Vers un corpus de textes d'élèves annoté en relations de discours

Myriam Bras, Laure Vieu, Maëlle Joret, Audrey Pépin-Boutin, Clamença Poujade, Charlotte Roze

Vers un corpus de textes d'élèves annoté en relations de discours

Towards a corpus of learner texts annotated with discourse relations[1]

Myriam Bras

CLLE, Université de Toulouse, CNRS

Laure Vieu
IRIT, CNRS, Université de Toulouse

Maëlle Joret, Audrey Pépin-Boutin, Clamença Poujade, Charlotte Roze

CLLE, Université de Toulouse, CNRS

Résumé

Cet article présente le processus d'annotation en relations de discours (RD) de textes d'élèves du corpus Résolco (Garcia-Debanc *et al.* 2017) produits selon une même consigne d'écriture. Nous procédons à une segmentation en Unités de Discours Elémentaires qui sont ensuite reliées entre elles par des RD lors de l'annotation. Le jeu de RD choisi est proche de celui de la Segmented Discourse Representation Theory (Asher & Lascarides 2003) qui offre une méthode opératoire de construction de représentations de discours cohérents. Elle est mise ici à l'épreuve pour la première fois sur des textes d'apprenants et étendue pour l'annotation de l'incohérence.

Mots-clés Cohérence textes d'élèves, Relations de Discours, SDRT, Segmentation, Corpus annoté

Abstract

This paper presents the process of annotating with discourse relations (DRs) student texts from the Résolco corpus (Garcia-Debanc *et al.* 2017), produced according to the same writing instruction. We first perform a segmentation into Elementary Discourse Units and then link them together with DRs during the annotation process. The set of DRs is similar to that of Segmented Discourse Representation Theory (Asher & Lascarides 2003), which offers an operational method for building representations of coherent texts. This theory is tested here for the first time on learner texts and extended to annotate incoherence.

Keywords Discourse Coherence in learner texts, Discourse Relations, SDRT, Segmentation, Annotated corpus

[1] Ce travail a été réalisé dans le cadre du projet Écriture scolaire et universitaire : Corpus, Analyses Linguistiques, Modélisations didactiques (E-CALM), projet ANR-17-CE28-0004 du programme Sociétés innovantes, intégrantes et adaptatives (DS08) 2017, Janvier 2018-Juin 2021, responsable Claire Doquet.

1. INTRODUCTION

Nous abordons dans cet article la question de la cohérence des textes du point de vue du récepteur, c'est-à-dire de celui qui cherche à comprendre un texte. Nous adoptons une approche représentationnelle dans laquelle l'interprétation d'un texte repose sur la possibilité d'en construire une représentation de type logique, comme dans la Discourse Representation Theory (DRT de Kamp & Reyle 1993) ou son extension, la Segmented Discourse Representation Theory (SDRT de Asher & Lascarides 2003) qui définit de façon formelle ce qu'est un discours cohérent, et que nous adoptons ici.

Nous nous plaçons ainsi dans la continuité de travaux proposant d'analyser la cohérence d'un texte en identifiant des relations de discours (comme Narration, Elaboration, Résultat, Explication, Contraste) entre segments de ce texte (Hobbs 1985, Mann &Thompson 1987, Kehler 2002, Asher &Lascarides 2003).

Dans ces approches, l'analyse de la cohérence porte généralement sur des textes attestés, et prend généralement pour objet des textes rédigés par des scripteurs experts. En témoignent les premiers corpus de textes annotés en relations de cohérence comme le Penn Discourse Tree Bank (Prasad *et al.* 2008) pour l'anglais ou le corpus ANNODIS (Afantenos *et al.* 2012) pour le français. Or, l'étude de la cohérence chez des scripteurs dont la compétence rédactionnelle est encore en cours d'acquisition, à la croisée de la didactique du français et de la linguistique, révèle un point de vue très intéressant sur la cohérence (Charolles 1978, Rondelli 2010).

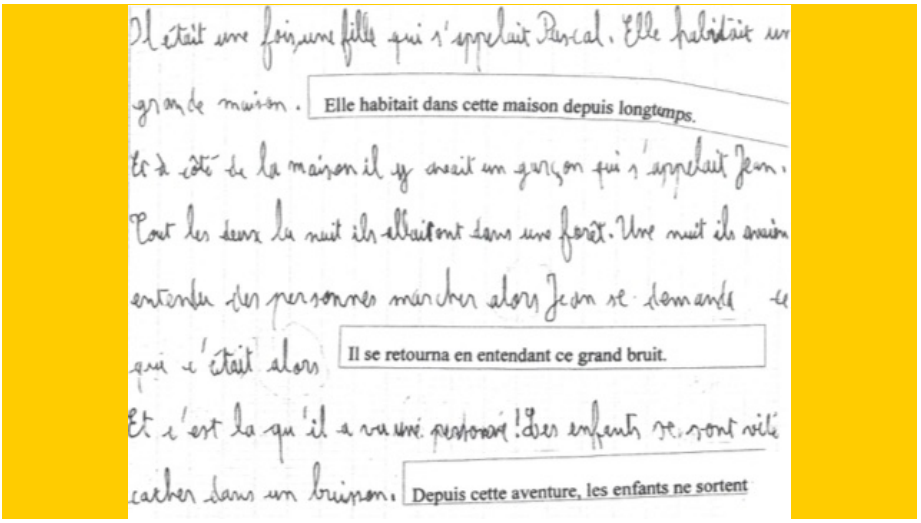
Les travaux que nous présentons ici s'inscrivent dans le cadre du projet ANR *Écriture scolaire et universitaire : Corpus, Analyses Linguistiques, Modélisations didactiques* (E-CALM). L'objectif de cet article est de présenter le processus d'annotation en relations de discours de textes d'élèves d'école primaire et de collège issus du corpus Résolco (Garcia-Debanc *et al.* 2017). Il s'agit de textes produits selon une même consigne d'écriture, une tâche-problème imposant aux élèves la production d'un texte narratif impliquant la résolution d'anaphores de divers types (Garcia-Debanc & Bonnemaison 2014 ; Garcia-Debanc & Bras 2016 ; Garcia-Debanc *et al.* ce volume) :

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

L'enjeu de l'entreprise est de mettre au jour les données et les mécanismes à l'œuvre dans l'interprétation de discours dont la cohérence, en tant que propriété de la réception des discours (Charolles 1995), varie d'un scripteur à l'autre, et évolue tout au long de la scolarité. Notre corpus est constitué de textes d'élèves de CE2, 6ème et 3ème — trois niveaux de classe correspondant aux fins des cycles 2, 3 et 4 —

afin d'observer d'éventuels paliers d'évolution. Nous donnons ci-dessous un exemple de texte produit par un élève de CE2.

Fig. 1. Texte de Manuel, élève de CE2



La SDRT est mise à l'épreuve pour la première fois sur des textes d'apprenants dans le projet E-CALM. La construction du corpus annoté a pour objectif de tester la capacité explicative de la théorie face à divers types d'incohérence et à une structure globale des textes plus ou moins complexe. Elle cherche aussi à évaluer des hypothèses sur l'évolution de la cohérence entre la fin du cycle 2 et la fin du cycle 4.

Cet article est consacré à la méthodologie d'annotation du corpus et à la campagne d'annotation. Nous présenterons d'abord brièvement le cadre théorique et la méthodologie choisies pour annoter la cohérence (section 2). Puis nous présentons la campagne d'annotation (section 3), avant de détailler les phases de segmentation (section 4) et d'annotation en relations de discours (section 5).

2. ANNOTER LA COHERENCE : CADRE THEORIQUE ET METHODOLOGIE

Dans notre cadre théorique, la SDRT (Asher & Lascarides 2003), un discours cohérent est un discours pour lequel on peut construire une représentation formelle. La représentation est construite grâce à des règles qui s'appuient à la fois sur des connaissances linguistiques, des connaissances extralinguistiques et des principes pragmatiques impliqués dans le processus d'interprétation. La SDRT est donc une théorie de l'interface sémantique/pragmatique qui cherche à prédire et à évaluer la

cohérence d'un discours par le calcul d'une représentation rendant compte de l'interprétation. La SDRT présente une dimension opératoire, au sens où elle donne une sorte d'algorithme de construction des représentations. Elle revêt également une dimension systématique et extensive, au sens où elle décrit le rôle de toutes les informations, quelle que soit leur nature, impliquées dans l'interprétation. C'est précisément ces trois dimensions – opératoire, systématique et extensive – qui en font une théorie pertinente pour analyser les textes d'élèves dans la perspective d'une meilleure appréhension des sources d'incohérence.

La cohérence est appréhendée en SDRT à travers les Relations de Discours (RD), appelées aussi relations de cohérence ou rhétoriques dans d'autres approches (Hobbs 1985, Mann & Thompson 1987, Sanders *et al.* 1992, Hovy & Maier 1992).

Selon Charolles (1995), la cohésion[2] s'articule avec la cohérence sans être ni une condition nécessaire ni une condition suffisante de la cohérence ; la cohérence est une propriété évaluée par le récepteur tout au long du processus d'interprétation du discours. La SDRT peut rendre compte de la cohésion et de la cohérence. Dans le processus d'annotation décrit, nous privilégions l'annotation de la cohérence, en annotant de façon secondaire les sources d'incohérence relevant de la cohésion.

En SDRT, un discours est représenté par une structure récursive, une Segmented Discourse Representation Structure (SDRS), constituée de Discourse Representation Structures (les DRS, ou représentations de la DRT) ou de SDRS, reliées entre elles par des RD comme Narration, Résultat, Contraste, Arrière-Plan, Élaboration, Explication … Une RD peut être subordonnante, au sens où elle installe une relation de domination entre les représentations de deux segments du discours, ou coordonnante quand la relation entre les représentations des segments reste sur le même plan (Asher & Vieu 2005)[3].

La cohérence occupe une place centrale dans le processus de construction des SDRS : un discours est évalué comme cohérent si la représentation de chaque segment de ce discours peut être rattachée à la représentation en cours de construction avec une RD et si toutes les relations de cohésion peuvent être résolues avec cet attachement et cette RD. Dans le cas où il est impossible d'attacher un segment à la structure existante, le discours est évalué comme étant incohérent et la construction de la représentation (donc l'interprétation) s'arrête.

La SDRT a déjà été appliquée à des données attestées, textes littéraires ou journalistiques, produits par des scripteurs experts, notamment pour la construction du corpus ANNODIS[4], premier corpus annoté en RD pour le français (Afantenos *et al.* 2012, Asher *et al.* 2017). Le corpus ANNODIS est constitué d'extraits de textes de presse, de wikipedia et d'articles scientifiques. Les RD sont celles de la SDRT, sachant que ces relations sont elles-mêmes partagées en grande partie avec d'autres théories du discours (Hovy & Maier 1992).

---

[2] Halliday et Hasan (1976) définissent la cohésion d'un texte comme un ensemble de liens de cohésion entre différents éléments du texte. Un lien de cohésion s'établit quand l'interprétation d'un élément dépend de celle d'un autre élément.

[3] Il n'est pas possible de présenter plus en profondeur le cadre théorique dans les limites de cet article. Nous renvoyons le lecteur à tous les articles cités.

[4] http://redac.univ-tlse2.fr/corpus/annodis/

Notre objectif ici est d'évaluer la possibilité pour la SDRT de rendre compte de textes d'apprenants, dont la compétence rédactionnelle est en cours d'acquisition. En l'état actuel de la théorie, le processus de construction des SDRS s'arrête à la première impossibilité d'attachement d'un segment à la représentation en cours de construction. Ce blocage du processus équivaut à évaluer le discours comme étant incohérent. Dans le processus d'annotation décrit dans cet article, nous continuons la construction au-delà des blocages pour tenter de mesurer des degrés d'incohérence, ce qui exige un enrichissement de la théorie, pour pouvoir notamment typer et quantifier les incohérences. Nous renvoyons le lecteur à (Bras & Vieu sous presse) pour une présentation plus détaillée du cadre théorique dans la perspective de l'annotation des textes d'élèves.

Notre travail sur ces textes s'organise en trois étapes. Il s'agit d'abord de segmenter les textes en Unités de Discours Élémentaires (UDE) sur la base de critères syntaxiques et sémantico-référentiels. L'étape suivante consiste à annoter les RD entre UDE, et les différents problèmes rencontrés au moment des décisions d'attachement de chaque UDE et des choix de RD. L'annotation des RD entre UDE permet de former des segments complexes et d'aboutir à une « structure de discours », schématisée sous forme de graphe, sans transcrire la représentation du contenu propositionnel des UDE, comme on le ferait en SDRT standard. L'annotation des problèmes rencontrés nous permettra de rendre compte des différents types d'incohérence afin de construire une typologie des problèmes entravant la construction de la représentation, c'est-à-dire l'interprétation du texte. Enfin, la dernière étape consiste en l'analyse des annotations obtenues pour y observer et éventuellement y quantifier les points d'incohérence relevés à l'étape précédente, afin de permettre la comparaison des structures de discours des textes de CE2, 6ème, 3ème.

3. CAMPAGNE D'ANNOTATION

La campagne d'annotation s'est déroulée sur une période de 2 ans (2020-2021). Elle a impliqué au total six annotatrices dont deux chercheuses et une ingénieure de recherche expertes, formant peu à peu les trois autres annotatrices, étudiantes en master ou en début de thèse, et ayant toutes suivi un enseignement sur la SDRT. Ce choix s'est fondé sur l'expérience du projet ANNODIS pour lequel les annotateurs étaient « naïfs », afin de tester la compréhension intuitive du sens des RD et de la structure hiérarchique. Dans le projet E-CALM, au contraire, nous voulons tester les limites du pouvoir d'expression des RD et de la SDRT, il est donc indispensable que les annotateurs soient initiés au cadre théorique.

La campagne s'est appuyée sur un guide de segmentation en UDE et sur un guide d'annotation en RD. Nous avons adapté les guides du projet ANNODIS (Muller *et al*. 2012) au corpus Résolco (Lala *et al.* 2017a-b), ces guides ont été complétés, notamment par l'annotation des problèmes d'interprétation donnant lieu à des incohérences (voir liste des « problèmes de cohérence » en section 5), et améliorés tout au long de la campagne (Bras *et al.* 2021a-b) pour prendre en compte les spécificités des textes d'élèves. Ces spécificités sont liées au fait que les auteurs des textes sont des apprentis-scripteurs. D'une part, ils ne maitrisent pas complètement les moyens de segmentation et d'organisation que sont la ponctuation

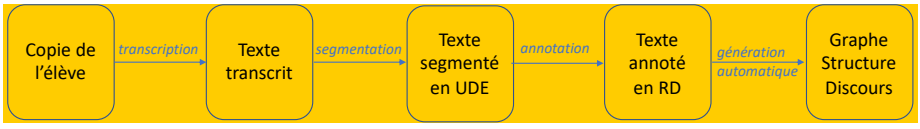
et syntaxe ; la segmentation nécessite donc un recours à la sémantique plus important que pour des textes d'experts-scripteurs. D'autre part, ils ne contrôlent pas bien les interactions entre les ingrédients variés de l'interface sémantique-pragmatique (marqueurs de discours, temps verbaux, anaphores, présuppositions…) permettant à un récepteur de reconstituer la structure discursive voulue par le scripteur ; l'annotation en RD doit prendre en compte des combinaisons de ces ingrédients inédites chez les experts-scripteurs.

Nous avons segmenté au total 115 textes, en commençant par 15 textes en quintuple annotation, lors d'une phase exploratoire progressive destinée à la production d'une segmentation de référence pour chaque texte et à la mise à jour progressive du guide de segmentation. La phase nominale de segmentation a été réalisée en double annotation avec adjudication collective pour les 100 autres textes. Cette phase a exigé plus de 60h de réunions collectives, auxquelles s'ajoutent les temps de segmentation de chaque annotatrice, soit près de 550 heures au total.

L'annotation en RD a également commencé par une phase exploratoire sur 12 textes. Une deuxième phase, pratiquement terminée, vise à annoter 12 textes supplémentaires. Étant donnée la complexité de la tâche, surtout quand les textes sont incohérents, ces deux phases se déroulent en quadruple annotation avec réunions d'harmonisation hebdomadaires, pour un total estimé de 330 heures pour les 24 premiers textes. Une dernière phase permettra d'obtenir un corpus final de 36 textes minimum (au moins 12 par niveau).

La Fig. 2 présente la chaine de traitement mise en œuvre pour la construction de ce corpus annoté (voir (Garcia-Debanc *et al.* 2017) pour la description de l'étape de transcription ; voir section 5 pour la génération automatique de la Structure de Discours).

Fig. 2. Chaine de traitement du corpus Résolco pour l'annotation de la Structure de Discours



4. SEGMENTATION

Les Unités de Discours Élémentaires (UDE) sont les segments minimaux des représentations du discours en SDRT. Ils correspondent dans les cas prototypiques à une « proposition syntaxique » au sens d'une structure prédicat/arguments accompagnée des compléments rattachés au verbe, souvent délimitée par des marques de ponctuation, introduisant un référent d'évènement ou d'état de fait. Les critères de délimitation de ces segments sont donc essentiellement ponctuationnels, syntaxiques et sémantico-référentiels.

L'expérience de segmentation sur des données attestées, dans le projet ANNODIS notamment, a permis d'affiner et d'élargir ces critères pour tenir compte

du rôle de certains éléments ayant une autonomie discursive tout en étant rattachés syntaxiquement à une proposition-hôte : les adverbiaux introduisant des cadres de discours (Charolles 1997), analysés en SDRT comme introduisant de nouveaux topiques évènementiels (Vieu *et al.* 2005), ou les appositions réalisant des élaborations d'entités (Prévot *et al.* 2009). Dans le manuel de segmentation d'ANNODIS, une UDE décrit un évènement ou un état de fait, qu'elle apparaisse sous la forme d'une proposition indépendante, principale, d'une proposition subordonnée conjonctive finie, gérondive, participiale, infinitive, d'une subordonnée relative non déterminative, d'une expression elliptique, d'un cadratif ou d'une apposition (Muller *et al.* 2012).

Les cadratifs temporels et les appositions sont fréquents dans notre corpus. Mais leur autonomie discursive n'est pas toujours matérialisée par la ponctuation. La difficulté de la tâche de segmentation des textes d'élèves réside souvent dans le fait qu'il est impossible de se fier à la ponctuation et à l'emploi des majuscules : il faut souvent ignorer ces marques ou bien au contraire les reconstruire pour segmenter, tout en identifiant les cas où la ponctuation est déterminante et ne peut être ignorée. Le manuel de segmentation (Bras *et al.* 2021a) préconise donc une prise en compte minimale de la ponctuation et de l'emploi des majuscules, tout en gardant une attention à celles-ci. Pour ces mêmes raisons, nous ne tenons pas compte des phrases graphiques.

Pour illustrer plus avant les consignes de segmentation ainsi définies, nous donnons ci-dessous la segmentation du texte de Manuel-CE2 vu en Fig. 1:

> Texte de Manuel (CE2) segmenté
> [Il était une fois, une fille]$_1$*P [qui s'appelait Pascal.]$_2$ [Elle habitait une grande maison.]$_3$ [Elle habitait dans cette maison depuis longtemps.]$_4$ [Et à côté de la maison]$_5$ [il y avait un garçon]$_6$ [qui s'appelait Jean.]$_7$ [Tout les deux la nuit ils allaient dans une forêt.]$_8$ [Une nuit]$_9$ [ils avaient entendu des personnes marcher]$_{10}$ [alors Jean se demanda]$_{11}$ [ce que c'était]$_{12}$ [alors Il se retourna]$_{13}$*C [en entendant ce grand bruit.]$_{14}$ [Et c'est la qu'il a vu une personne !]$_{15}$ [Les enfants se sont vite cacher dans un buisson.]$_{16}$ [Depuis cette aventure,]$_{17}$ [les enfants ne sortent plus la nuit.]$_{18}$

Cet exemple permet d'illustrer la segmentation des cadratifs (segments 5, 9 et 17), celle de différents types de propositions subordonnées (segments 2, 7, 14), et celle des segments décrivant l'argument d'un verbe de parole ou d'attitude propositionnelle (segment 12).

Nous notons *P (Ponctuation) seulement quand la ponctuation interfère avec la segmentation, c'est-à-dire dans les cas où nous sommes contraintes de ne pas respecter la ponctuation qui indiquerait une fin de segment pour suivre la dimension syntaxique ou sémantique, qui prévalent dans ce cas sur la segmentation, voir segment 1 du texte de Manuel ci-dessus et l'exemple (1) :

1. [Il était une fois .un petit garçon [nommé Roméo.]$_2$ Et sa voisine Julliette]$_1$*P [qui abitait un manoir.]$_3$ (Texte de Gabriel, CE2)

Le texte de Manuel ne respecte pas strictement la consigne dans la mesure où la deuxième phrase imposée est précédée de « alors », ce que nous notons par le code *C (Consigne).

Nous utilisons aussi le code *S (Syntaxe) dans les cas où il manque un mot pour constituer un segment complet, comme illustré en (2) :

2. [Et un moment]$_9$*S [Roméo entendit un grand bruit]$_{10}$ (Texte de Gabriel, CE2)

5. ANNOTATION EN RELATIONS DE DISCOURS ET PROBLEMES DE COHERENCE

L'annotation en RD en SDRT est récursive, segment après segment, et consiste principalement à déterminer le point d'attachement s_n du segment courant s_m ainsi que la RD R réalisant cet attachement. L'annotation proprement dite est alors une séquence de $R(s_n,s_m)$ où R est une RD et s_n et s_m sont les numéros des segments reliés. Lorsqu'on est en présence de segments complexes, s_m est amené à être remplacé par un ensemble de segments noté $[s_i, s_j, \ldots]$ ou $[s_i\text{-}s_j]$.

Nous avons dû étendre ce schéma afin que l'annotation rende compte de divers problèmes de cohérence.[5] Tout d'abord, pour les cas où nous ne sommes pas en mesure de déterminer le point d'attachement du segment courant s_m, ou la RD réalisant cet attachement, ou les deux, nous introduisons des annotations du type : $R(\ ?,s_m)$, $?(s_n,s_m)$ ou $?(\ ?,s_m)$. Ensuite, même si l'interprétation du texte permet d'identifier un attachement et une RD, et donc de produire $R(s_n,s_m)$, de nombreux problèmes peuvent se présenter, qui seront alors annotés en ajoutant sur la même ligne #problème(s_i) (éventuel argument), où s_i est le segment, en général s_m, contenant l'élément générant le problème de cohérence.

Le jeu de 24 RD que nous utilisons est celui du projet ANNODIS (Muller *et al.* 2012), augmenté par la relation de Résultat Faible (Bras *et al.* 2009) et par les relations de dialogue de la SDRT (Asher & Lascarides 2003), dont la nécessité s'est manifestée lors de l'annotation d'éléments de dialogue dans certains textes du corpus[6] :

Acquiescement (ACQ), Alternance (ALT), Arrière-plan (ARP), Attribution (ATT), But (BUT), Cadre (CAD), Commentaire (COM), Conditionnel (CND), Continuation (CTN), Contraste (CTR), Correction (COR), Elaboration (ELB), Elaboration d'entité (EEL), Elaboration de question (QEL), Explication (EXP), Fusion (FUS), Localisation temporelle (TMP), Narration (NAR), Parallèle (PAR), Question de clarification (QCL), Question-Réponse (QRP), Résultat (RES), Résultat faible (RSF), Retour-arrière (RAR).

[5] Comme indiqué en section 2, la possibilité même de construire une représentation du discours correspond en SDRT à la possibilité d'interpréter un discours et donc à la reconnaissance de sa cohérence sémantico-pragmatique. Cette vision est nécessairement abandonnée ici.

[6] Nous renvoyons le lecteur intéressé par la description de ces RD, qui ne sont pas objet d'étude ici, aux travaux cités ci-dessus ainsi qu'à des travaux dans d'autres théories basées sur des relations de discours ou de cohérence comme ceux de Sanders *et al*. (1992), Hovy & Maier (1992) ou Kehler (2002).

Le jeu d'étiquettes pour typer et annoter les problèmes de cohérence est construit de façon incrémentale lors de la phase pilote d'annotation ; il semble désormais stable, mais il est possible qu'il soit encore étendu. Il contient actuellement 8 types :

- #tense(s_i) (temps inapproprié / temps requis) : le temps verbal de l'éventualité principale du segment s_i est incompatible avec la RD annotée.
- #tense/alt+(s_i) : le segment s_i induit une alternance de temps non canonique mais acceptable, e.g., une alternance passé simple / présent ou passé simple / passé composé est souvent acceptable dans les textes narratifs.
- #tense/alt-(s_i) : le segment s_i induit une alternance inacceptable, i.e., une alternance similaire aux alternances acceptables, mais employée à mauvais escient en contexte.
- #anaphore(s_i) (élément anaphorique) : soit il est impossible de trouver un antécédent à l'élément anaphorique de s_i indiqué en argument, soit l'anaphore est ambigüe, i.e., il y a plusieurs antécédents possibles mais aucun moyen de déterminer lequel serait celui voulu par l'auteur (dans ce second cas, on note comme argument élément/amb)
- #presupposition(s_i) (élément présuppositionnel) : on ne peut accommoder en contexte l'élément présuppositionnel de s_i indiqué en argument
- #semantique(s_i) : les effets sémantiques de la RD sont incohérents avec certains éléments du contexte. Les cas de problèmes sémantiques sont variés, on complète donc l'annotation avec une explication.
- #structure d'information(s_i) : la structure d'information du segment s_i est inappropriée en contexte. On complète ici aussi l'annotation avec une explication.
- #progthem(s_i) : le segment s_i induit une rupture dans la progression thématique du texte.

Le processus d'annotation est décrit dans le manuel d'annotation (Bras *et al.* 2021b) et illustré en détails dans (Bras & Vieu sous presse). Le texte de Manuel CE2 segmenté à la section précédente est annoté ainsi :

EEL(1,2) % une fille[7]
ARP(1,[3,4])
ELB(1,5)
ARP(5,6)
EEL(6,7) % un garçon
ARP(1,8)
ELB(1,[9-18])
CAD(9,[10-16])
RES(10,11) #tense(10) (PQP/PS)
NAR(10,11)
ATT(11,12)
RES(11,13)
NAR(11,13)
EXP(13,14) #anaphore(14) (ce grand bruit)
RSF(13,15) #tense/alt-(15)
NAR(13,15)
RES(15,16)
NAR(15,16)

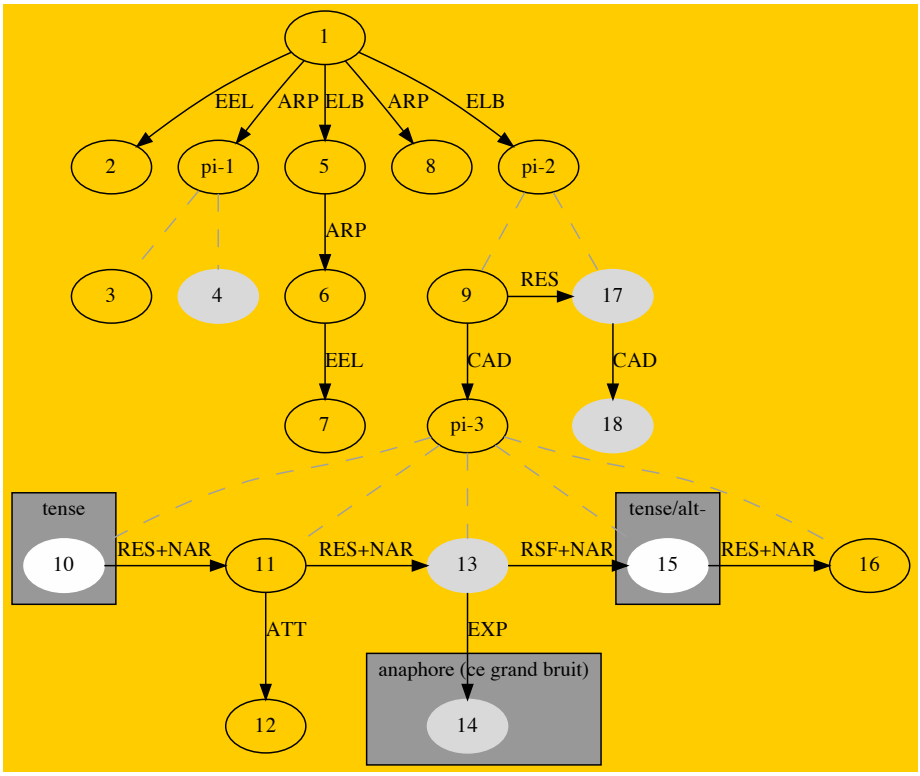
[7] Certaines RD requièrent l'annotation d'éléments supplémentaires (e.g., %<GN de sn élaboré> pour la relation d'Elaboration d'entité).

RES(9,17)
CAD(17,18)

Nous avons développé un script Python permettant de générer automatiquement le graphe correspondant à l'annotation, graphe dont les nœuds sont les représentations des segments (UDE ou segments complexes), les arcs étiquetés les RD (horizontaux pour les RD coordonnantes), et les arcs en pointillés les liens entre un segment complexe et ses sous-segments. Ceci est non seulement utile pour appréhender la structure globale du texte et notamment visualiser son niveau de complexité structurelle, mais aussi plus prosaïquement pour repérer facilement d'éventuelles coquilles au cours du processus d'annotation même.

La Fig. 3 présente le graphe généré pour le texte de Manuel annoté ci-dessus. Les segments complexes sont notés pi-n, les segments des phrases imposées sont colorés en gris clair et les problèmes de cohérence sont visualisés par des rectangles gris foncé encadrant les segments sources de l'incohérence.

Fig. 3. Graphe d'annotation en relation de discours du texte de Manuel (CE2)



Ce graphe illustre une structure globale fréquemment rencontrée dans le corpus : après une première partie (segments 1-8) décrivant la situation initiale à l'aide principalement de relations d'Arrière-Plan et d'Elaboration d'entité, on voit la

complication qui est introduite par un cadre (9) et qui s'organise autour de relations de Narration et des relations causales Résultat, Résultat faible, et Explication. La situation finale, ici portée par la dernière phrase imposée, s'attache par la relation Résultat au cadre-topique (9) qui domine toute la complication et qui comporte par construction un événement la synthétisant, disponible pour résoudre l'anaphore résomptive du GN démonstratif *cette aventure* du segment 17.

Ce texte comporte cependant plusieurs problèmes de cohérence. Le premier est dû au temps verbal de 10, un plus-que-parfait, incompatible avec les relations de Résultat et Narration signalées ici par le connecteur *alors*. Le deuxième est un problème récurrent dans notre corpus (Garcia-Debanc 2020) : le GN démonstratif anaphorique *ce grand bruit* de la deuxième phrase imposée est sans antécédent, malgré les efforts de l'auteur pour introduire un son auparavant (en 10). Enfin, l'alternance des temps verbaux entre 13 et 15 est maladroite.

Pour des raisons d'espace, nous ne pouvons pas décrire plus avant le processus d'annotation, mais le lecteur peut se reporter à (Bras & Vieu sous presse) pour une illustration détaillée de l'annotation de trois textes du corpus.

Après annotation de 24 textes (8 par niveau), nous pouvons déjà observer certaines tendances d'évolution entre niveaux, sur la base de premières statistiques rassemblées dans le Tableau 1.

Tableau 1 : premiers résultats

	Longueur en nombre de segments : moyenne (écart-type)	Nombre de RD distinctes[8] : moyenne (écart-type)	Taux d'incohérence relatif à la longueur : moyenne (écart-type)
CE2	16,8 (6,6)	7,4 (2,3)	0,49 (0,39)
6ème	32,5 (12,7)	10,9 (1,8)	0,16 (0,15)
3ème	44,5 (20,5)	12,3 (0,4)	0,16 (0,08)

La longueur des textes augmente fortement en moyenne ; ce paramètre reste toutefois très variable (l'écart-type est élevé). Le nombre de RD distinctes utilisées par texte augmente régulièrement, sans que cela soit corrélé à la longueur des textes, et l'usage est de plus en plus homogène (l'écart-type baisse fortement). Le taux de problèmes de cohérence par segment baisse considérablement entre le CE2 et la 6ème, après quoi le taux reste stable tout en s'homogénéisant (l'écart-type baisse de moitié entre 6ème et 3ème). Ces observations montrent une maitrise croissante de la cohérence discursive, avec l'exploitation d'une richesse grandissante de moyens. Au-delà des résultats de ce tableau, on observe des évolutions dans l'usage de certaines RD, par exemple, la présence de Résultat et d'Elaboration diminue régulièrement entre CE2 et 3ème, au profit notamment d'Explication et d'Elaboration d'Entité qui deviennent plus fréquentes en 3ème. Continuation est en forte augmentation entre le CE2 et la 6ème et 3ème, ce qui traduit une présence croissante de segments complexes.

[8] Nous avons écarté dans ces calculs les RD internes à la deuxième et troisième phrases imposées, systématiquement les mêmes (EXP(13,14) et CAD(17,18) dans l'exemple traité).

6. Conclusions d'etape et Perspectives

Nous avons présenté dans cet article notre méthodologie de segmentation et d'annotation en RD. L'annotation en cours, dans le cadre du projet E-CALM qui se termine en 2021, vise un corpus d'au moins 36 textes de CE2, 6ème et 3ème extrait du corpus Résolco. Le corpus sera ensuite étendu dans le cadre d'une thèse en cours (Pépin-Boutin, en prép.).

Nous avons privilégié la qualité des annotations et la rigueur du manuel pour la reproductibilité de l'expérience. Celle-ci a pu être mesurée pour la segmentation, pour laquelle les accords inter-annotatrices se situent autour de 90% dans la phase nominale. Cela ne veut pas dire pour autant que la segmentation est automatisable, du fait de la ponctuation largement défaillante et du recours fréquent à des critères sémantiques. L'automatisation serait envisageable avec un corpus de taille suffisante pour pouvoir entrainer des algorithmes d'apprentissage, ce qui est hors de portée dans le cadre de ce projet, étant donné le temps requis pour la segmentation.

Pour l'annotation en RD, nous n'avons pas pu calculer d'accord inter-annotatrices dans cette phase exploratoire. L'établissement de l'annotation de référence n'est pas un simple choix de la meilleure solution parmi les annotations des annotatrices car beaucoup de décisions ne sont pas seulement locales. L'annotation finale est souvent obtenue après une phase de discussion collective approfondie conduisant à l'enrichissement continu du manuel d'annotation et à l'évolution de la typologie des problèmes de cohérence. On sait également que de tels calculs d'accord sont délicats (Asher *et al.* 2017) parce qu'il existe des équivalences entre structures de discours même si dans le cadre de cette campagne d'annotation nous avons normé l'annotation des segments complexes en privilégiant la portée la plus large[9], notamment pour garantir le fonctionnement optimal du script générant les graphes.

Nous n'avons pas dans ce projet d'objectif d'annotation automatique des textes du corpus, contrairement au projet ANNODIS. L'objectif est avant tout de faire progresser la connaissance mutuelle et la coopération entre linguistique théorique et didactique de l'écrit, à travers l'établissement d'une cartographie des indicateurs de maitrise de la cohésion et de la cohérence chez les élèves (Garcia-Debanc & Bras 2016).

Le choix de privilégier la qualité des annotations nous a conduites à des analyses approfondies de phénomènes non encore abordés dans le cadre d'une théorie du discours telle que la SDRT et nous a permis à la fois d'enrichir le cadre théorique et de dégager de futures pistes de recherche. Nous avons montré que la SDRT ainsi étendue permettait de rendre compte de façon satisfaisante de textes d'apprenants présentant des degrés d'incohérence variés.

Il sera très instructif de continuer à observer et analyser les régularités présentes, notamment sur les RD utilisées ou la complexité structurelle des graphes selon les niveaux de compétence scripturale des auteurs, dans ce corpus comme dans d'autres corpus. En effet, la méthodologie d'annotation des textes d'élèves du corpus Résolco définie ici est indépendante de la consigne qui a permis sa production. Elle pourra donc être utilisée pour annoter d'autres corpus de textes d'élèves, directement pour des textes relevant du genre narratif, et avec quelques aménagements éventuels pour d'autres genres.

[9] Dans l'annotation de l'exemple, cela signifie que nous annotons ELB(1,[9-18]) au lieu de ELB(1,[9,17]).

Références bibliographiques

AFANTENOS S. ; ASHER N. ; BENAMARA F. ; BRAS M. ; FABRE C. ; HO-DAC M. ; LE DRAOULEC A. ; MULLER P. ; PÉRY-WOODLEY M.-P. ; PRÉVOT L. ; REBEYROLLE J. ; TANGUY L. ; VERGEZ-COURET M. & VIEU L. (2012), "An empirical resource for discovering cognitive principles of discourse organisation: the ANNODIS corpus", *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, ELRA (en ligne)

ASHER N., LASCARIDES A. (2003), *Logics of Conversation*. Cambridge: Cambridge University Press.

ASHER N., MULLER P., BRAS M., HO-DAC L.-M., BENAMARA F., AFANTENOS S. & VIEU L. (2017), "ANNODIS and related projects: case studies on the annotation of discourse structure" In N. Ide and J. Pustejovsky (Eds.): *Handbook of Linguistic Annotation*, pp. 1241–1264. Springer.

ASHER N. & VIEU L. (2005), "Subordinating and coordinating discourse relations », *Lingua* 115(4), 91-610.

BRAS M. & VIEU L. (sous presse), « Segmenter et annoter les relations de cohérence dans des textes narratifs d'élèves de 9 à 15 ans : quels apports d'une théorie de l'interface sémantique/pragmatique pour les enseignants ? », in Longhi B. & Lewi O. (éds) *Segmenter, connecter, ponctuer à l'écrit : un défi pour l'enseignement*. Peter Lang

BRAS M., LE DRAOULEC A. & ASHER N. (2009), "A formal analysis of the French Temporal Connective 'alors'", in Behrens, B. & Fabricius-Hansen, C. (éds.) *Structuring information in discourse: the explicit/implicit dimension*, *OSLa Oslo Studies in Language* 1, 149-170 (en ligne).

BRAS M., VIEU L. POUJADE & C. ROZE C. (2021a), *Manuel de segmentation en unités de discours élémentaires du corpus Résolco.* Rapport interne CLLE-ERSS, Toulouse.

BRAS M., VIEU C. ROZE JORET M. & PEPIN-BOUTIN A. (2021b), *Manuel d'annotation en relations de discours du corpus Résolco*. Rapport interne, CLLE-ERSS, Toulouse.

CHAROLLES M. (1978), « Introduction au problème de la cohérence des textes », *Langue Française* 78, 7-41.

CHAROLLES M. (1995), « Cohésion, cohérence et pertinence du discours », *Travaux de linguistique* 29, 125-151.

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

GARCIA-DEBANC C. (2020), « Ecrire et réécrire pour résoudre des problèmes de cohésion textuelle : quel est donc *ce grand bruit* dans le corpus RESOLCO ? Analyse de récits d'élèves de 9 à 15 ans », SHS Web Conf. 78 (2020) 07021 (en ligne).

GARCIA-DEBANC C. & BONNEMAISON K. (2014), « La gestion de la cohésion textuelle par des élèves de 11-12 ans : réussites et difficultés », in *Actes du 4e Congrès Mondial de Linguistique Française (CMLF 2014)*, Juillet 2014, Berlin, Allemagne (en ligne).

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 9-12 ans : indicateurs de maitrise et progressivité », in S. Plane, C. Bazerman, F. Rondelli, C. Donahue, A. N. Applebee, C. Boré, P. Carlino, M. Marquilló Larruy, P. Rogers & D. Russell (éds) *Recherches en écritures : regards pluriels*, *Recherches textuelles* 13, Metz : Université de Loraine, 39-62.

GARCIA-DEBANC C., HO-DAC M., BRAS M. & REBEYROLLE J. (2017). « Vers l'annotation discursive de textes d'élèves », *Corpus* 16 | 2017 (en ligne).

GARCIA-DEBANC C., HO-DAC M. & REBEYROLLE J. (ce volume), « La continuité référentielle dans le corpus Résolco ».

HALLIDAY M. & HASAN R. (1976), *Cohesion in English*. London**,** Longman.

HOBBS, J. R. (1985). *On the coherence and structure of discourse.* Rapport technique CSLI-85-37, Center for Study of Language and Information.

HOVY H. & MAIER E. (1992), *Parsimonious or Profligate: How Many and Which Discourse Structure Relations?* Rapport technique ISI/RR-93-373, USC Information Sciences Institute.

KAMP H. & REYLE U. (1993), *From Discourse to Logic*. Kluwer.

KEHLER A. (2002), *Coherence, Reference, and the Theory of Grammar*. CSLI Publications.

LALA M., BRAS M. & GARCIA-DEBANC C. (2017a), *Manuel de segmentation en unités de discours élémentaires du projet Résolco*. Rapport interne**,** CLLE-ERSS, Toulouse.

LALA M., BRAS M. & GARCIA-DEBANC C. (2017b), *Manuel d'annotation en relations de discours du projet Résolco*. Rapport interne. CLLE-ERSS, Toulouse.

MANN W. & THOMPSON S. (1987), *Rhetorical Structure Theory: A Theory of Text Organization*. Rapport technique Reprint Series ISI/RS-87-1190, Information Sciences Institute.

MULLER P., VERGEZ-COURET M., PREVOT 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, Toulouse.

PEPIN-BOUTIN A. (en préparation), *La cohérence dans les récits de sujets présentant une pathologie neurodéveloppementale : analyse sémantique et pragmatique du discours.* Thèse de l'Université Toulouse Jean Jaurès.

PRASAD R., DINESH N., LEE A., MILTSAKAKI E., ROBALDO L., JOSHI A. & WEBBER B. (2008), "The Penn Discourse Treebank 2.0.", in *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC-2008)***,** Marrakech, Morocco (en ligne).

PREVOT L., VIEU L. & ASHER N. (2009), « Une formalisation plus précise pour une annotation moins confuse : la relation d'élaboration d'entité », *Journal of French Language Studies* 19(2), 207-228.

RONDELLI F. (2010), « La cohérence textuelle : pratiques des enseignants et théories de référence », *Pratiques* 145-146 | 2010, 55-84.

SANDERS T., SPOOREN W. & NOORDMAN L. (1992), "Towards a taxonomy of coherence relations", *Discourse Processes* 15, 1-35.

VIEU L., BRAS M., ASHER N. & AURNAGUE M. (2005), "Locating Adverbials in Discourse", *Journal of French Language Studies* 15, 173-193.