s'est montré parfaitement applicable aux textes du corpus, même s'il sera utile d'en rédiger une version adaptée pour l'annotation du corpus « Charolles » dans son ensemble.

Sur le plan théorique, d'autre part, la tâche d'analyse sous-jacente a permis de mettre à l'épreuve le cadre théorique de la SDRT sous-jacent à la tâche d'annotation, et notamment

l'aspect prédictif de la théorie. A titre d'exemple, nous pouvons mentionner la prédiction correcte d'un défaut de cohérence dans le texte de Yassim au niveau de la mise en relation du segment 5 avec son contexte gauche, mais la difficulté à traiter des segments répétés comme 7 et 10, ou la nécessité d'une analyse spécifique des segments initiaux (segment 2 du texte de Lucie, segments 1 et 2 du texte de Yassim).

De manière générale, l'expérience a montré que l'analyse de ces textes pourrait nourrir de façon pertinente un double processus de validation empirique et d'évolution des théories et des modèles de la cohérence du discours.

4. Vers une annotation des chaînes de référence

Le modèle d'annotation en structures multi-échelles du projet ANNODIS 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 de référence (cf. Ho-Dac & Péry-Woodley, 2014, p.2649). Le motif qui nous intéresse ici est le second dans la mesure où la consigne d'écriture proposée aux élèves contient des expressions référentielles qui exigent d'être introduites et peuvent s'inscrire dans des chaînes (cf. section 1).

4.1 Objectifs de l'annotation des chaînes de référence dans l'expérience ANNODIS

Dans le guide d'annotation en structures multi-échelles (Colléter et alii, 2012), une chaîne de référence est définie comme un type particulier de chaîne de cohésion où les éléments sont des unités contenant un même référent occupant la position sujet, et où le segment résultant est constitué de l'ensemble des unités connectées. Avant de pouvoir procéder à l'annotation proprement dite d'une chaîne, l'annotateur – qui, dans l'expérience ANNODIS, est un annotateur naïf (étudiant en troisième année de licence de Sciences du Langage) – doit identifier, dans les vastes corpus de textes qui lui sont soumis, des zones présentant une forte cohérence référentielle. Cette

tâche est facilitée par le pré-marquage automatique de formes théoriquement susceptibles d'orienter vers ces zones. Il s'agit de formes décrites dans la littérature comme des marques discursives participant à la signalisation de la structure à l'étude : pronoms personnels clitiques de 3e personne, pronoms démonstratifs, syntagmes nominaux comportant un déterminant possessif ou démonstratif, syntagmes nominaux ayant à leur tête un nom déjà présent dans un autre syntagme de la section ou figurant dans le titre de la section (dans le cas des sections titrées). Afin de ne pas surcharger le texte d'indices pré-marqués et de permettre de cibler des zones contenant des structures de haut-niveau, seuls les indices apparaissant en position initiale de phrase ont été pris en compte. La visualisation de ces indices par le biais de l'interface GLOZZ permet aux annotateurs de survoler le texte en ciblant des zones textuelles présentant une forte concentration des indices potentiels. Une fois une zone de forte concentration identifiée, la tâche d'annotation consiste d'abord à délimiter l'unité « chaîne », puis à indiquer les indices de surface permettant la reconnaissance des composants de la chaîne et enfin à réunir ces différents éléments au sein d'un même schéma. Au total 588 chaînes et 3456 expressions coréférentielles ont été annotées. L'extrait ci-dessous fournit un exemple d'une chaîne annotée où les indices annotés figurent en gras :

(4) **Exemple de chaîne topicale annotée dans la ressource ANNODIS**

[**César** désigna dans son testament trois héritiers, les petits-fils de ses soeurs, à savoir Octave, Lucius Pinarius Scarpus et Quintus Pedius. **Il** légua les trois quarts de son héritage au premier et le quart restant aux deux autres. Dans la dernière clause de son testament, **César** adopta Octave, le futur empereur Auguste, et lui donna son nom. Enfin, **il** légua au peuple romain ses jardins près du Tibre et trois cents sesterces par tête.]

Notons que les nombreux syntagmes comportant des déterminants possessifs n'ont pas été annotés et ce bien qu'ils participent à la chaîne, contribuant à la cohérence référentielle du passage. Ce défaut de l'annotation s'explique par les objectifs et la procédure d'annotation retenue. Dans l'expérience ANNODIS,

en effet, l'objectif de l'annotation des chaînes est avant tout de fournir au linguiste une grande quantité de portions de textes organisées autour d'un même référent. La tâche des annotateurs (non experts), guidés par des indices de surface, consiste donc principalement à délimiter les passages potentiellement intéressants, revient ensuite au linguiste l'analyse comparée des types de chaînage, du nombre de maillons, etc., en tenant compte des types de textes. Nous allons voir maintenant que bien que l'objet soit le même, les objectifs de l'annotation des chaînes de référence dans les textes d'élèves sont assez différents. L'annotation des chaînes de référence dans les textes scolaires (issus du corpus « Charolles ») nous semble toutefois pouvoir largement bénéficier de l'expérience d'annotation de ce même phénomène dans ANNODIS.

4.2 Objectifs de l'annotation des chaînes de référence dans les textes d'élèves

Comme nous venons de le souligner, dans ANNODIS, l'objectif de l'annotation, dans un très vaste volume de textes, de zones textuelles fortement cohérentes du point de vue référentiel est de fournir au linguiste un matériau riche permettant une analyse fine des types de chaînes en tenant compte notamment du genre de discours – une telle analyse exigeant un autre niveau d'annotation qui intervient dans un second temps. C'est précisément à ce niveau que se situe l'objectif de l'annotation des textes d'élèves. Dans la mesure où l'objectif est d'élaborer une cartographie des moyens linguistiques choisis par l'élève pour introduire un nouveau référent et pour le reprendre dans son discours, il ne s'agit pas de repérer des zones dans un corpus mais d'annoter pour chaque expression référentielle les marqueurs utilisés par chaque élève (*elle, il, cette maison, ce grand bruit, cette aventure, les enfants*).

Plus précisément, la tâche d'annotation des chaînes dans les textes produits par les élèves selon la consigne présentée dans la section 1 consiste à identifier les expressions référentielles donnant accès aux référents contenus dans les trois phrases proposées et à lister, le cas échéant, les moyens utilisés pour

construire des chaînes de référence. Selon les niveaux scolaires, les textes contiennent des expressions référentielles variées : en première mention, les syntagmes nominaux sont principalement réalisés sous la forme de noms propres (référant alors de manière directe et non ambiguë à leur référent). Dans les chaînes, on trouve des syntagmes nominaux définis ou possessifs ou encore des pronoms personnels de 3^{e} personne. Dans le texte de Lucie (cf. exemple (3) plus haut), on constate la grande homogénéité référentielle du texte produit. Le personnage principal est le personnage féminin qui apparaît dans la première phrase de la consigne (après avoir été annoncé dans le titre) : la chaîne comprend un grand nombre de maillons et est réalisée par un ensemble d'expressions coréférentielles variées : les syntagmes nominaux sujet (*la fillette, la petite fille, elle*) se combinent avec les syntagmes possessifs (*sa mère*),

L'annotation requise est donc une annotation experte : les annotateurs sont des linguistes (cf. thèse de Karine Bonnemaison, en cours). Comme le dit Charolles (2014, p.93), ce type de tâche d'annotation « ne peut à l'évidence être confiée qu'à des experts disposant d'une solide formation linguistique dans le domaine et d'un manuel d'annotation aussi précis et exhaustif que possible pour qu'ils puissent faire face aux innombrables difficultés qu'ils ne manqueront pas de rencontrer ». En effet, dans les textes d'élèves comme dans les œuvres littéraires étudiées par Charolles et ses collègues – mais probablement pour des raisons bien différentes - il arrive très souvent que l'on ne voie pas clairement « à quel référent introduit ou repris dans le discours précédent ou suivant peut renvoyer une expression anaphorique » (Charolles, 2014, p.55). Dans les textes d'élèves, on peut s'attendre à ce que les difficultés soient nombreuses : il y aura certainement de nombreuses situations où il faudra choisir une interprétation. Dans ces situations, l'annotateur devra se demander ce que l'élève a voulu dire, comprendre sa logique. En (3), on peut, par exemple, se demander comment s'établit le lien entre le syntagme nominal défini du titre, *le bandit*, et la chaîne composée de deux maillons (*un jeune homme vêtu de noir et armé d'un pistolet, L'homme*)

qui sert d'antécédent au pronom *il* de la deuxième phrase de la consigne.

Un autre objectif de l'annotation est de mettre à l'épreuve les très nombreuses études linguistiques qui se sont intéressées à l'interprétation des expressions référentielles et anaphoriques à des écrits non normés d'apprenants. Il s'agit de confronter ces écrits spécifiques aux descriptions théoriques des principaux facteurs pouvant peser sur le choix et l'interprétation de telle ou telle expression référentielle en fonction du contexte dans lequel elle apparaît. Les anaphores et les chaînes de coréférence doivent donc annotées (et enregistrées dans un fichier à part, une base de données ou une feuille de calcul, cf. section 2.2) selon leur composition : type de syntagme composant chacun des maillons (noms propres, pronoms, syntagmes définis, démonstratifs ou possessifs, etc.), nombre de maillons (dans le cas d'une anaphore, il y aura deux maillons seulement dont l'un est composé par l'expression référentielle se trouvant dans la phrase à insérer), distance entre les maillons et fonction syntaxique des syntagmes composant les maillons.

On peut d'ores et déjà lister deux grands types de questions auxquelles l'annotation des chaînes de coréférence permettra d'apporter des éléments de réponse (à la fois d'un point de vue qualitatif et quantitatif) :

Quelle est la forme des expressions référentielles utilisées par les élèves ? On se demandera notamment comment est introduit, dans le texte rédigé par l'élève, le référent du pronom personnel anaphorique de la première phrase : *elle*. Le référent de ce premier personnage du récit à construire est-il introduit par un nom propre ? Comment sont introduits les référents des syntagmes nominaux démonstratifs : *cette maison*, *ce grand bruit*, *cette aventure* (ce dernier syntagme étant résomptif). Tous les élèves utilisent-ils les mêmes stratégies ? Y a-t-il des différences selon le niveau scolaire ?

Quelles sont les relations entre les expressions référentielles (anaphores, chaînes de référence) ? Pour parler de chaînes, nous considèrerons au moins trois expressions

coréférentielles, suivant en cela Schnedecker, Longo (2012) selon lesquelles en deçà de ce nombre la notion de chaîne n'est pas pertinente puisque celle de coréférence ou d'anaphore suffit à rendre compte de la relation entre les expressions (cf. Corblin, 1995 ; Schnedecker, 1997).

5. Conclusion

L'objectif de la constitution de grands corpus d'écrits scolaires a été rappelé en introduction. Il s'agit principalement de faire progresser les connaissances sur la mise en place des compétences rédactionnelles des élèves par l'étude d'un grand nombre de productions écrites diversifiées. Seul l'examen d'un nombre important de textes d'élèves permet en effet de déterminer des indicateurs de complexification permettant d'évaluer les compétences rédactionnelles d'élèves de niveaux scolaires différents.

L'objectif de l'annotation discursive de ces corpus, quant à lui, est double : il s'agit non seulement de rendre possibles des analyses qualitatives, en confrontant des modèles théoriques existants à un nouveau type de données ; mais aussi (et peut-être surtout) quantitatives, en mesurant les indicateurs de compétences rédactionnelles pertinents. Au final, la mise en regard des différentes annotations discursives devrait permettre de dessiner une cartographie de l'acquisition des compétences discursives et textuelles des rédacteurs des textes et de croiser ces indicateurs avec ceux qui relèvent d'autres niveaux d'analyse linguistique, tels que l'orthographe ou la syntaxe.

La mise au point d'une méthodologie pour l'annotation discursive des textes d'élèves répond donc à des enjeux scientifiques à la fois en analyse du discours et en didactique du français langue première. Nous avons montré comment des données éloignées de la norme comme le sont des textes de jeunes élèves posent des questions aux modèles linguistiques qui s'efforcent de rendre compte de la cohésion. En didactique du français langue première, une meilleure connaissance des moyens linguistiques mis en jeu par les élèves des différents niveaux scolaires pour assurer la cohésion textuelle par une étude

systématique d'un ensemble important de textes peut ainsi conduire à la formulation d'indicateurs de progressivité (Nonnon, 2010). L'annotation discursive des grands corpus permet en effet de recenser les moyens linguistiques utilisés selon les niveaux scolaires de façon isolée ou combinés et d'observer l'évolution de leur mise en place tout au long du curriculum. Toutefois, de premières analyses ont montré que l'expertise rédactionnelle n'est pas uniquement liée au niveau scolaire (Garcia-Debanc, 2010) et que les marqueurs linguistiques doivent être traités conjointement en faisceau d'indicateurs de maitrise de la cohésion textuelle. L'annotation de grands corpus devrait permettre de confirmer ou de nuancer ces premières observations. La détermination du degré d'expertise de chaque élève peut s'appuyer sur ces annotations : elle suppose que soient mis en regard les moyens linguistiques utilisés par chaque rédacteur pour la résolution des différents problèmes de cohésion textuelle. L'enseignant dispose ainsi d'éléments permettant de caractériser le degré de maitrise de la cohésion textuelle pour chaque rédacteur et de l'aider à progresser en diversifiant les moyens linguistiques mobilisés. L'élaboration de ces indicateurs de progressivité doit être réalisée avec prudence : la maitrise de la cohésion textuelle ne se juge pas au nombre de marques linguistiques présentes dans un texte. En revanche, la présence de certaines structures - telles que la subordination inverse en *quand* ou les cadratifs spatio-temporels - est généralement assortie de la maitrise d'autres marques de cohésion. On peut alors utiliser ces résultats pour déterminer des indicateurs qui seraient particulièrement discriminants pour définir un degré d'expertise du rédacteur. De tels éléments peuvent être précieux pour proposer aux enseignants des repères pour l'évaluation des compétences de leurs élèves. Le repérage de ces indicateurs modifie le regard sur les textes d'élèves, désormais analysés comme moments d'un développement plus que comme comportant des erreurs (Roubaud, Garcia-Debanc, 2014).

Les annotations réalisées dans le cadre d'ANNODIS ainsi que les premiers chantiers d'annotation portant sur les textes

d'élèves nous permettent de dégager un certain nombre de principes de travail.

L'annotation des textes doit porter sur la version initiale telle qu'elle a été produite par l'élève. On ne procèdera à aucune normalisation syntaxique, orthographique ou ponctuationnelle, dans la mesure où ces éléments ne sont pas directement pris en compte dans les analyses. On fera en sorte qu'il soit toujours possible de revenir à tout moment au texte original, notamment pour apprécier la mise en espace du texte ou les différentes ratures, traces de son élaboration.

Les expériences de segmentation et d'annotation de textes d'élèves en Relations de Discours nous ont permis de valider la transférabilité effective du guide d'annotation d'ANNODIS, dans la mesure où la procédure d'annotation – partant de la segmentation en UDE vers une structure complexe formée de segments simples ou complexes reliées entre eux par des Relations de Discours – est applicable dans les deux cas, et où les textes sont de longueurs comparables.

S'agissant de l'annotation des chaînes de référence, le guide et les procédures mises en œuvre pour ANNODIS ne peuvent pas être réutilisées tels quels et nécessitent d'être complétés. L'analyse linguistique des annotations réalisées par les annotateurs naïfs n'est pas achevée, mais, dans la mesure où on dispose de 588 chaînes et de 3 456 expressions coréférentielles annotées, on peut raisonnablement penser qu'une fois l'analyse terminée, elle viendra nourrir les travaux déjà disponibles sur les chaînes d'expressions référentielles, leurs compositions, leurs variations en fonction du genre de discours. Les marqueurs linguistiques utilisés par les élèves pourront ainsi être confrontés aux procédures utilisées par les experts.

De même que le projet ANNODIS a permis de confronter des modes d'annotation discursive différents et complémentaires, la mutualisation des corpus entre chercheurs en didactique du français s'intéressant à des niveaux d'analyse différents (lexique, syntaxe, cohésion textuelle, structuration temporelle) est essentielle pour essayer de croiser ces indicateurs

en vue de définir des indicateurs pertinents de compétences textuelles.

Références bibliographiques

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

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

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

Aurnague, M., Garcia-Debanc, C. (2016). « Les verbes de déplacement comme contenu d'enseignement du lexique à l'école primaire : modélisation sémantique et analyse de productions d'élèves », *Congrès Mondial de Linguistique Française (CMLF 2016)*.

Andersen, H. L., Birkelund, M., Leblay C., Auriac-Slusarczyk (2010). « Acquisition et enseignement en production écrite. » *Synergies Pays Scandinaves*, Numéro 5 / Année 2010, Revue du Gerflint, Aarhus, Danemark.

Bonnet C., Corblin C., Elalouf M.-L. (1998). *Les procédés d'écriture chez les élèves de 10 à 13 ans, un stade de développement*, Lausanne, CVRP.

Boré C., Elalouf M.L. (2007). « Construction et exploitation de corpus d'écrits scolaires », *Revue Française de Linguistique Appliquée*, XII, 1, « Corpus : état des lieux et perspectives » : 53-70.

Boré C. (2011). » L'énonciation des brouillons et la question du sujet scolaire », *Pratiques,* 149-150 : 71-90.

Bras, M. (2015-2016). « Discours : sémantique et pragmatique », cours de M2, Master Sciences du Langage, Université Toulouse Jean-Jaurès.

Cappeau P., Roubaud M.-N. (2005). *Enseigner les outils de la langue avec les productions d'élèves*. Paris : Bordas.

Charolles, M. (2014). « Annotation des expressions référentielles et profondeur de traitement », *in* M. Fossard, M.-J. Béguelin (éds.) *Nouvelles perspectives sur l'anaphore. Points de vue linguistique, psycholinguistique et acquisitionnel*. Collection : Sciences pour la communication - volume 111. Peter Lang : Bern, Berlin, Bruxelles, Frankfurt am Main, New York, Oxford, Wien, 55-98.

Charolles, M. (1997). « L'encadrement du discours - univers, champs, domaines et espace », *Cahiers de recherche linguistique*, 6 : 1-73.

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

Charolles, M. (1988a). « Les plans d'organisation textuelle : périodes, chaines, portées et séquences », *Pratiques*, 58 : 3-13.

Charolles, M. (1988b). « La gestion des risques de confusion entre personnages dans une tâche rédactionnelle » *Pratiques*, 60 : 75-97.

Charolles, M. (1978). « Introduction aux problèmes de la cohérence des textes », *Langue Française*, 38, 7-41.

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

Elalouf M.-L., Boré C. et alii. 2005. *Ecrire entre 10 et 14 ans. Un corpus, des analyses, des repères pour la formation*, Paris, Scérén, CRDP Versailles, CDDP Essonne.

Elalouf M.-L. (2011). « Constitution de corpus scolaires et universitaires : vers un changement d'échelle », *Pratiques* 149-150 : 56-70.

Fabre-Cols C. et alii (2000). *Apprendre à lire des textes d'enfants*. Bruxelles : De Boeck-Duculot.

Fayol, M. (1986). « Les connecteurs dans les récits écrits », *Pratiques,* 49, 101-113.

Garcia-Debanc C. (2010). « Segmentation, connexion et indexation dans des productions écrites d'élèves de 9 à 13 ans de deux genres textuels. » *Synergies Pays Scandinaves* n° 5 : 81-96.

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

Garcia-Debanc C., Bras M. (2016). « Vers une cartographie des compétences de cohérence et de cohésion textuelle dans une tâche-problème de production écrite réalisée par des élèves de 9 à 12 ans : indicateurs de maitrise et progressivité ». In S. Plane, C. Bazerman, F. Rondelli, C. Donahue et al. (eds). *Recherches en écriture : regards pluriels*, Actes du Colloque Writing Research Across Boarders 2014, *Recherches textuelles* n° 13, Metz, Université de Lorraine : 39-62.

Garcia-Debanc C., Gangneux, M. (2015) « L'Enseignement de la synonymie à l'école primaire. État des lieux et recherches innovantes pour une articulation entre enseignement du lexique et production écrite ». *Etudes de linguistique appliquée*, 178 : 143-164.

Geoffre, T. (2014). « Profils d'acquisition de la morphographie au cycle 3. Vers une caractérisation des parcours d'élèves ?», *Repères*, 49 : 147-168.

Halliday, M.A.K. (1977). « Text as Semantic Choice in Social Contexts ». In T. Van Dijk & J. Petöfi (eds), *Grammars and Descriptions*. Berlin: Walter de Gruyter, 176-226.

Hayes, JR, Flower, L.S (1980). « Identifying the organization of writing processes », *in* Gregg, L.W, Steinberg, E.R. (eds), *Cognitive processes in writing*. N. J : Hillsdale, LEA, 3-30.

Ho-Dac L.-M., Péry-Woodley M.-P. (2014). « Annotation des structures discursives : l'expérience ANNODIS. » In Franck Neveu, Peter Blumenthal, Linda Hriba, Annette Gerstenberg, Judith Meinschaefer, Sophie Prévost (Eds) *4e*

Congrès Mondial de Linguistique Française (CMLF 2014), Juillet 2014, Berlin, Allemagne, 2647-2661.

Ho-Dac, L-M., Fabre, C., Péry-Woodley, M-P., Rebeyrolle, J., Tanguy, L. (2012). « An empirical approach to the signalling of enumerative structures », *Discours. Revue de linguistique, psycholinguistique et informatique*, 10. <http://discours.revues.org/8611>

Masseron C. (2005). « Indicateurs langagiers et stratégies scripturales. Du discours à la langue », *Pratiques*, 125-126 : 205-249.

Mélanie-Becquet F., Landragin F. (2014). « Linguistique outillée pour l'étude des chaînes de référence : questions méthodologiques et solutions techniques », *Langages*, 195 : 117-137.

Muller, P., Vergez-Couret, M., Prévot, L., Asher, N., Benamara, F., Bras, M., Le Draoulec, A., Vieu, L. (2012). « Manuel d'annotation en Relations de Discours du projet ANNODIS », *Carnets de Grammaire*, 21, rapport interne CLLE-ERSS.

Nonnon É. (2010). « La notion de progression au cœur des tensions de l'activité d'enseignement », *Repères*, 41 : 5-34.

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.

Rebeyrolle, J., Péry-Woodley, M.-P. (2014). « Énumération et structuration discursive. In Actes *du 4e Congrès Mondial de Linguistique Française (CMLF 2014),* Juillet 2014, Berlin, Allemagne, 3183-3196.

Roubaud M.-N, Garcia-Debanc C. (2014). « L'approche d'"anomalies" dans des textes narratifs d'élèves de fin d'école primaire (10-11 ans). Quelques pistes pour la lecture des textes par les enseignants », *in* Avanzi M. *et alii* (eds.). *Enseignement du français : les apports de la recherche en linguistique*, Peter Lang, Bruxelles-Bern, 307-326.

Schnedecker, C. (1997). *Noms propres et chaînes de référence*. Université de Metz : Recherches Linguistiques.

TEI Guidelines Version 2.9.1. Last updated on 15th October 2015, revision 46ac023. This page generated on 2015-10-15T20:09:00Z.

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

Victorri, B. (2012). *ANALEC : logiciel d'annotation et d'analyse de corpus écrits.* Logiciel téléchargeable sur le site Web du laboratoire Lattice, http://www.lattice.cnrs.fr/ANALEC

Vieu, L, Bras, M., Asher, N., Aurnague, M. (2005). « Locating Adverbials in Discourse ». *Journal of Language Studies*, 15 : 173-193.

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