ResolCo un corpus de manuscrits d'élèves et d'étudiants pour l'étude de la cohérence

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

▶ **To cite this version:**

Lydia-Mai Ho-Dac, Silvia Federzoni, Myriam Bras, Josette Rebeyrolle, Claudine Garcia-Debanc. ResolCo un corpus de manuscrits d'élèves et d'étudiants pour l'étude de la cohérence. 10èmes Journées Internationale de la Linguistique de Corpus, Nov 2019, Grenoble, France. hal-02877122

HAL Id: hal-02877122
https://hal.science/hal-02877122

Submitted on 22 Jun 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

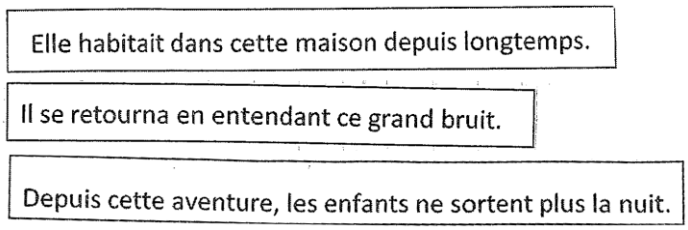
L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

ResolCo un corpus de manuscrits d'élèves et d'étudiants pour l'étude de la cohérence

Lydia-Mai Ho-Dac, Silvia Federzoni, Myriam Bras, Josette Rebeyrolle et Claudine Garcia-Debanc
Laboratoire CLLE, Université Toulouse Jean Jaurès
lydia-mai.ho-dac@univ-tlse2.fr, silvia.ferderzoni@univ-tlse2.fr, myriam.bras@univ-tlse2.fr, claudine.garcia-debanc@univ-tlse2.f, josette.rebeyrolle@univ-tlse2.fr

La ressource ResolCo[1] est un corpus constitué de transcriptions de manuscrits d'élèves et d'étudiants enrichi par des annotations concernant les traces du processus d'écriture, les variantes orthographiques observées et certaines structures discursives. L'originalité de la ressource ResolCo réside dans son protocole de collecte et principalement dans sa consigne qui est la suivante :

Racontez une histoire dans laquelle vous insérerez, séparément et dans l'ordre donné, les trois phrases suivantes



(découpez et collez les bandelettes dans votre texte) :

Cette consigne a été imaginée pour provoquer la mise en œuvre de stratégies de Résolution de problèmes de Cohérence (Charolles 1994, Garcia-Debanc et al. 2017). Le corpus ResolCo fournit ainsi un terrain privilégié pour l'étude de l'organisation du discours et des indices de cohésion à différents âges d'acquisition de la langue écrite. Les stratégies mis en jeux sont principalement :

- les stratégies utilisées pour introduire des référents de type variés (humains – *Elle, Il, les enfants* ; inanimés – c*ette maison, ce grand bruit*; événementiel – *cette aventure*) et gérer la compétition et l'interférence entre les continuités référentielles (cf. Ariel 1990 et Givon 1983) ;
- des stratégies de planification du discours (amorçage de la phrase-fermoir (cf. Marandin 1986) et gestion de l'anaphore résumante – *cette aventure*) ;
- des stratégies de gestion des temps verbaux, chaque phrase présentant un temps du récit différent ;
- la production d'un texte de type narratif sans contrainte de genre.

Les productions écrites ont été recueillies dans différentes écoles primaires, collèges et universités des régions Occitanie et Île de France. Les points de collectes retenus permettent de disposer d'un nombre comparable de copies par niveau et pour représenter différents

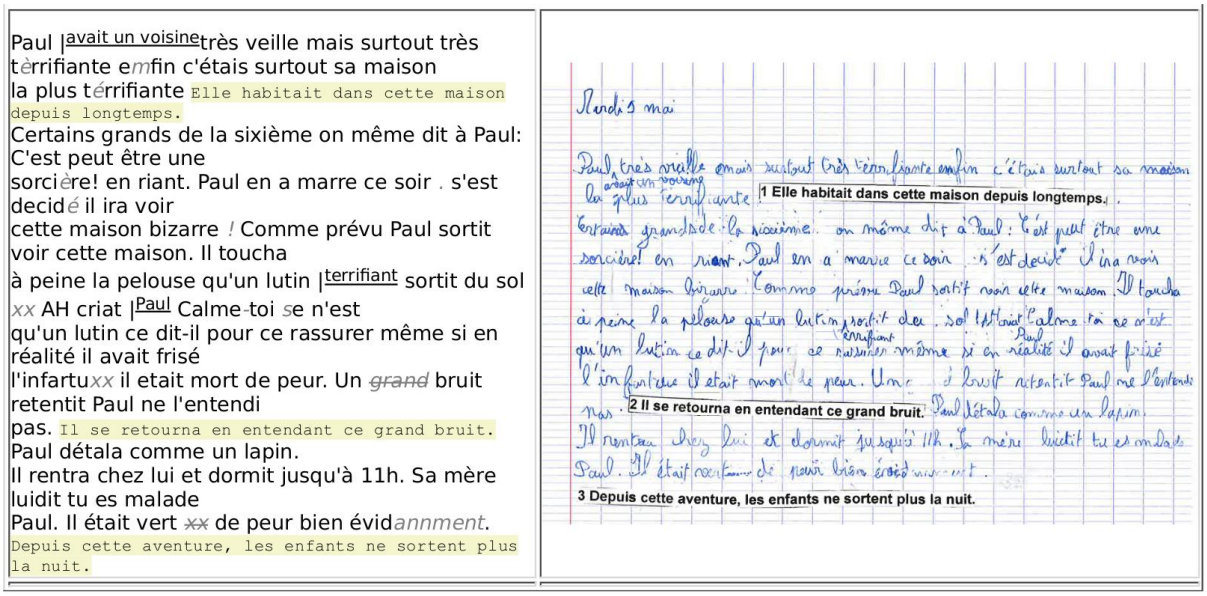
1 La ressource ResolCo fait partie du projet ANR É:Calm, Écritures scolaires : Corpus, Analyses Linguistiques, Modélisations didactiques. Voir http://e-calm.huma-num.fr/

milieux scolaires (urbain, rural, ZEP). Toutes les données récoltées seront mises à disposition de la communauté sous licence Creative Commons By-NC-SA 3.0 (Paternité, usage non commercial, partage à l'identique).

La première étape de constitution du corpus consiste en la transcription et l'annotation des traces écrites avec pour objectif de reproduire le plus fidèlement possible la copie originale et sa mise en page. Pour ce faire, les transcriptions sont accompagnées d'annotations indiquant d'éventuelles ratures, la présence des dessins des élèves, etc. Le format choisi pour la transcription est le format XML et la norme TEI-P5 qui fournit à la fois un modèle pour l'encodage des métadonnées et de nombreux éléments dédiés à l'annotation des traces écrites.

Les métadonnées retenues concernent le contexte de production de la copie (établissement, niveau scolaire, année scolaire, consigne d'écriture, etc.). La norme TEI-P5 permet également d'indiquer les étapes de digitalisation des données. Concernant le corps de texte, en plus de l'indication de l'emplacement des bandelettes, les annotations signalent des phénomènes de mise en page et d'écriture manuscrite.

Les principales conventions de codage TEI-P5 des traces écrites adoptées sont les suivantes : chaque ligne sur la copie est encodée par l'élément <lb>, les paragraphes correspondent à l'élément <p>, tout trace de texte révisé (insertion et/ou suppression par rature ou effacement visible) est renseignée par l'élément <mod> et visualisée au moyen d'un texte en exposant et/ou barré, toute portion de texte illisible ou à la transcription incertaine est indiquée (<gap>



ou <unclear>) est affichée sous forme d'italiques ou de séquences de "xx".

figure . 1 Transcription des traces d'écriture et scan d'une copie

Afin de s'assurer de la qualité des données, toutes les transcriptions ont été vérifiées par un annotateur différent du transcripteur. Cette phase de vérification s'est appuyée sur un affichage simultané du scan de la copie originale et d'une visualisation du texte transcrit obtenue par transformation XSLT depuis le fichier XML, comme l'illustre la figure 1.

La transcription XML est ensuite converti au format Glozz (Widlöcher & Mathet 2009) pour permettre l'annotation des variantes orthographiques et la génération d'une version normée permettant l'application d'outils du TAL et l'annotation des structures discursives. Le tableau

1 donne un aperçu quantitatif de l'état actuel du corpus ResolCo en indiquant pour chaque niveau : le nombre de textes transcrits, de ratures, de textes normés et d'erreurs d'orthographe.

	ratures	textes transcrits	ratures/texte	corrections	textes normés	corrections/texte
total	**1590**	**350**	**5**	**1336**	**132**	**10**
CE2	97	36	3	199	13	15
CM1	214	39	5	29	25	1
CM2	299	94	3	371	35	11
6EME	396	85	5	687	29	24
4EME	204	47	4	0	0	na
3EME	276	36	8	50	17	3
Master	91	13	7	0	13	0

Tableau 1 : aperçu quantitatif de l'état actuel de la ressource ResolCo

La version normée est également associée à un étiquetage des catégories morphosyntaxiques et des relations syntaxiques produit par l'outil Talismane (Urieli 2013). La constitution d'un treebank par correction manuelle des analyses proposées par Talismane a permis une évaluation sur un échantillon de 13 220 tokens dont 11 706 mots (hors ponctuations).

Les résultats obtenus sont tout à fait acceptables avec une exactitude de 95,7 pour l'attribution des catégories morphosyntaxiques (11 203 token correctement étiquetés sur 11 706) et une efficacité au niveau de l'analyse des relations syntaxiques (i.e. détecter le gouverneur et le type de relation entre chaque token) de 97,5 pour la détection du bon gouverneur et de 90,7 pour la caractérisation des relations.

Cette ressource permet un ensemble d'analyses fournissant des données originales sur l'évolution des compétences scripturales liées notamment à l'orthographe, la syntaxe et le discours au fil de l'acquisition de l'écriture à l'école.

Références bibliographiques

Ariel, M. (1990). *Accessing noun phrase antecedents*. Routledge: London

Charolles, M. (1994). Cohésion, cohérence et pertinence du discours. *Travaux de Linguistique*, 29, 125-151

Garcia-Debanc, C., Ho-Dac, L.-M., Bras, M., Rebeyrolle, J. (2017). Vers l'annotation discursive de textes d'élèves. *Corpus*, 16.

Givón, T. (1983). Topic continuity in discourse: an introduction. In T. Givon (ed) *Topic continuity in discourse: a quantitative cross-language study*. John Benjamins: Amsterdam/Philadelphia, pp. 1-42

Marandin, J. M. (1986). *Ce est un autre*. L'interprétation anaphorique du syntagme démonstratif. *Langages*, 81, pp. 75-89.

Urieli, A. (2013). *Robust French syntax analysis: reconciling statistical methods and linguistic knowledge in the Talismane toolkit.* Thèse de doctorat, Université Toulouse le Mirail-Toulouse II.

Widlöcher, A., & Mathet, Y. (2009). La plate-forme Glozz: environnement d'annotation et d'exploration de corpus. In *Actes de la 16e Conférence Traitement Automatique des Langues Naturelles (TALN'09)*.