Annotation de la cohérence dans un corpus de textes d'élèves d'école et collège

Annotation of coherence in a corpus of texts from primary and middle school

Myriam BRAS – CLLE, Université de Toulouse, CNRS, Toulouse, France
Laure VIEU – IRIT, CNRS, Université de Toulouse, Toulouse, France

Résumé : Cet article traite de la question de la cohérence dans des productions écrites d'élèves d'école et collège. Le corpus analysé est issu du corpus RESOLCO, il est constitué de 36 textes d'élèves d'école primaire et de collège (CE2, 6ème et 3ème) produits selon une même consigne d'écriture, une tâche-problème demandant aux élèves la production d'un texte narratif impliquant la résolution d'anaphores de divers types. Il a été annoté dans le cadre du projet E-Calm avec l'objectif d'analyser la cohérence à partir de l'identification de relations entre segments de texte. Le schéma d'annotation est basé sur la Segmented Discourse Representation Theory qui définit de façon formelle ce qu'est une représentation de discours cohérente. L'article présente la méthodologie de segmentation des textes en Unités de Discours Élémentaires et d'annotation en Relations de Discours et en problèmes de cohérence (incohérences locales, impossibilités d'inférer une Relation de Discours, ou impossibilités d'attacher la représentation d'un segment au reste de la représentation), réalisant une extension originale de la SDRT. L'analyse des annotations met en évidence une tendance à la baisse de ces trois types d'indicateurs, traduisant une progression de la cohérence du CE2 à la 3ème, avec un palier significatif entre le CE2 et la 6ème, à corréler avec un saut majeur de compétences rédactionnelles entre la fin du cycle 2 et la fin du cycle 3. La théorie, mise à l'épreuve ici pour la première fois sur des textes d'apprenants, se trouve ainsi validée empiriquement dans sa version étendue.

Mots-Clés : cohérence du discours, textes d'élèves, corpus annoté en relations de discours

Abstract: This article deals with the question of coherence in the written productions of primary and secondary school pupils. The corpus analyzed comes from the RESOLCO corpus, and comprises 36 texts by primary and secondary school pupils (CE2, 6ème and 3ème) produced according to the same writing instruction, a problem-task requiring pupils to produce a

narrative text involving the resolution of anaphora of various types. It was annotated as part of the E-Calm project, with the aim of analyzing coherence by identifying relationships between text segments. The annotation scheme is based on Segmented Discourse Representation Theory, which formally defines what a coherent discourse representation is. The paper presents the methodology for segmenting texts into Elementary Discourse Units and annotating them into Discourse Relations and coherence problems (local inconsistencies, impossibilities of inferring a Discourse Relation, or impossibilities of attaching the representation of a segment to the rest of the representation), achieving an original extension of SDRT. Analysis of the annotations reveals a downward trend in these three types of indicator, reflecting a progression in coherence from CE2 to 3ème, with a significant plateau between CE2 and 6ème, to be correlated with a major leap in writing skills between the end of cycle 2 and the end of cycle 3. The theory, tested here for the first time on learners' texts, is thus empirically validated in its extended version.

Keywords: discourse coherence; student texts; annotated corpus in discourse relations

1. Analyser la cohérence des textes d'élèves : pourquoi et comment ?

Cette contribution est consacrée à l'analyse de la cohérence dans un corpus de productions écrites d'élèves d'école et de collège. La cohérence est envisagée comme une propriété de la représentation que se construit celui qui interprète le texte tout au long de sa lecture, se situant ainsi du côté du processus cognitif de réception des textes (Charolles, 1995). Le corpus analysé est issu du corpus RESOLCO[1] constitué de textes d'élèves d'école primaire et de collège produits selon une même consigne d'écriture, une tâche-problème demandant aux élèves la production d'un texte narratif impliquant la résolution d'anaphores de divers types (Garcia-Debanc et Bonnemaison, 2014 ; Garcia-Debanc et al, 2017). Nous y avons sélectionné trois niveaux correspondant aux fins des cycles 2, 3 et 4 – CE2, 6ème et 3ème – afin d'observer d'éventuels paliers d'évolution. Ce corpus a été annoté dans le cadre du projet E-Calm avec l'objectif d'analyser la cohérence à partir de l'identification de relations de cohérence, ou relations de discours, entre segments de texte. Le schéma d'annotation est basé sur une théorie de

[1] http://redac.univ-tlse2.fr/corpus/resolco

l'interface sémantique/pragmatique, la Segmented Discourse Representation Theory (Asher et Lascarides, 2003) qui définit de façon formelle ce qu'est une représentation de discours cohérente. La SDRT a déjà été mise à l'épreuve des données sur des textes d'experts (voir par exemple le projet ANNODIS, Asher et al. 2017). Elle a été évaluée pour la première fois sur des textes de scripteurs dont la compétence rédactionnelle est encore en cours d'acquisition, dans le cadre du projet E-Calm (Bras et Vieu, 2022). Il s'agissait pour nous d'évaluer la possibilité pour cette théorie de rendre compte de textes d'apprenants, présentant potentiellement plus de problèmes de cohérence que des textes de scripteurs experts. Nous verrons dans cet article que la SDRT permet de construire des représentations structurées de façon hiérarchique. Nous souhaitions nous appuyer sur cette propriété des représentations pour (i) observer l'évolution du profil des représentations obtenues par l'annotation tout au long du cursus scolaire de l'école à la fin du collège ; (ii) évaluer l'hypothèse d'une corrélation du degré de structuration de la représentation du texte d'un scripteur avec son degré de compétence textuelle.

Nous présenterons dans un premier temps notre cadre théorique (section 2), puis notre méthodologie d'annotation du corpus (section 3) illustrée sur des textes d'élèves (section 4). Dans un second temps, nous donnerons les résultats de l'exploitation des annotations produites pour évaluer la cohérence des textes (section 5), prolongeant ainsi les deux publications réalisées au cours du projet E-Calm (Bras et al. 2021b, Bras et Vieu 2022).

2. Cadre théorique : une théorie de l'interface sémantique/pragmatique

Notre cadre théorique, la Segmented Discourse Representation Theory (Asher et Lascarides 2003)[2], est une théorie logico-référentielle de l'interface sémantique/pragmatique, prolongeant la Discourse Representation Theory (Kamp et Reyle 1993). Elle définit de façon formelle ce qu'est une représentation de discours cohérente et offre une méthode opératoire de construction des représentations des textes. Les représentations des textes, segmentés en Unités de Discours Élémentaires (UDE), sont articulées par des Relations de Discours (RD). La construction de la représentation se fait de façon récursive, UDE après UDE, et consiste principalement à déterminer le point d'attachement de l'UDE courante, ainsi que la RD réalisant cet attachement. Les RD sont classées en deux grandes catégories : les RD subordonnantes

[2] Voir (Busquets et al. 2001) pour une présentation en français.

(Explication, Elaboration, Arrière-Plan...) qui confèrent à la représentation du discours, avec les enchâssements d'UDE dans des segments complexes, sa structure hiérarchique ; les RD coordonnantes (Narration, Continuation, Résultat...) qui permettent le développement de la représentation dans un même niveau hiérarchique. L'attachement d'une UDE par une RD à la partie de la représentation déjà construite permet ensuite de procéder à la résolution des anaphores et autres équations incomplètes (traduisant les relations de cohésion) figurant dans la représentation du contenu propositionnel de l'UDE en cours de traitement. On représente ainsi les liens de cohérence et de cohésion à deux niveaux différents tout en rendant compte de leur interdépendance (qui n'est pas une condition nécessaire et suffisante, cf. Charolles 1995).

Comme indiqué en section 1, la SDRT a été mise à l'épreuve pour la première fois sur des textes d'apprenants dans le cadre du projet E-Calm (Bras et Vieu 2022). Il s'agissait pour nous d'évaluer la possibilité pour cette théorie de rendre compte de tels textes, et d'exploiter ses propriétés pour observer les représentations obtenues grâce aux annotations. En l'état actuel de la théorie, le processus de construction des SDRS s'arrête à la première impossibilité d'attachement d'une UDE à la représentation en cours de construction. Ce blocage du processus équivaut à évaluer le texte comme étant incohérent du point de vue du récepteur. Dans le processus d'annotation mis au point dans le projet E-Calm, nous continuons la construction au-delà des blocages pour tenter de mesurer le degré d'incohérence, ce qui impose d'étendre la théorie, pour pouvoir notamment typer et quantifier les points d'incohérence (Bras et Vieu 2022). Cela nous conduit à distinguer trois types de problèmes de cohérence : l'impossibilité d'attacher une UDE au reste de la structure ; l'impossibilité d'inférer une Relation de Discours entre un point d'attachement valide et une UDE ; l'identification d'une ou de plusieurs incohérences locales. Nous donnerons des exemples de ces problèmes en sections 4 et 5.

3. Méthodologie d'annotation de la cohérence dans les textes d'élèves

Le corpus annoté comporte 12 textes par niveau choisi, CE2, 6ème et 3ème, soit un échantillon de 36 textes choisis au sein du corpus RESOLCO rassemblant 385 textes manuscrits transcrits et codés en XML. L'annotation proprement dite procède en quatre étapes au sein de la chaine globale de traitement de chaque texte (cf. Figure 1) :

- La première étape consiste en une segmentation en Unités de Discours Élémentaires (UDE). La méthode de segmentation définie dans des projets antérieurs pour des textes

d'experts comme ANNODIS (Muller et al. 2012) a été largement modifiée pour tenir compte de la fiabilité moindre des critères ponctuationnels et syntaxiques dans les textes d'élèves, au bénéfice des critères sémantico-référentiels (description d'un événement ou d'un état, structure prédicative, éléments dotés d'une certaine autonomie discursive comme les cadratifs au sens de (Charolles 1997) traités en SDRT comme des introducteurs de nouveaux topiques (Vieu et al. 2005)).

- La deuxième étape, l'annotation en Relations de Discours (RD) permet ensuite de relier entre elles les UDE, i.e. les segments de base, pour construire des segments complexes éventuellement reliés entre eux aussi par des RD. Le jeu de 24 relations choisi est proche de celui de la SDRT et de celui utilisé dans le projet ANNODIS.
- Dans la troisième étape, parallèle à la deuxième, sont annotés différents problèmes de cohérence localisés sur des Points d'Incohérence (PI). Nous avons défini à cet effet un jeu d'étiquettes pour 12 types de Points d'Incohérence locale.
- La quatrième étape correspond à la construction d'une représentation graphique : nous avons développé un script Python permettant de générer automatiquement le graphe correspondant à l'annotation, à la fois pour appréhender la structure globale du texte et notamment visualiser son niveau de complexité structurelle, mais aussi pour repérer facilement d'éventuelles coquilles au cours du processus d'annotation même.

Figure 1 : Chaine de traitement des textes du corpus RESOLCO pour l'annotation de la cohérence

L'annotation du corpus de 36 textes a commencé par une phase exploratoire sur 12 textes, suivie d'une phase nominale sur 24 textes. Étant donnée la complexité de la tâche, surtout quand les textes sont incohérents, nous avons procédé par quadruple annotation avec harmonisation collective (au total environ 1000h d'annotation ont été décomptées). La phase exploratoire a permis la mise au point des manuels de segmentation en UDE et d'annotation en RD et PI (Bras et al. 2020-2022 a et b), à partir d'une première adaptation de la méthodologie d'annotation du projet ANNODIS (Muller et al. 2012) au corpus RESOLCO (Lala et al. 2017 a et b). Nous renvoyons le lecteur à (Bras et al. 2021b, Bras et Vieu 2022) pour une description plus complète du processus d'annotation. Nous présentons dans les sections suivantes les jeux

d'étiquettes utilisés (section 4) puis quelques exemples de textes segmentés et annotés en RD et PI (section 5).

4. Jeux d'étiquettes

4.1. Relations de Discours

Le jeu d'étiquettes de RD utilisé comporte 24 étiquettes issues 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 et Lascarides 2003) :

> Acquiescement (ACQ), Alternance (ALT), Arrière-plan (ARP), Attribution (ATT), But (BUT), Cadre (CAD), Commentaire (COM), Conditionnel (CND), Continuation (CTN), Contraste (CTR), Correction (COR), Élaboration (ELB), Élaboration d'entité (EEL), Élaboration 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).

Nous donnons ci-dessous des exemples d'annotation avec les RD les plus fréquentes dans le corpus.

Narration et Elaboration

[Le groupe courrue]31 [en crient]32 [et se refugit dans une grotte]33 (Joachim, 3ème)

NAR(31,33)

ELB(31,32)

Arrière-Plan et Elaboration d'Entité

[Il était une fois un pirate]1 [qui avait pour nom " le pirate blanc ",]2 [il avait l'œil borgne et un bras en moin,]3 [il avait sur le port les plus beau bateau de toute le continent.]4 (Robin, 6ème)

EEL(1,[2-3]) % un pirate

ARP(1,4)

Résultat

[La porte grincait.]2 [Elle avait trés peur.]3 (Pauline, CE2)

RES(2,3)

[Elle habitait dans cette maison depuis longtemps.]13[Elle croyait donc]14 [tout connaitre sur elle dans ses moindre secret.]15 (Joachim, 3ème)

RES(13,14)

Explication

[il y a une lois qui a été votée,]12 [depuis un évennement tragique.]13 (Charline, 3ème)

EXP(12,13)

[les enfants ne sortent plus la nuit.]7 [Parce qu'ils entander des bruit de grosse bête.]8 (Coline, CE2)

EXP(7,8)

4.2. Incohérences locales

Les Incohérences locales sont annotées avec un jeu de 12 étiquettes, elles codent les types de problèmes de cohérence locale les plus fréquents dans les textes, avec des formules de la forme :

RD(s_i, s_{i+1}) #type_incohérence($s_{i/i+1}$) [%% info]

où RD est l'étiquette de la relation de discours entre les segments s_i et s_{i+1} et où le symbole #permet d'identifier l'étiquette du type d'incohérence locale qui suit et le segment, s_i ou $s_{i+1,}$ sur lequel elle porte. Selon le type d'incohérence, des informations supplémentaires peuvent être annotées après le symbole %%. Nous donnons ci-dessous des exemples d'annotation des plus fréquentes.

#tense(s_i) (temps inapproprié / temps requis) : le temps verbal de l'éventualité principale du segment s_i est incompatible avec la RD

[Il se retourna]7 [en entendent ce grand bruit.]8 [Le plus âgée de fils appelle son père]9 [quand il vit ça.]10 (Oscar, 3ème)

NAR(7,9)

EXP(9,10) #tense(9)(PRE/PS)

#anaphore(s_i) (élément anaphorique) : il est impossible de trouver un antécédent à l'élément anaphorique de s_i indiqué en argument

[Il se retourna]7 [en entendent ce grand bruit.]8 [Le plus âgée de fils appelle son père]9 [quand il vit ça.]10 (Oscar, 3ème)

EXP(7,8) #anaphore(8) (ce grand bruit)

NAR(7,9)

EXP(9,10) #anaphore(10) (ça) %% cataphore avec ça impossible

#anaphore/amb(s_i) (élément anaphorique) : cas d'anaphore ambiguë, i.e. plusieurs antécédents possibles mais aucun moyen de déterminer lequel serait celui voulu par l'auteur

[Le soir venu]9 [tout le monde alla se couché]10 [mais Isabelle avait une idée derière la tête,]11 [amené Nils et Michelle dans La fôrét à côté de la maison]12 [Elle habitait dans cette maison depuis longtemps.]13 [Elle croyait donc]14 [tout connaitre sur elle dans ses moindre secret.]15 (Joachim, 3ème)

RES(13,14)

ATT(14,15) #anaphore/amb(15) (elle) %% maison ou forêt ?

#presupposition(s_i) (élément présuppositionnel) : il est impossible d'accommoder en contexte l'élément présuppositionnel de s_i indiqué en argument

[Il était une fois .un petit garçon [nommé Roméo.]2 Et sa voisine Julliette]1*P…. [Depuis que les deux enfants se sont vu]5 (Gabriel, CE2)

ELB(1,5) #presupposition(5) (les enfants se sont vus)

[Depuis cette aventure,]10 [les enfants ne sortent plus la nuit.]11 (Pauline, CE2)

CAD(10,11) #presupposition(11) (ne plus) %% absence d'evt. de type sortir la nuit

Nous ne pouvons illustrer les emplois des autres étiquettes dans l'espace limité de cet article, nous nous contentons de les lister ci-dessous et de renvoyer à (Bras et Vieu 2022, Bras et al. 2021b) pour plus d'explications :

- **#sem(s_i) :** effets sémantiques de la RD incohérents avec certains éléments du contexte
- **#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.
- **#structure d'information(s_i)** : structure d'information du segment s_i inappropriée en contexte

- **#progthem(s_i) :** le segment s_i induit une rupture dans la progression thématique du texte
- **#implicite(s_i) :** information implicite dans s_i non récupérable
- **#syntaxe(s_i) :** par exemple erreur sur sujet de l'infinitive s_i qui n'est pas celui qu'on reconstruit avec la principale (incompatibilité avec la RD qu'on rétablit en changeant le sujet)
- **#pop-up(s_i) :** blocage d'un attachement de s_i plus haut dans la structure par un élément incompatible

La section suivante propose trois exemples de textes annotés incluant certains des extraits analysés dans cette section.

5. Exemples d'annotation

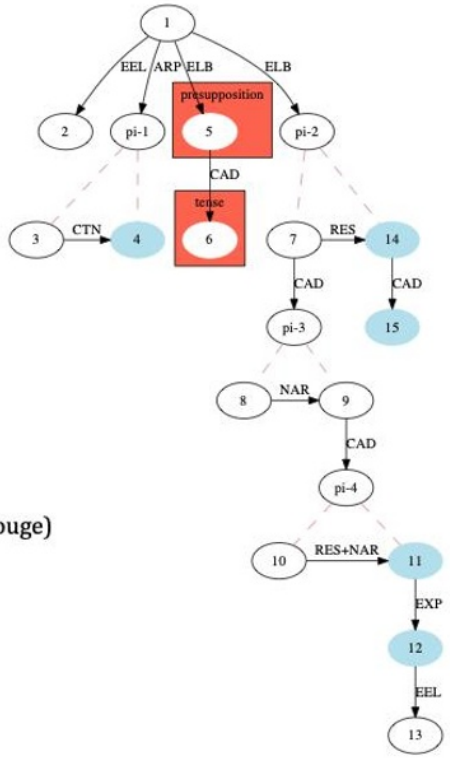
Le texte de Gabriel, CE2, ci-dessous, nous permet d'illustrer les principes de segmentation en UDE utilisés : sont segmentés les propositions, les adverbiaux détachés, les appositions, sans nécessairement tenir compte de la ponctuation présente. Ainsi, quand la ponctuation de l'élève aurait induit une segmentation différente de celle qui a été retenue, parce que la syntaxe et/ou la sémantique nous donnait des indices plus fiables, nous avons décidé de ne pas tenir compte de cette ponctuation et de l'annoter avec le symbole *P. Ici, par exemple, nous avons considéré que "Et sa voisine Juliette" était dans la portée de "Il était une fois". Le symbole *S permet de signaler un problème d'ordre syntaxique, par exemple un mot manquant (cf. UDE 9 où nous avons supposé que l'élève voulait dire "Et à un moment").

Texte de Gabriel, CE2
[Il était une fois .un petit garçon [nommé
Roméo.]2 Et sa voisine Julliette]1*P [qui abitait un
manoir.]3 [Elle habitait dans cette maison depuis longtemps.]4
[Depuis que les deux enfants se sont vu]5
[ils sont tombé directement amoureux.]6 [Un jour]7
[ils sont aller dans la forêt en pleine nuit.]8
[Et un moment]9*S [Roméo entendit un grand bruit]10
[Il se retourna]11 [en entendant ce grand bruit.]12 [C'était un loup]13
[Depuis cette aventure,]14 [les enfants ne sortent plus la nuit.]15

Nous donnons en Fig.2 l'annotation en RD, en partie gauche, avec en rouge l'annotation des incohérences locales (les % signalent des commentaires d'annotation). Dans la partie droite de la Fig.2, nous avons le graphe généré automatiquement à partir de cette annotation, dans lequel les segments complexes, qui regroupent les segments qu'ils dominent, sont étiquetés « pi-n ». Les ovales bleus signalent les UDE représentant les phrases imposées par la consigne et les rectangles rouges les endroits où ont été annotées des incohérences locales. Nous pouvons observer la structure assez profonde du schéma, et le petit nombre d'incohérences locales. Ce texte est un des plus cohérents parmi les textes de CE2 de notre corpus (à titre de comparaison, voir le texte de Pauline, CE2, analysé dans (Garcia-Debanc et al., ce volume)).

Figure 2 : Annotation du texte de Gabriel, CE2, et schéma de la structure hiérarchique de la représentation

EEL(1,2) % un petit garçon
ARP(1,[3-4])
ELB(1,5) #presupposition(5) (les enfants se sont vus)
CAD(5,6) #tense(6) (PC/PRE)
ELB(1,[7-15]) %% sans continuation entre 5 et 7
CAD(7,[8-13])
NAR(8,9)
CAD(9,[10-13])
RES(10,11)
EXP(11,12)
EEL(12,13) % ce grand bruit (avec présupposition que le loup bouge)
RES(7,14)
CAD(14,15)



Nous poursuivons avec un texte d'élève de 3ème, Joachim, trois fois plus long que le texte de Gabriel en nombre d'UDE, ce qui n'est pas surprenant. Dans le graphe généré à partir de son annotation, en Fig.3, on observe une structure hiérarchique beaucoup plus marquée que dans celle du texte de Gabriel en Fig. 2. On remarque aussi que les points d'incohérence, en rouge, se situent au niveau des 2ème et 3ème phrases imposées, ce qui est fréquent dans le corpus.

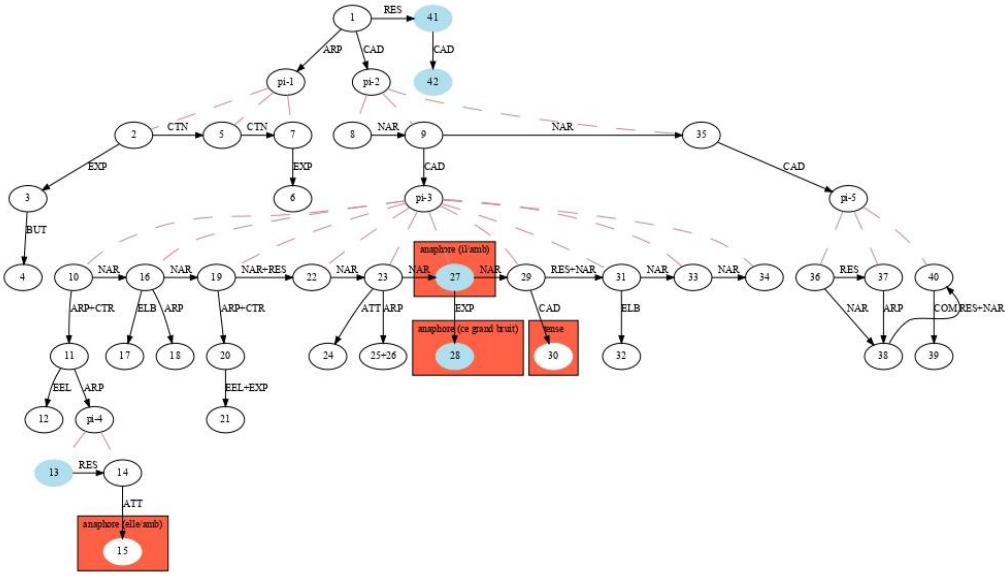
Texte de Joachim, 3ème

[Ce weak-end,]1 [Isabelle était si content,]2 [ses cousin venaint]3
[lui rendre visite.]4 [Elle avait tous préparé,]5 [impatiente]6
[elle regardé par la fenêtre leur arrivé.]7 [Ses cousins
Nils et Michelle vinirent pour le repas.]8 [Le soir venu]9
[tout le monde alla se couché]10 [mais Isabelle avait une
idée derière la tête,]11 [amené Nils et Michelle dans
La fôrét à côté de la maison]12 [Elle habitait dans cette maison depuis longtemps.]13
[Elle croyait donc]14 [tout connaitre sur elle dans
ses moindre secret.]15 [Les 3 enfamts sortire par la
fenêtre]16 [sans que le parents ne le remarque,]17 [il faiit
nuit noir.]18 [Le groupe s'enfonça dans la fôret]19 [mais Michelle
[pas trés courageux]21 avait peur,]20 [il commenca à paniqué]22
[Il dit]23 [entendre un bruit,]24 [Nils le prennait juste pour
une chochotte]25 [qui voulait juste fair demi-tour.]26
[Il se retourna]27 [en entendant ce grand bruit.]28 [Avec leur imagination
d'enfant,]29 [chacun pensaient a un monstre ou un esprit
surnaturel.]30 [Le groupe courrue]31 [en crient]32 [et se refugit
dans une grotte]33 [où ils passerent la nuit.]34 [Le matin]35 [les
parent retrouvèrent les lits des enfant vide]36 [inquiet]37 [ils
cherchèrent dans la fôret]38 [et par chance]39 [retrouverent les
enfants :]40 [Depuis cette aventure,]41 [les enfants ne sortent plus la nuit.]42

Les textes de Gabriel et Joachim permettent d'illustrer l'identification de quelques-uns des types d'incohérences locales listés en section 4. Nous illustrons ci-dessous, avec un dernier exemple, le texte d'Oscar, 3ème, un autre type d'incohérence possible (cf. fin de la section 2), à savoir l'impossibilité d'attacher un segment de discours au reste de la structure, qu'une RD soit identifiable ou non, ce qui se traduit par les formules suivantes dans l'annotation :

$?(?, s_{i+1})$ ou $RD(?, s_{i+1})$

Figure 3 : Schéma de la structure hiérarchique de la représentation du texte de Joachim, 3ème

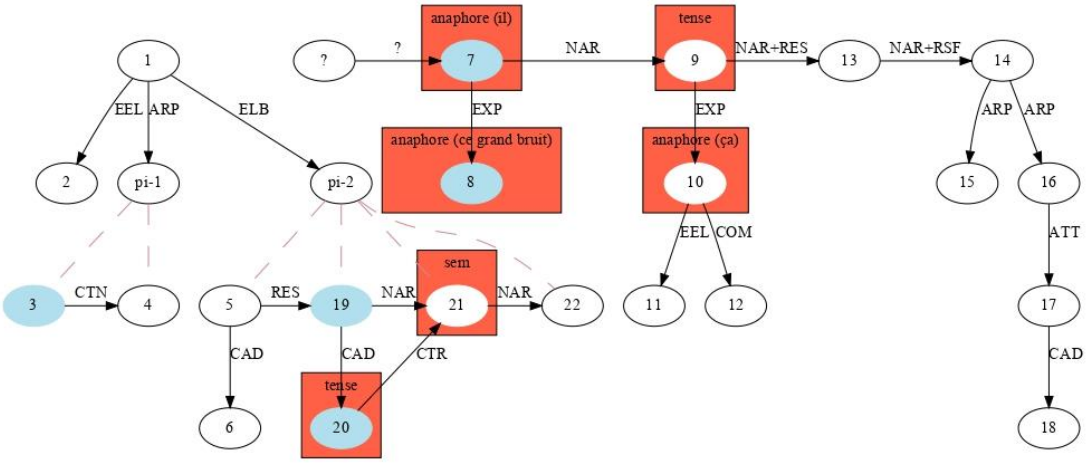


Texte d'Oscar, 3ème

[Il était une fois une grande famille et une grande maison]1 [perdu dans une foret.]2 [Elle habitait dans cette maison depuis longtemps.]3 [Il arrivait parfois des bruit retentissait.]4*S [Un jour,]5 [il se passa quel chose d'estraordinaire.]6*S [Il se retourna]7 [en entendent ce grand bruit.]8 [Le plus âgée de fils appelle son père]9 [quand il vit ça.]10 [Les arbres montait au ciel]11 [cela arriva pendant la nuit.]12 [Le père regarde par la feunetre]13 [il voit une soucoupe]14 [et les arbre montait droit dedans.]15 [Le père sachant]16*S [que des nuit au par avant]17 [des amis avait vue une soucoupe.]18 [Depuis cette aventure,]19 [les enfant ne sortent plus la nuit.]20 [Jusqu'au jour ou le fils le plus jeune sorti pendant la nuit]21 [et se fit enlevé par la soucoupe volante]22

L'annotation du texte d'Oscar (Fig. 4) met en évidence une impossibilité d'attacher la représentation de la deuxième phrase imposée par la consigne, les UDE 7 et 8, au reste de la structure, sans possibilité d'inférer une RD. Cela traduit une rupture majeure dans la cohérence, fréquente dans ce corpus, parce que provoquée par la contrainte d'insertion d'une phrase requérant la résolution de deux expressions anaphoriques (cf. (Bras et Vieu 2022) pour une explication détaillée des mécanismes d'inférence conduisant à cette situation et (Garcia-Debanc 2020) au sujet de la résolution de "ce grand bruit").

Figure 4 : Schéma de la structure hiérarchique de la représentation du texte de Oscar, 3ème



6. Exploitation des annotations de la cohérence

L'annotation des 36 textes a abouti à la délimitation de 1042 UDE, à l'identification de 1017 RD et de 207 Points d'Incohérence (locale, d'attachement et de RD). En considérant de façon globale l'ensemble des UDE du corpus, le nombre d'UDE concernées par des problèmes de cohérence s'élève donc à environ 20 %. Ces annotations ouvrent la voie à différents types d'explorations que nous listons et résumons ci-après avant de nous focaliser sur l'analyse des mesures de la cohérence.

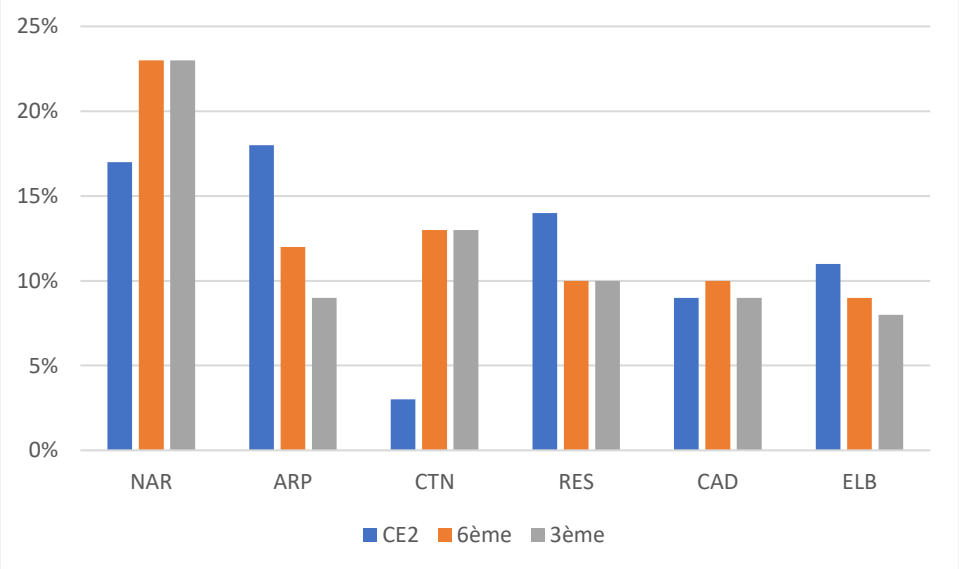
6.1. Longueur des textes en Unités de Discours Élémentaires

La segmentation des textes permet de mesurer leur longueur en nombre d'UDE, plutôt qu'en nombre de mots. La longueur moyenne ainsi calculée est de 28,9 UDE par texte sur l'ensemble du corpus, elle augmente fortement avec l'avancée en niveau scolaire : 16,5 UDE en CE2 ; 30 en 6ème et 40,2 en 3ème, mais l'écart-type élevé (6,3 en CE2 ; 11,6 en 6ème et 17,9 en 3ème) montre que la longueur des textes est très variable et que plus les textes sont longs, plus leur longueur varie d'un élève à l'autre. Le nombre très modeste de textes par niveau nous invite cependant à la prudence sur ces conclusions.

6.2. Annotation en Relations de Discours

Il est possible d'étudier l'évolution du jeu des relations de discours impliquées dans l'interprétation des textes aussi bien du point de vue de sa diversité que du changement de proportion de certaines familles de relations par rapport à d'autres. Ainsi on peut mettre en évidence par exemple que le nombre moyen de relations distinctes par texte augmente entre le CE2 et la 3ème (7,4 RD en CE2 ; 10,3 en 6ème ; 11,6 en 3ème sur les 21 RD distinctes utilisées dans tout le corpus) sans que cela soit corrélé à la longueur des textes, et avec un usage de plus en plus homogène (l'écart-type baisse fortement : 2,33 RD en CE2 ; 1,79 en 6ème ; 1,26 en 3ème). Du point de vue de la diversité des relations employées, les résultats montrent que les relations majoritaires sont celles du système narratif (22 % pour Narration, 12 % pour Arrière-Plan, 11 % pour Continuation et Résultat, 9 % pour Cadre et Élaboration), sans surprise eu égard à la tâche d'écriture (Cf. Fig. 5). Au-delà des relations majoritaires, on observe des évolutions dans l'usage de certaines RD, par exemple, la présence de Résultat et d'Élaboration diminue régulièrement entre CE2 et 3ème, au profit notamment d'Explication et d'Élaboration d'Entité (qui n'apparaissent pas dans la Fig. 5 parce que moins représentées) qui deviennent plus fréquentes en 3ème. Le nombre de relations causales distinctes augmente, ce qui se traduit par une utilisation plus fréquente de la relation d'Explication en 3ème qu'en CE2 (Bras et al. 2021a). Continuation est en forte augmentation entre le CE2 et la 6ème et 3ème, ce qui traduit une présence croissante de segments complexes.

Figure 5 : Relations de Discours majoritaires



6.3. Annotation en Points d'Incohérence

Les annotations produites permettent aussi d'analyser les types de problèmes de cohérence rencontrés à la lecture des textes et d'observer leur évolution de la fin du cycle 2 à la fin du cycle 4. Nous nous sommes appuyées pour cette analyse sur trois types d'indicateurs de problèmes de cohérence, ou points d'incohérence : le nombre d'incohérences locales annotées avec les étiquettes présentées en section 4, le nombre d'UDE non rattachées à une autre UDE, le nombre d'impossibilités d'inférence d'une relation de discours.

6.3.1. Taux d'incohérence par texte

Nous avons calculé pour chaque texte un **«** taux d'incohérence **»** égal au nombre de points d'incohérence rencontrés, tous types confondus, par UDE. La moyenne de ces taux d'incohérence diminue fortement entre les textes de CE2 et ceux de 6ème, elle est stable entre les textes de 6ème et ceux de 3ème ; 0,5 en CE2 ; 0,16 en 6ème et 3ème, avec un écart-type qui se resserre (0,39 en CE2 ; 0,12 en 6ème ; 0,08 en 3ème), révélant une homogénéisation du taux et moins de variation d'un élève à l'autre en 3ème.

6.3.2. Répartition des points d'incohérence

Le tableau 1 fournit la répartition des points d'incohérence analysés en 3.3.1 (en colonne 3) pour chacun des trois types d'indicateurs (en colonnes 4, 5, 6). Il montre (i) que le nombre d'incohérences locales par UDE diminue de façon très significative entre les textes de CE2 et ceux de 6ème ; (ii) que le nombre de RD non inférables diminue aussi entre les textes de CE2 et ceux de 6ème ; (iii) que le nombre de segments non attachables diminue encore plus entre les textes de CE2 et ceux de 6ème, cette dernière observation étant à compléter par la donnée du nombre de textes qui présentent des problèmes d'attachement (8 sur 12 en CE2 ; 2 sur 12 en 6ème ; 3 sur 12 en 3ème), également en forte baisse.

Tableau 1 : Répartition des Points d'Incohérence

	Longueur moyenne des textes en UDE	Nombre de Points d'Incohérence par UDE	Dont incohérences locales #	Dont RD non inférables	Dont Segments non attachables
CE2	16,5	0,5	0,41	0,03	0,07
6ème	30	0,16	0,14	0,01	0,02
3ème	40,2	0,16	0,14	0,01	0,01
Corpus	28,9	0,28	0,24	0,02	0,03

Les incohérences locales se répartissent comme indiqué dans le tableau 2 (avec un taux par UDE basé sur la totalité des étiquettes # par niveau, et pas sur la moyenne des taux par texte comme dans le tableau 1). Pour les trois types d'incohérence les plus fréquents présentés, on constate comme précédemment des diminutions très significatives des incohérences entre CE2 et 6ème. La baisse continue entre 6ème et 3ème, sauf pour les temps verbaux, pour lesquels la prise de risque est plus importante dans les textes de 3ème avec des textes plus longs et respectant généralement mieux les contraintes de la consigne induisant l'emploi de l'imparfait, du passé simple pour terminer au présent.

Tableau 2 : Répartition des incohérences locales

	incohérences locales #	#anaphore	#tense	#présupposition
CE2	0,34	0,17	0,08	0,04
6ème	0,12	0,05	0,02	0,02
3ème	0,12	0,004	0,05	0,01

Conclusion

Nous avons présenté dans cet article la méthodologie et les étapes de traitement qui ont abouti à la création d'un corpus de 36 textes d'élèves de CE2, 6ème et 3ème issus du corpus RESOLCO, segmentés en Unités de Discours Élémentaires, annotés en Relations de Discours et en problèmes de cohérence, et accompagnés des graphes présentant leur structure de discours.

Ce corpus annoté permet une analyse de l'évolution de la cohérence dans les textes entre l'école primaire et le collège à partir de l'annotation d'incohérences locales, d'impossibilités d'inférer une Relation de Discours, ou d'impossibilités d'attacher la représentation d'un segment au reste de la représentation. Ces indicateurs permettent de mesurer le degré de cohérence de chaque texte. L'analyse des annotations met en évidence une tendance à la baisse de ces trois types d'indicateurs cumulés en un taux de « points d'incohérence », rapporté à la longueur du texte en nombre d'UDE. La diminution du taux d'incohérence présente un palier

significatif entre le CE2 et la 6ème, qui est à corréler avec un saut majeur de compétences rédactionnelles entre la fin du cycle 2 et la fin du cycle 3. On observe par ailleurs une augmentation du nombre de Relations de Discours distinctes de discours par texte.

Les mesures de cohérence montrent une maitrise croissante de la compétence rédactionnelle induisant une amélioration de la cohérence discursive. Le nombre de textes annotés, très modeste pour une analyse quantitative, appelle à la prudence. Nous avons néanmoins constaté que les tendances observées sur les 24 premiers textes annotés étaient confirmées sur les 36 textes du corpus final, ce qui nous conduit à valider l'hypothèse d'une évolution du profil des représentations obtenues par l'annotation tout au long du cursus scolaire (école, collège).

Une hypothèse connexe reste à évaluer, c'est celle d'une corrélation du degré de structuration de la représentation du texte d'un scripteur avec son degré de compétence textuelle. Nous pouvons pour le moment nous appuyer sur le nombre de relations de Continuation, qui témoignent d'une structure hiérarchique de plus en plus marquée. Il faudra développer aussi un indicateur prenant en compte la profondeur des graphes obtenus pour évaluer cette hypothèse.

Un autre prolongement de ce travail est la confrontation de nos annotations avec des annotations/interventions d'enseignants afin d'évaluer la pertinence de notre méthodologie, et d'envisager de la mettre au service des enseignants. Nous renvoyons le lecteur à (Garcia-Debanc et al., ce volume) pour une telle expérience avec des enseignants en formation continue.

Au-delà de l'analyse de l'évolution de la cohérence dans les textes d'élèves, notre ambition était aussi d'évaluer la capacité de la SDRT à rendre compte de textes d'apprenants, présentant plus de problèmes de cohérence que des textes de scripteurs experts. Nous avons, à travers la contribution au projet E-Calm rapportée dans cet article, montré que cette théorie pouvait permettre d'analyser de tels textes moyennant une mise en œuvre d'un principe de coopérativité étendu, et un enrichissement significatif du dispositif d'annotation, avec l'annotation des Points d'Incohérence. Cette extension de la théorie à travers la méthodologie d'annotation développée nous semble représenter une avancée significative pour l'analyse de la cohérence et pour la SDRT elle-même.

Bibliographie

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 & J. Pustejovsky (Eds.), *Handbook of Linguistic Annotation*, Springer, pp. 1241–1264.

Bras M., Le Draoulec A., Asher N. (2009), "A formal analysis of the French Temporal Connective *alors*", in B. Behrens & C. Fabricius-Hansen (eds.), *Structuring Information in Discourse: The Explicit/Implicit Dimension*, OSLa, 1, 149-170.

Bras, M., Joret, M., Pépin-Boutin, A., Vieu, L. (2021a), « Annotation des relations causales dans un corpus de textes d'élèves d'école et collège », Colloque international *L'expression de la causalité en langue maternelle et en langue étrangère*, Lublin, Pologne, 20-21 mai 2021.

Bras, M., Vieu L., Joret, M., Pépin-Boutin, A., Poujade, C., Roze, C. (2021b), « Vers un corpus de textes d'élèves annoté en relations de discours », *Langue Française, 211-3, 115-130.*

Bras, M., Vieu L., (2022), « 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) *Connecter et segmenter à l'écrit. Ponctuation et opérateurs linguistiques : deux défis pour l'enseignement.* Peter Lang, Berne, pp. 25-55.

Bras, M., Vieu, L., Poujade, C., Roze, C. (2020-2022a). *Manuel de segmentation en unités de discours élémentaires du corpus RESOLCO,* Rapport interne, Toulouse : CLLE-ERSS.

Bras, M., Vieu, L., Roze, C., Joret, M., Pépin-Boutin, A. (2020-2022b). *Manuel d'annotation en relations de discours du corpus RESOLCO*, Rapport interne, Toulouse : CLLE-ERSS.

Busquets, J., Vieu, L., Asher, N. (2001), « La SDRT : Une approche de la cohérence du discours dans la tradition de la sémantique dynamique », *Verbum*, XXIII, 1,73–101.

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., 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., Ho-Dac, M., Bras, M., Rebeyrolle, J. (2017) « Vers l'annotation discursive de textes d'élèves », *Corpus* [En ligne], 16 | 2017.

Garcia-Debanc, C., (2020), « Écrire et réécrire pour résoudre des problèmes de cohésion textuelle : quel est donc ce grand bruit dans le corpus RÉSOLCO ? Analyse de récits d'élèves de 9 à 15 ans », in F. Neveu et al. (éds), 7e *Congrès Mondial de Linguistique Française* (Montpellier, France), SHS Web of Conferences 78, Les Ulis, EDP Sciences, #07021.

Garcia-Debanc, C., Bras, M., Vieu, L. (2023), « Annotation de la cohérence dans des textes d'élèves et jugement de cohérence d'enseignants du primaire », ce volume.

Kamp, H. & Reyle, U. (1993). *From Discourse to Logic.* Kluwer.

Lala, M., Bras, M., Garcia-Debanc, C. (2017a). *Manuel de segmentation en unités de discours élémentaires du projet RESOLCO* (Rapport interne), Toulouse : CLLE-ERSS.

- (2017b). *Manuel d'annotation en relations de discours du projet RESOLCO* (Rapport interne), Toulouse : CLLE-ERSS.

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.

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