*Actes de CORIA-TALN-RJCRI-RECITAL 2025*
*Marseille, France, 30 juin au 4 juillet 2025*

# Analyse de la continuité référentielle dans le corpus d'écrits scolaires français et italien *Scolinter*

Martina Barletta[1, 2] Claude Ponton[1]
(1) Univ. Grenoble Alpes, LIDILEM, F-38000 Grenoble, France
(2) University of Milano-Bicocca, Department of Human Sciences for Education "Riccardo Massa", 20126 Milan, Italie
martina.barletta@univ-grenoble-alpes.fr,
claude.ponton@univ-grenoble-alpes.fr

## RÉSUMÉ

Cet article présente une étude sur la continuité référentielle dans des écrits scolaires en français et en italien, en s'appuyant sur le corpus Scolinter. L'objectif est d'analyser les mécanismes de cohérence textuelle à l'école primaire et de comparer les stratégies utilisées dans les deux langues à travers l'annotation et l'analyse des chaines de continuité référentielle. Une campagne d'annotation a été menée sur 150 textes par langue (CE1 et CE2), et le corpus de référence obtenu suite à l'adjudication a fait l'objet d'une analyse présentée ici. Les résultats montrent des différences notables. Par exemple, en français, les pronoms personnels sont privilégiés, tandis qu'en italien, l'anaphore zéro est plus fréquente. L'étude met également en évidence une tendance commune dans l'introduction des référents, souvent par des syntagmes nominaux indéfinis suivis d'une reprise pronominale. En revanche, la densité référentielle ne varie pas significativement entre les niveaux scolaires. Ces analyses apportent un éclairage sur le développement des compétences rédactionnelles et les spécificités linguistiques influençant la gestion de la référence dans chaque langue.

## ABSTRACT

**Analysing referential continuity in the French and Italian corpus of school writing *Scolinter***

This article presents a study of topic continuity in French and Italian school texts, based on the Scolinter corpus. The study aims to analyze the mechanisms of textual coherence in primary school writing and to compare the strategies used in the two languages. An annotation campaign was carried out on 150 texts per language (CE1 and CE2), and the reference corpus obtained as a result of the curation process is the subject of the analysis presented here. The results show some remarkable differences. For example, personal pronouns are preferred in French, whereas zero anaphora is more frequent in Italian. The study also reveals a common trend in the introduction of referents, often by indefinite noun phrases followed by a pronominal mention. On the other hand, referential density does not vary significantly between grade levels. These analyses shed light on the development of writing skills and the linguistic peculiarities that influence reference management in each language.

MOTS-CLÉS : Cohérence textuelle, continuité référentielle, écrits scolaires, corpus annoté.

KEYWORDS: Textual coherence, topic continuity, student writings, annotated corpus.

Contribution originale

Actes des 32ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN), volume 1 : articles scientifiques originaux, pages 134–153.

# 1 Introduction

La constitution d'un corpus d'écrits scolaires représente une tâche complexe à plusieurs égards : d'un côté, le recueil nécessite souvent l'implication de plusieurs chercheurs et enseignants, en plus des élèves et des institutions ; de l'autre, il est souvent nécessaire de numériser ces écrits pour les rendre exploitables à l'aide d'outils de Traitement Automatique des Langues, notamment lorsque les textes sont manuscrits. En plus de la sauvegarde de l'image du texte sous forme de scan, cette étape de numérisation consiste en une transcription souvent associée à la création d'une version normalisée du texte [1]. Ces traitements, bien qu'essentiels, demandent un investissement conséquent en termes de temps, de ressources humaines et de moyens techniques, ce qui constitue un frein majeur à leur mise en œuvre systématique.

La description d'un phénomène dans son contexte réel de production constitue cependant un enjeu essentiel pour la linguistique et, dans le cas d'écrits représentant des phases d'apprentissage, aussi pour la didactique. Ces descriptions fondées sur l'analyse de corpus permettent notamment d'éclairer le développement des compétences rédactionnelles des élèves et d'aider à ajuster les attentes des enseignants quant aux normes d'écriture propres à chaque niveau scolaire (Garcia-Debanc *et al.*, 2021), ou même de soutenir l'enseignement de stratégies efficaces pour l'apprentissage de l'écriture. Mais cela n'est possible que si l'on parvient à établir « une cartographie de l'emploi de ces formes linguistiques (...) en fonction du niveau scolaire et du degré d'expertise du rédacteur » (Garcia-Debanc *et al.*, 2021, p. 2).

Dans cette perspective de description linguistique, notre étude s'appuie sur le corpus Scolinter [2] et porte sur l'annotation de la continuité référentielle sur une partie des textes qui composent ce corpus, en vue d'une analyse du développement des phénomènes de cohérence textuelle en français et en italien à l'école primaire. Ce travail nous permettra de saisir un des niveaux qui permettent de caractériser et donc de décrire les mécanismes de fonctionnement de la cohérence textuelle. De plus, cette analyse comparative met en évidence tant des similarités que des divergences dans les stratégies de construction de la cohérence textuelle dans les deux langues.

# 2 Annotation de la continuité référentielle dans un corpus d'écrits scolaires

## 2.1 Définition de la continuité référentielle

Si de nombreux corpus et travaux, en France et à l'international, ont eu comme objectif l'annotation des expressions réalisant la présence des référents dans les textes [3], très peu de projets ont concerné l'analyse de ce phénomène dans des corpus de scripteurs en phase d'acquisition de la langue écrite. À notre connaissance, un seul projet de ce type a été mené en France : le corpus RésolCo (Garcia-Debanc *et al.*, 2019, 2021). Ce corpus se compose de 400 textes recueillis dans des classes allant du CE2 à l'université, produits à partir d'une consigne visant une tâche de résolution de problèmes de

1. Dans le cadre de notre projet, la normalisation consiste en une réécriture au plus proche du texte de l'élève avec un simple rétablissement des normes orthographiques et de segmentation (Wolfarth *et al.*, 2018b).

2. Le corpus est disponible à la consultation au lien suivant `https://scoledit.org/scolinter/`.

3. En France, les projets plus récents ont abouti aux corpus Annodis (Péry-Woodley *et al.*, 2011), Ancor (Muzerelle *et al.*, 2013) et Democrat (Landragin, 2016). À l'international, on peut notamment citer le corpus OntoNotes 5.0 (Weischedel, Ralph *et al.*, 2013), ARRAU (Poesio, Massimo *et al.*, 2013) et ANCORA (Recasens & Martí, 2010).

cohésion textuelle. Il a ensuite été annoté en continuité référentielle, en se limitant aux trois référents imposés par la consigne (Garcia-Debanc *et al.*, 2021) [4]. Nous nous sommes appuyés sur le modèle d'annotation adopté dans ce corpus pour construire le protocole d'annotation décrit par la suite. Nous définissons la continuité référentielle comme l'ensemble des expressions permettant au lecteur d'accéder à une entité représentée dans l'univers du texte — qu'il s'agisse de syntagmes nominaux, de pronoms, de verbes où le sujet est omis (ou anaphores zéro), etc. Notre annotation vise l'ensemble de ces expressions, définies comme maillons de la chaine de continuité référentielle ou mentions. En particulier, l'annotation concerne les mentions renvoyant à un ou plusieurs personnages animés présents dans le texte, qu'ils soient imposés par la consigne ou introduits par l'élève (Barletta, 2024). Cette notion s'inspire des travaux de Givón (1983) et de Ariel (2014) sur la continuité référentielle et sur la théorie de l'accessibilité. Cette définition, bien qu'elle s'écarte de la notion stricte de coréférence, permet d'inclure dans l'annotation des mentions qui n'y répondent pas pleinement, tout en rendant possible, à terme, une comparaison avec les résultats issus du corpus RésolCo.
Le schéma d'annotation que nous avons développé peut être illustré par l'exemple de texte annoté dans la plateforme INCEpTION (cf. Figure 1). Ici nous avons annoté les expressions liées aux référents animés présents dans le texte (la sorcière, le chat et la maîtresse du chat), comme les syntagmes nominaux y compris leurs adjectifs préposés ou postposés, les anaphores zéro et les déterminants possessifs [5].

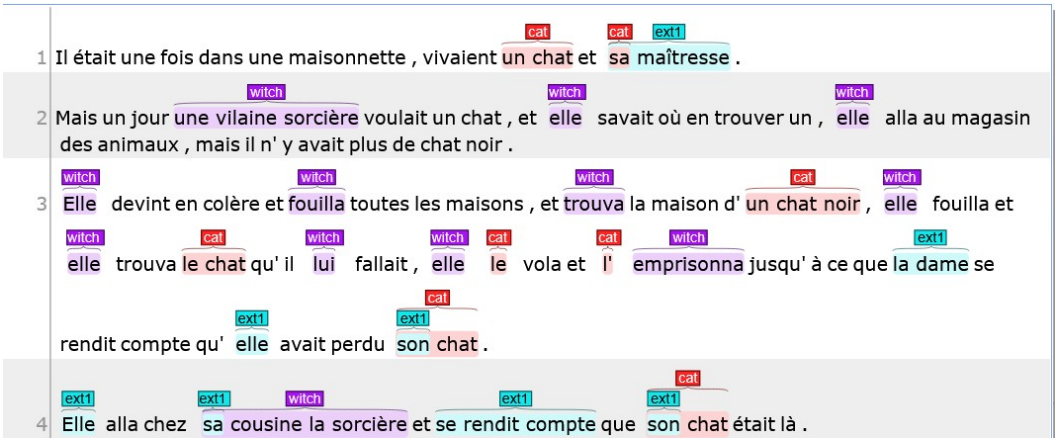


FIGURE 1 – Texte normalisé annoté de niveau CE1, élève 841.

## 2.2 Le corpus Scolinter

Alors que le manque de corpus d'écrits scolaires, c'est-à-dire de corpus d'écrits d'apprenants L1 à différentes étapes de leur scolarité (Elalouf & Boré, 2007; Wolfarth *et al.*, 2017) était autrefois un frein aux recherches véritablement empiriques dans ce domaine, ces dernières années ont été marquées par de nombreux travaux visant à combler ce manque de corpus transcrits et librement accessibles. Les différents projets réalisés en France (Doquet *et al.*, 2021, 2023) ou en Italie (Brunato & Dell'Orletta, 2015; Barbagli *et al.*, 2016; Grandi *et al.*, 2024; Revelli, 2024) témoignent de l'intérêt porté dans plusieurs pays sur l'étude de ce genre d'écriture. Ces efforts de recueil et de traitement de ces données

4. La consigne est disponible en Annexe cf. A.1.
5. Il faut remarquer qu'ici les mentions qui indiquent un référent générique comme "un chat", "des animaux" et "de chat noir" (P2) ne sont pas annotées, comme prévu par le guide d'annotation, et la locution verbale a été annoté dans son entièreté. Le guide est disponible au lien suivant : https://hal.science/view/index/docid/5082398.

ont parfois été réalisés dans le cadre de projets plus larges, comme dans le cas de corpus (Littéracie avancée, Scoledit, Ecriscol, RésolCo) constitués ou intégrés dans le projet E-Calm (Jacques & Rinck, 2017; Wolfarth, 2019; Doquet, 2020; Garcia-Debanc *et al.*, 2021), ou dans celui de recherches plus réduites voire limitées au travail d'une doctorante ou d'un doctorant (Mazziotti, 2021; Pallanti, 2021; Savin, 2024). Dans ce contexte varié est né le corpus Scolinter (Farina & Teruggi, 2021; Ponton *et al.*, 2021). L'objectif de ce corpus est de comparer le développement de l'écriture à l'école primaire en France, en Italie et en Espagne à travers le recueil d'un corpus longitudinal et comparable dans les trois langues. Ce corpus hérite des protocoles de recueil et de traitement de données déjà adoptés pour la constitution du corpus Scoledit (Wolfarth *et al.*, 2017, 2018a), en élargissant le contexte de recueil aux écoles primaires italiennes et espagnoles.
Dans les classes de CE1 à CM2, la consigne invite les élèves à sélectionner un ou plusieurs personnages parmi ceux présentés sur des vignettes (un chat, un loup, un robot et une sorcière), puis à rédiger une histoire les mettant en scène [6]. L'application de cette consigne sur ces quatre niveaux scolaires permet de comparer les productions réalisées au fil des années et entre textes provenant des différents pays (Wolfarth, 2019; Barletta, 2024). Le corpus Scolinter comprend, en l'état actuel : les 1 685 textes qui forment la partie longitudinale du corpus Scoledit, ainsi que 1 487 textes en italien et 813 textes en espagnol déjà traités. Le corpus français est complet sur les 5 niveaux de primaire, alors qu'une partie des textes des corpus italien et espagnol est encore en phase de transcription et de normalisation, notamment pour les niveaux 2, 3 et 5 en Italie (équivalents aux niveaux CE1, CE2 et CM2), et pour les niveaux 3, 4 et 5 en Espagne (équivalents aux niveaux CE2, CM1 et CM2). Il faut noter que les corpus italien et espagnol ne sont pas parfaitement longitudinaux car une partie aurait dû être recueillie dans les années 2020 et 2021, alors que les écoles n'étaient pas ou peu accessibles à cause de la pandémie de COVID.

## 2.3 Caractérisation du corpus français

Les manuscrits qui constituent le corpus français ont été recueillis entre 2014 et 2018 dans quatre académies différentes en France. Le corpus longitudinal se compose de 337 textes par niveau, rédigés par les mêmes élèves du CP au CM2. Au total, ces 1 685 textes comportent 178 183 tokens, avec une longueur moyenne de 105,75 tokens par texte. La longueur moyenne des textes semble croître proportionnellement au niveau scolaire, avec toutefois un saut plus marqué entre le CE1 et le CE2, niveaux intermédiaires et retenus pour nos annotations. Les phases de recueil et de constitution de ce corpus ont été décrites en détail dans les différents travaux de Wolfarth *et al.* (2017); Ponton *et al.* (2019); Wolfarth (2019).

| Niveau | Nb textes | Nb tokens | Longueur moyenne | Étendue |
|---|---|---|---|---|
| CP (1) | 337 | 8 931 | 26,50 | 3–71 |
| CE1 (2) | 337 | 22 241 | 66 | 6–212 |
| CE2 (3) | 337 | 39 582 | 117,45 | 6–349 |
| CM1 (4) | 337 | 48 357 | 143,49 | 25–398 |
| CM2 (5) | 337 | 59 072 | 175,29 | 19–591 |
| **CP–CM2 (1–5)** | **1 685** | **178 183** | **105,75** | **3–591** |

TABLE 1 – Caractérisation du corpus longitudinal français.

6. Des textes ont été également collecté dans des classes de CP en utilisant une consigne différente. Les deux consignes sont disponibles en Annexe (cf. A.2).

## 2.4 Caractérisation du corpus italien

Le corpus italien a été collecté entre 2018 et 2022. Le recueil a eu lieu dans différentes écoles situées dans la région de Lombardie. Nous disposons de 447 textes longitudinaux sur les niveaux 1, 2 et 5 (CP, CE1 et CM2), dont les textes de 5e sont actuellement en phase de traitement. Un recueil parallèle au principal, mené en 2018 et 2019, nous a permis d'avoir à disposition aussi des textes des classes de 2e et 3e années. Étant donné que les textes de 2e du corpus longitudinal étaient déjà transcrits et que nous disposions de ressources et de temps limités, nous avons choisi d'annoter une partie de ces textes ainsi que ceux de 3e du *minicorpus* (longitudinal en 2e et 3e, recueillis entre 2018 et 2019).
À ce jour, 1 487 textes ont ainsi été traités, dont 702 en CP, 672 en CE1, 96 en CE2 et 17 en CM2, les niveaux CE2 et CM1 étant encore en phase de traitement (les statistiques ont été incluses à titre informatif car pas encore représentatives de la totalité des niveaux). Sur un total de 1 487 textes disponibles pour 119 873 tokens, la longueur moyenne des textes en italien est alors de 80,61 tokens.

| Niveau | Année recueil | Nb textes | Nb tokens | Longueur moyenne | Étendue |
|---|---|---|---|---|---|
| CP (1) | 2018 | 702 | 27 505 | 39,18 | 2–165 |
| CE2 (2) | 2018 + 2019 | 672 | 72 182 | 107,41 | 12–552 |
| CE2 (3) | 2019 | 96 | 15 812 | 164,7 | 42–451 |
| CM2 (5) | 2022 | 17 | 4 374 | 257,29 | 134–394 |
| **CP–CM2 (1–5)** | **–** | **1 487** | **119 873** | **80,61** | **2–552** |

TABLE 2 – Caractérisation du corpus italien.

## 2.5 Campagne d'annotation en français et italien

Les résultats de la première campagne d'annotation visant à tester l'applicabilité du guide d'annotation proposé pour le français ont été présentés dans Barletta (2024). Une nouvelle campagne a été ensuite menée pour comparer les niveaux CE1 et CE2 en français et en italien. Pour ce faire, nous avons appliqué la méthodologie décrite par Barletta (2024); Barletta & Ponton (2025) pour le prétraitement des données et la sélection des textes à annoter. Nous avons ainsi sélectionné 150 textes par langue, 75 du niveau 2 (CE1) et 75 du niveau 3 (CE2). Ces textes ont préalablement été soumis à une étape de tokenisation et de parsing, réalisée à l'aide des librairies spacy-conll et Spacy 3.3 (Honnibal *et al.*, 2020) [7]. Les annotations ont été effectuées sur la plateforme INCEpTION (Klie *et al.*, 2018) entre septembre 2024 et janvier 2025 par une paire d'annotateurs experts par langue (une doctorante en Sciences du Langage et TAL qui a annoté les deux corpus et des étudiants en Master de Sciences du Langage). Le guide, conçu pour être appliqué aux deux langues, a été modifié à la suite des observations faites lors de la campagne précédente [8] Contrairement à cette campagne précédente qui s'appuyait sur le schéma d'annotation de la coréférence intégré à la plateforme, nous avons conçu ici un schéma spécifique répondant à nos besoins (calcul automatique de l'accord inter-annotateurs et outillage de l'adjudication) ainsi qu'un jeu d'étiquettes stabilisé [9]. Le corpus annoté ainsi constitué est décrit dans la Table 3.

7. Nous avons utilisé les modèles fr_core_news_lg pour le français et it_core_news_lg pour l'italien.

8. L'étape d'annotation s'est déroulée en suivant la méthodologie décrite par (Fort, 2012), en prévoyant des étapes d'amélioration du guide d'annotation sur la base des remarques effectuées par les annotateurs dans leurs journaux de bord. Ces remarques portaient principalement sur des doutes relatifs à l'annotation de l'anaphore zéro en français et la délimitation des mentions verbales dans le cas des locutions verbales, et a servi à apporter davantage d'exemples pour leur annotation.

9. Le jeu d'étiquettes est décrit en Annexe cf. A.3

| Niveau | Nb textes | | Total tokens | | Longueur moy. token | |
|---|---|---|---|---|---|---|
| | fr | it | fr | it | fr | it |
| CE1 (2) | 74 | 75 | 5 298 | 8 812 | 71,59 | 117,49 |
| CE2 (3) | 74 | 75 | 8 753 | 12 737 | 118,28 | 169,83 |
| **CE1-CE2 (2–3)** | **148** | **150** | **14 051** | **21 549** | **94,94** | **143,66** |

TABLE 3 – Caractérisation du corpus annoté par niveau scolaire - tokens par texte.

Concernant l'accord inter-annotateurs, nous nous sommes servis du calcul de l'alpha de Krippendorff (Krippendorff, 2013), disponible sur la plateforme d'annotation utilisée. Pour le français, la moyenne est de 0,86 et pour l'italien, de 0,75. Bien que l'on ait observé une légère dégradation de l'accord dans le cas de l'italien, l'accord observé reste cependant acceptable. Toutefois, dans une perspective proche, nous pensons mesurer la fiabilité de nos annotations à l'aide de la métrique *gamma* (Mathet *et al.*, 2015) car elle permet de mieux gérer les frontières des mentions. Par la suite, la phase d'adjudication des annotations a été réalisée par l'une des annotatrices et deux enseignants-chercheurs, responsables de la constitution des corpus français et italien. Deux textes peu ou pas cohérents du corpus français ont été retirés de l'annotation lors de cette étape, car ils étaient très difficiles à annoter en raison de leur faible cohérence, mais ils seront conservés pour des analyses qualitatives sur les dysfonctionnements observés.

# 3 Résultats

## 3.1 Hypothèses de recherche

Dans cette étude, nous nous attendons à trouver à la fois des similitudes et des différences significatives entre les corpus français et italien, notamment en ce qui concerne la structure des chaines référentielles et la nature morphosyntaxique des maillons utilisés dans ces chaines. Plus précisément, nous formulons les hypothèses suivantes :

1. **La taille des chaines évolue en fonction de la longueur des textes narratifs, par conséquent la densité référentielle reste stable.** La densité référentielle est définie comme le ratio entre le nombre d'expressions référentielles et le nombre total de tokens dans un texte (Boudreau & Kittredge, 2005; Schnedecker, 2017). Même si entre le CE1 et CE2 on observe une augmentation moyenne de la longueur des textes, la taille des chaines liées aux personnages de l'histoire augmente aussi entre niveaux scolaires car ceux-ci sont des référents saillants donc persistants dans le texte (Givón, 1983; Schnedecker, 2021). Cela devrait mener à une densité référentielle stable entre niveaux scolaires.
2. **La variété lexicale à l'intérieur des chaines référentielles devrait être relativement limitée.** Nous utilisons le coefficient de Perret (2000) pour mesurer la variation dans les chaînes de coréférence. Également appelé stabilité référentielle, ce coefficient est défini comme le rapport entre le nombre total d'anaphores nominales et le nombre de désignations différentes pour un même référent (Perret, 2000, p. 17). Rousier-Vercruyssen & Landragin (2019) l'interprètent comme un indice d'instabilité référentielle, calculé comme le ratio entre les dénominations uniques et le nombre total de syntagmes nominaux, multiplié par 100 pour rendre les chaînes

comparables [10]. Cet indice devrait fournir un aperçu de la variété lexicale à l'intérieur des chaines d'un texte. Les élèves ont tendance à reproduire des caractéristiques des contes de fées dans leurs écrits narratifs. Dans ce genre textuel, la faible variation lexicale faciliterait l'identification des personnages (Schnedecker, 2017; Rousier-Vercruyssen & Landragin, 2019), en réception/lecture comme en écriture. Cela devrait se traduire par une stabilité référentielle plutôt élevée [11].

3. **La distribution des types de mentions diffère entre le français et l'italien.** En français, les pronoms personnels sont fréquemment utilisés pour reprendre un référent, tandis qu'en italien, cette reprise se fait souvent par le biais du verbe, donnant lieu à une anaphore zéro. Cette différence devrait se traduire par une proportion plus élevée de mentions pronominales par rapport aux autres catégories de mentions en français, comme déjà observée en français par Federzoni *et al.* (2020), et par un usage plus important des anaphores zéro en italien, lié à la structure à sujet nul de l'italien (Ferrari, 2014).
4. **Des similitudes existent dans le démarrage des chaines référentielles entre les deux langues.** Nous prévoyons des similitudes entre le français et l'italien quant à la structure morphosyntaxique des chaines référentielles. À ce stade, les élèves devraient produire des chaines référentielles semblables à celles observées dans les écrits narratifs de scripteurs experts, notamment en respectant les critères de démarrage des chaines, à travers l'utilisation de syntagmes nominaux indéfinis. Ce type de marqueur indique, dans les deux langues, l'introduction cognitive d'un référent nouvellement mentionné dans le texte (Cornish, 1998; Salles, 2015), destiné à devenir topique par la suite (Kleiber, 1994; Cornish, 1998). Nous nous attendons donc à un pourcentage élevé de syntagmes nominaux indéfinis en position 1 des chaines dans les deux langues.

Ces hypothèses permettront de mieux comprendre les spécificités des chaines référentielles et la gestion des référents par les élèves dans les textes en français et en italien.

## 3.2 Méthodologie et résultats

Les travaux issus du projet Democrat (Landragin, 2016), qui fait partie de nos références et des références du corpus RésolCo (Garcia-Debanc *et al.*, 2021) au niveau de l'annotation et des analyses, ont conduit à définir une variété d'indicateurs permettant de caractériser les chaines de référence (Schnedecker & Landragin, 2014; Landragin *et al.*, 2024). Afin de vérifier nos hypothèses, nous avons développé différents scripts Python calculant certains de ces indicateurs sur le corpus annoté. Les résultats présentés ici ont été calculés à partir des exports au format UIMA CAS 1.0 de notre adjudication réalisée sur la plateforme INCEpTION.

Nous avons retenu les indicateurs suivants et précisé les méthodes de calcul employées : certains indices décrivent les mentions et leurs caractéristiques, tandis que d'autres portent sur les chaines et leur contenu (Oberlé *et al.*, 2018) :

— le nombre de mentions ou maillons annotés ;

10. Dans notre calcul, en suivant la définition de Perret, nous avons exclu du calcul la première mention de la chaine si elle était représentée par un syntagme nominal indéfini (par exemple, *un petit chat*). Nous n'avons cependant pas pu exclure les mentions nominales incluses dans les titres des textes dans cette version du calcul, comme prévu dans la définition originelle de l'indice.

11. Cela était le cas dans Schnedecker (2017), qui a observé que les contes de fées se caractérisent par une stabilité référentielle légèrement supérieure à celle observée dans les récits de faits divers, bien qu'aucune mesure chiffrée précise n'ait été fournie.

— le nombre de chaines annotées, comptées en distinguant trois cas : les chaines proprement dites, correspondantes à des ensembles de trois mentions ou plus reliées au même référent ; les chaines anaphoriques, lorsque deux mentions sont associées ; et les singletons, dans le cas de mentions uniques ;
— la longueur des chaines, c'est-à-dire le nombre de mentions qui indiquent le même référent ou référents dans un texte. Ici nous avons inclus dans le calcul seulement les chaines proprement dites (au-delà de trois mentions) ;
— le nombre moyen de chaines ou de référents par texte. Nous distinguons ici les chaines de personnages au singulier de celles au pluriel, et nous excluons du calcul les anaphores et les singletons ;
— la densité référentielle (définie en 3.1) ;
— l'interdistance moyenne, calculée comme la distance entre mentions de la même chaine en nombre de tokens divisé par le nombre de maillons de la chaine (Rousier-Vercruyssen & Landragin, 2019; Landragin *et al.*, 2024) ;
— l'instabilité référentielle de Rousier-Vercruyssen & Landragin (2019) (définie en 3.1) ;
— la distribution des mentions selon leur typologie et leur structure interne. En nous appuyant sur les travaux de Tutin (2002), Schnedecker (2017) et Federzoni *et al.* (2021), nous avons analysé la répartition des mentions en fonction de leur composition interne (SN défini, SN indéfini, nom propre, pronom, anaphore zéro, etc.), en exploitant le POS-tagging effectué au préalable sur nos textes (cf. 2.5). Cette classification a été effectuée à l'aide d'un algorithme à base de règles, puis affinée manuellement en cas d'erreurs d'analyse. Nous avons également examiné la distribution des mentions en position initiale et en deuxième position dans les chaines.

## 3.3 Observations

Dans notre analyse des annotations des corpus français et italien, nous avons observé les éléments suivants : dans le corpus français, nous avons identifié un total de 3 141 mentions, représentant 4 941 tokens au total, ce qui correspond à 35,2 % des tokens du corpus annoté. Dans le corpus italien, nous avons annoté 4 829 mentions, pour un total de 7 941 tokens, ce qui représente 36,9 % des tokens du corpus annoté (cf. Table 4).

| Niveau | Tot. mentions | | Tot. chaines | | Chaines (3+) | | Chaines == 2 | | Singletons (1) | |
|---|---|---|---|---|---|---|---|---|---|---|
| | fr | it | fr | it | fr | it | fr | it | fr | it |
| CE1 (2) | 1 197 | 1 992 | 222 | 275 | 150 | 194 | 31 | 28 | 41 | 53 |
| CE2 (3) | 1 944 | 2 837 | 283 | 330 | 192 | 216 | 41 | 43 | 50 | 71 |
| **CE1–CE2 (2–3)** | **3 141** | **4 829** | **505** | **605** | **342** | **410** | **72** | **71** | **91** | **124** |

TABLE 4 – Caractérisation du corpus annoté par niveau scolaire - mentions et chaines de continuité référentielle.

Les données révèlent une répartition assez similaire des mentions dans les deux corpus, bien qu'une légère différence soit observée dans la longueur moyenne des textes, dans le nombre de mentions et des chaines, le corpus italien affichant des valeurs légèrement plus élevées.
Concernant la taille des chaines, celle-ci varie peu entre les niveaux scolaires et entre les langues. En terme de densité référentielle, contrairement à nos hypothèses initiales, nous pouvons observer une baisse de la densité référentielle entre les deux niveaux. Cette baisse n'est pas statistiquement

significative selon la p-valeur obtenue à partir du test t de Student, utilisé pour comparer les niveaux CE1 et CE2 dans chaque langue [12]. En même temps, l'interdistance moyenne augmente entre les deux niveaux (cf. Table 5) de deux mentions par langue. Pour justifier ce résultat, nous postulons qu'entre deux niveaux scolaires très rapprochés, la variation en terme de densité est probablement moins marquée qu'entre niveaux scolaires plus éloignés.

| Niveau | Nb moy référents/texte | | Len moy chaines | | Len max chaines | | Densité (%) | | Interdistance | |
|---|---|---|---|---|---|---|---|---|---|---|
| | fr | it | fr | it | fr | it | fr | it | fr | it |
| CE1 (2) | 3 | 3,66 | 8,50 | 9,71 | 30 | 58 | 22,12 | 22,12 | 6,33 | 7,83 |
| CE2 (3) | 3,82 | 4,4 | 7,29 | 12,41 | 35 | 35 | 21,17 | 21,63 | 8,28 | 9,35 |
| **CE1-CE2** | **3,41** | **4,03** | **9,44** | **11,16** | **35** | **58** | **21,65** | **21,88** | **7,42** | **8,63** |

TABLE 5 – Caractérisation du corpus annoté par niveau scolaire - indicateurs des mentions et chaines de continuité référentielle.

Pour vérifier notre hypothèse par rapport à la variété lexicale à l'intérieur des chaines, en suivant la littérature existante, nous avons appliqué la définition d'instabilité référentielle décrite par Rousier-Vercruyssen & Landragin (2019). Nous avons obtenu des indices d'instabilité de 71,68% pour le français et de 72,56% pour l'italien. Ces valeurs très élevées d'instabilité contrastent avec notre hypothèse initiale ainsi qu'avec les observations faites lors de l'annotation du corpus [13]. Cependant, cet indice ne semble pas rendre compte de la relative stabilité des chaines dans nos textes et le calcul de l'indice néglige la présence de noms propres et la similarité entre syntagmes nominaux qui partagent la même tête lexicale. Cet indice reste donc peu adapté à rendre compte de la réelle variété de syntagmes contenus dans des chaines (Schnedecker, 2021). Cela est évident dans le cas de notre corpus où la nature des maillons que l'on retrouve majoritairement dans les textes ne rend pas toujours possible l'application de cet indice [14]. Ainsi, en français, 1 116 mentions ont été prises en compte pour ce calcul sur les 3 141 annotées, et l'instabilité a pu être calculée sur 140 textes au total sur les 148 annotés, tandis qu'en italien l'instabilité a pu être calculée sur un total de 1 673 mentions dans 146 textes sur les 150 textes annotés.

12. Dans le deux cas, $p > 0,05$.

13. Ces valeurs s'éloignent aussi du 40% d'instabilité constaté par Rousier-Vercruyssen & Landragin (2019) dans une partie des textes narratifs du corpus Democrat

14. Nous retrouvons dans notre échantillon des chaines composées de SN indéfinis suivis par anaphores zéro et/ou des pronoms. Cela est fréquent dans le cas du corpus italien.

| Type ER | Mentions | | Position 1 | | Position 2 | |
|---|---|---|---|---|---|---|
| | fr | it | fr | it | fr | it |
| Pronom | **1307** | 945 | 22 | 11 | **169** | **121** |
| SN défini | 859 | 1250 | 78 | 118 | 75 | 76 |
| SN indéfini | 306 | 397 | **199** | **209** | 36 | 89 |
| SN possessif | 313 | 308 | 21 | 23 | 36 | 24 |
| SN démonstratif | 20 | 35 | 3 | 0 | 3 | 5 |
| SN sans déterminant | 42 | 56 | 0 | 1 | 4 | 5 |
| Nom propre | 123 | 323 | 5 | 11 | 6 | 11 |
| Anaphore zéro | 132 | **1473** | 0 | 28 | 10 | 75 |
| Autre | 39 | 42 | 14 | 9 | 3 | 4 |
| **Tot.** | **3141** | **4829** | **342** | **410** | **342** | **410** |

TABLE 6 – Distribution des mentions par typologie et par position dans les chaines de continuité référentielle.

En ce qui concerne la distribution des mentions par typologie, dans le cas du français la catégorie majoritaire est celle des pronoms, suivie des syntagmes nominaux définis, tout comme dans les observations précédentes réalisées par Federzoni *et al.* (2020). En italien, les anaphores zéro constituent la catégorie la plus fréquente, suivies également des syntagmes nominaux définis. Ces résultats confirment notre hypothèse d'une similitude dans la répartition des mentions avec les corpus annotés de scripteurs experts pour le français, ainsi que l'hypothèse d'une prédominance des anaphores zéro au sein des chaines en italien. En analysant les mentions en position 1 et 2 dans les chaines, nous retrouvons une similarité évidente entre les deux langues : effectivement, dans les deux corpus la catégorie la plus fréquente en position 1 est celle du SN indéfini (ex. "un chat"), alors que la plus fréquente en position 2 est celle du pronom, qu'il soit personnel ou relatif. Ces résultats semblent corroborer certaines observations empiriques de Salles (2015), qui mettent en évidence la présence dominante des pronoms en position 2 dans les chaînes narratives.

# 4 Analyse qualitative

Une analyse qualitative menée en parallèle sur les deux corpus nous a permis d'identifier des types de phénomènes, définis comme anomalies par Roubaud & Garcia-Debanc (2014) et qui sont présents dans notre corpus annoté. Entre autres, le double marquage des référents (ex. lui le lapin, cf. Annexe A.5.1) ou encore l'absence d'un référent explicite dans le texte (cf. Annexe A.5.2).

Dans les deux corpus nous avons aussi pu identifier deux stratégies de constructions de chaines qui semblent fréquemment utilisés : une qu'on appelle à prévalence pronominale et l'autre à prévalence nominale. Dans les deux langues, nous avons observé des textes où la gestion du ou des référents est effectuée de manière prévalente avec une des deux stratégies, soit à travers une utilisation des syntagmes nominaux quasiment tout au long de la chaine, soit par une première mention nominale suivie quasi seulement de pronoms personnels, ce qui se traduit souvent en italien par l'utilisation des verbes à sujet omis (pour des exemples des deux stratégies, voir Annexe A.5.3). Nous nous proposons par la suite d'approfondir l'analyse de ces stratégies dans les différents niveaux scolaires et entre langues.

# 5 Conclusion et perspectives

Cette campagne d'annotation, menée sur 148 textes français et 150 textes italiens d'élèves de niveaux CE1 et CE2, nous a permis de réfuter et de confirmer certaines hypothèses quant au fonctionnement de la continuité référentielle dans des productions écrites à l'école primaire. Nous avons pu confirmer que la densité référentielle reste relativement stable entre niveaux scolaires pour chaque langue. De plus, cette densité est comparable entre les deux langues. Cependant, il serait nécessaire de mener une analyse sur des textes de niveau CM1 et CM2 dans les deux langues afin de déterminer si ces résultats restent les mêmes pour des textes de longueur significativement supérieure. Concernant la variété lexicale à l'intérieur des chaines, nous obtenons des indices d'instabilité très hauts qui pourraient indiquer une variété très grande dans les syntagmes nominaux utilisés. Cependant, l'indice proposé prend en considération seulement une partie restreinte du matériel qui compose les chaines et nous semble décrire de manière réductive la réelle variation rencontrée. De nouveaux calculs de cet indice pourraient intégrer la prise en compte de la tête du syntagme nominal dans la variation, et également, l'inclusion des noms propres dans les mentions prises en compte. Enfin, la distribution des types de mentions (y compris les mentions en dehors des chaines) confirme celle attendue pour les deux langues. Nous retrouvons également la distribution attendue de SN indéfinis en démarrage de chaine dans ce genre textuel, ainsi qu'un pourcentage élevé de pronoms en position 2 dans les chaines.

Outre la reprise des éléments de calculs mentionnés, pour la suite de ce travail, nous prévoyons trois grandes directions. La première concerne une comparaison de ces résultats avec ceux du niveau CM2 pour mieux mesurer l'évolution des élèves dans la gestion de la continuité référentielle, ainsi qu'une comparaison avec les résultats obtenus sur le corpus RésolCo. La seconde prévoit de compléter les analyses quantitatives par une approche qualitative plus approfondie. Enfin, le corpus Scolinter étant un corpus trilingue, une phase d'annotation et d'exploitation du corpus espagnol devrait démarrer prochainement. Le corpus annoté que nous avons présenté ici sera déposé sur la plateforme Ortholang courant 2025.

# Remerciements

Nous souhaitons remercier l'équipe responsable du recueil du corpus, dont Catherine Brissaud, Lilia Teruggi, Elisa Farina et Rafaela Gutiérrez Cáceres, ainsi que les annotateurs qui ont rendu possible ce travail, Paola Zancanaro et Baptiste Delpech.

# Références

ARIEL M. (2014). *Accessing noun-phrase antecedents*. Routledge.

BARBAGLI A., LUCISANO P., DELL'ORLETTA F., MONTEMAGNI S. & VENTURI G. (2016). CItA : an L1 Italian Learners Corpus to Study the Development of Writing Competence. In N. CALZOLARI, K. CHOUKRI, T. DECLERCK, S. GOGGI, M. GROBELNIK, B. MAEGAARD, J. MARIANI, H. MAZO, A. MORENO, J. ODIJK & S. PIPERIDIS, Éds., *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, p. 88–95, Portorož, Slovenia : European Language Resources Association (ELRA).

BARLETTA M. (2024). Annotation de la continuité référentielle dans un corpus scolaire – premiers résultats. In M. BALAGUER, N. BENDAHMAN, L.-M. HO-DAC, J. MAUCLAIR, J. G. MORENO & J. PINQUIER, Éds., *35èmes Journées d'Études sur la Parole (JEP) 31ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN) 26ème Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RECITAL)*, p. 28–41, Toulouse, France : ATALA & AFPC. HAL : hal-04622985.

BARLETTA M. & PONTON C. (2025). La question de la normalisation des écrits scolaires pour leur traitement automatique. Le cas de l'omission de mots. *Corpus*, (26). DOI : 10.4000/1364v, HAL : hal-04916955.

BOUDREAU S. & KITTREDGE R. (2005). Résolution des anaphores et détermination des chaines de coréférences : Différences entre variétés de textes. *Traitement Automatique des Langues*, **46**(1), 41–70.

BRUNATO D. & DELL'ORLETTA F. (2015). ISACCO : a corpus for investigating spoken and written language evelopment in Italian school–age children. In C. BOSCO, S. TONELLI & F. M. ZANZOTTO, Éds., *Proceedings of the Second Italian Conference on Computational Linguistics CLiC-it 2015 : 3-4 December 2015, Trento*, Collana dell'Associazione Italiana di Linguistica Computazionale, p. 62–66. Torino : Accademia University Press. DOI : 10.4000/books.aaccademia.1324.

CORNISH F. (1998). Les 'chaines topicales' : Leur rôle dans la gestion et la structuration du discours. *Cahiers de Grammaire*, **23**, 19–40. HAL : hal-03773261v1.

DOQUET C. (2020). Analyser linguistiquement l'écriture à l'école : EcriScol, un corpus génétique. In *CLUB Working Papers in Linguistics*, volume 4, p. 127–140. Bologna : CLUB - Circolo Linguistico dell'Università di Bologna. HAL : hal-02883152.

DOQUET C., HO-DAC L.-M. & PONTON C. (2023). Un corpus de référence pour l'écriture de l'école à l'université : la ressource É-Calm. In *Jounées Linguistique de Corpus*, Grenoble, France. HAL : halshs-04212830.

DOQUET C., REVELLI L. & MOYSAN A. (2021). Écriture et forme scolaire : spécificités de transcription et de traitement. *Langue francaise*, **211**(3), 21–36. HAL : halshs-03664021.

ELALOUF M.-L. & BORÉ C. (2007). Construction et exploitation de corpus d'écrits scolaires. *Revue française de linguistique appliquée*, (1), 53–70. Publications linguistiques, Section : Sciences de l'éducation, DOI : 10.3917/rfla.121.0053, HAL : halshs-00156962.

FARINA E. & TERUGGI L. (2021). Un corpus di testi trilingue per promuovere la riflessione sulla pratica didattica. In *Ricerca e didattica per promuovere intelligenza comprensione e partecipazione*, volume I, panel 1-2-3, p. 263–279, Lecce/Brescia : Pensa MultiMedia.

FEDERZONI S., HO-DAC L.-M. & FABRE C. (2021). Coreference Chains Categorization by Sequence Clustering. In C. BRAUD, C. HARDMEIER, J. J. LI, A. LOUIS, M. STRUBE & A. ZELDES, Éds., *Proceedings of the 2nd Workshop on Computational Approaches to Discourse*, p. 52–57, Punta Cana, Dominican Republic and Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.codi-main.5, HAL : hal-03513356.

FEDERZONI S., HO-DAC L.-M. & REBEYROLLE J. (2020). Les chaines topicales dans la ressource ANNODIS. *SHS Web of Conferences*, **78**. DOI : 10.1051/shsconf/20207811005, HAL : hal-02890989.

FERRARI A. (2014). *Linguistica del testo. Principi, fenomeni, strutture*. Roma : Carocci.

FORT K. (2012). *Les ressources annotées, un enjeu pour l'analyse de contenu : vers une méthodologie de l'annotation manuelle de corpus*. Thèse de doctorat, Université Paris-Nord - Paris 13. HAL : tel-00797760.

GARCIA-DEBANC C., HO-DAC L.-M., FEDERZONI S., BRAS M. & REBEYROLLE J. (2019). ResolCo un corpus de manuscrits d'élèves et d'étudiants pour l'étude de la cohérence. HAL : hal-02877122.

GARCIA-DEBANC C., REBEYROLLE J. & HO-DAC L.-M. (2021). La continuité référentielle dans le corpus RÉSOLCO : méthode d'annotation et premières analyses. *Langue francaise*, **211**(3), 99–114. DOI : 10.3917/lf.211.0099, HAL : hal-03559961.

GIVÓN T. (1983). *Topic continuity in discourse : a quantitative cross-language study*. Amsterdam ; Philadelphia : J. Benjamins Pub. Co.

GRANDI N., BALLARÈ S., MARTARI Y. & MIOLA E. (2024). Univers-ITA. Descrizione e primi risultati di uno studio dell'italiano scritto di studenti universitari. *ITALIANO A STRANIERI*, **35**, 19–26.

HO-DAC L.-M., FEDERZONI S., BRAS M., REBEYROLLE J. & GARCIA-DEBANC C. (2019). ResolCo un corpus de manuscrits d'élèves et d'étudiants pour l'étude de la cohérence. In *10èmes Journées Internationale de la Linguistique de Corpus*, Grenoble, France.

HONNIBAL M., MONTANI I., VAN LANDEGHEM S. & BOYD A. (2020). spaCy : Industrial-strength Natural Language Processing in Python. DOI : 10.5281/zenodo.1212303.

JACQUES M.-P. & RINCK F. (2017). Un « corpus de littéracie avancée : résultat et point de départ. *Corpus*, (16). DOI : 10.4000/corpus.2806.

KLEIBER G. (1994). *Anaphores et pronoms*. Louvain-la-Neuve : Duculot.

KLIE J.-C., BUGERT M., BOULLOSA B., CASTILHO R. E. D. & GUREVYCH I. (2018). The INCEpTION Platform : Machine-Assisted and Knowledge-Oriented Interactive Annotation. In *Proceedings of the 27th International Conference on Computational Linguistics : System Demonstrations*, p. 5–9 : Association for Computational Linguistics. Santa Fe, USA.

KRIPPENDORFF K. (2013). *Content Analysis : An Introduction to Its Methodology*. Thousand Oaks, Calif. : Sage.

LANDRAGIN F. (2016). Description, modélisation et détection automatique des chaines de référence (DEMOCRAT). *Bulletin de l'Association Française pour l'Intelligence Artificielle*, (92), 11. HAL : hal-01347949.

LANDRAGIN F., GLIKMAN J., SCHNEDECKER C. & TODIRASCU A. (2024). Chaines de référence dans le corpus Democrat : une analyse en diachronie longue. *Corpus*, (25). Bases, corpus et langage - UMR 6039, DOI : 10.4000/corpus.8581.

MATHET Y., WIDLÖCHER A. & MÉTIVIER J.-P. (2015). The Unified and Holistic Method Gamma ($\gamma$) for Inter-Annotator Agreement Measure and Alignment. *Computational Linguistics*, **41**(3), 437–479. DOI : 10.1162/COLI_a_00227.

MAZZIOTTI S. (2021). *L'incidence du système linguistique : étude des postures de correction des enseignants et des modalités de réécriture à l'école primaire en France et en Italie*. Thèse de doctorat, Université de la Sorbonne Nouvelle - Paris III ; Università degli Studi di Bologna - Italie. HAL : tel-03737461.

MUZERELLE J., LEFEUVRE A., ANTOINE J.-Y., SCHANG E., MAUREL D., VILLANEAU J. & ESHKOL I. (2013). ANCOR, premier corpus de français parlé d'envergure annoté en coréférence et distribué librement. In ATALA, Éd., *TALN'2013, 20e conférence sur le Traitement Automatique des Langues Naturelles*, p. 555–563, Les Sable d'Olonne, France. HAL : hal-01016562.

OBERLÉ B., SCHNEDECKER C., BAUMER E., CAPIN D., GLIKMAN J., GUO C., REVOL T., TODIRASCU A. & TUSHKOVA J. (2018). Les chaines de référence dans les textes encyclopédiques du 12e au 21e siècle : étude longitudinale. *Travaux de linguistique*, **77**(2), 67–141. DOI : 10.3917/tl.077.0067.

PALLANTI L. (2021). *Travailler les compétences rédactionnelles à l'ÉNEPS. Conception et mise en oeuvre d'un système didactique expérimental*. Thèse de doctorat, Université Grenoble Alpes.

PERRET M. (2000). Quelques remarques sur l'anaphore nominale aux xiv$^{e}$ siècle et xv$^{e}$ siècle. *L'information grammaticale*, **87**(1), 17–23. DOI : 10.3406/igram.2000.2740.

POESIO, MASSIMO, ARTSTEIN, RON, URYUPINA, OLGA, RODRIGUEZ, KEPA, DELOGU, FRANCESCA, BRISTOT, ANTONELLA & HITZEMAN, JANET (2013). The ARRAU Corpus of Anaphoric Information. Artwork Size : 184570 KB Pages : 184570 KB, DOI : 10.35111/Y3MR-HE10.

PONTON C., GUTIÉRREZ-CACERES R., TERUGGI L., FARINA E., BRISSAUD C. & WOLFARTH C. (2021). Scolinter : un corpus trilingue. L'exemple de la segmentation en mots. *Langue française*, **211**(3), 37–50. DOI : 10.3917/lf.211.0037, HAL : halshs-00168567.

PONTON C., WOLFARTH C. & BRISSAUD C. (2019). Premières explorations textométriques d'un corpus scolaire longitudinal (CP-CM1). In *Journées Linguistique de Corpus (JLC2019)*, Grenoble, France. HAL : halshs-03205233.

PÉRY-WOODLEY M.-P., AFANTENOS S. D., HO-DAC L.-M. & ASHER N. (2011). La ressource ANNODIS, un corpus enrichi d'annotations discursives. *Révue TAL - Traitement Automatique des Langues*, **52**(3), 71–101. France, HAL : halshs-00935201.

RECASENS M. & MARTÍ M. A. (2010). AnCora-CO : Coreferentially annotated corpora for Spanish and Catalan. *Language Resources and Evaluation*, **44**(4), 315–345. DOI : 10.1007/s10579-009-9108-x.

REVELLI L. (2024). Tra norma, usi e sentire comune : pseudoregole dell'italiano attraverso il CoDiSSc. In *CLUB Working Papers in Linguistics*, volume 8, p. 47–61. Bologna : CLUB - Circolo Linguistico dell'Università di Bologna. DOI : 10.6092/unibo/amsacta/8065.

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. Peter Lang. DOI : 10.3726/978-3-0352-6464-7.

ROUSIER-VERCRUYSSEN L. & LANDRAGIN F. (2019). Interdistance et instabilité au sein des chaines de référence : indices textuels ? *Discours. Revue de linguistique, psycholinguistique et informatique*, (25). Presses universitaires de Caen, DOI : 10.4000/discours.10522.

SALLES M. (2015). Chaines de référence : la deuxième mention. L'exemple des entités inanimées dans les narrations littéraires. *Travaux de linguistique*, **71**(2), 111–133. DOI : 10.3917/tl.071.0111.

SAVIN H. (2024). *Vous écrivez ? Anaphorez ! ou les anaphoriques dans les productions écrites de lycéens-nes*. Thèse de doctorat, Université Grenoble Alpes. HAL : tel-04888254.

SCHNEDECKER C. (2017). Les chaines de référence : une configuration d'indices pour distinguer et identifier les genres textuels. *Langue française*, **195**(3), 53–72. DOI : 10.3917/lf.195.0053, HAL : hal-01591017.

SCHNEDECKER C. (2021). *Les chaines de référence en français*. Paris : OPHRYS.

SCHNEDECKER C. & LANDRAGIN F. (2014). Les chaines de référence : présentation. *Langages*, **195**(3), 3–22. DOI : 10.3917/lang.195.0003.

TUTIN A. (2002). A corpus-based study of pronominal anaphoric expressions in French. In A. BRANCO, T. MCENERY & R. MITKOV, Éds., *Proceedings of the 4th Discourse Anaphora and Anaphora Resolution Colloquium (DAARC 2002)*, Lisbonne : Edições Colibri.

WEISCHEDEL, RALPH, PALMER, MARTHA, MARCUS, MITCHELL, HOVY, EDUARD, PRADHAN, SAMEER, RAMSHAW, LANCE, XUE, NIANWEN, TAYLOR, ANN, KAUFMAN, JEFF, FRANCHINI, MICHELLE, EL-BACHOUTI, MOHAMMED, BELVIN, ROBERT & HOUSTON, ANN (2013). OntoNotes Release 5.0. DOI : 10.35111/XMHB-2B84.

WOLFARTH C. (2019). *Apport du TAL à l'exploitation linguistique d'un corpus scolaire longitudinal*. Thèse de doctorat, Université Grenoble Alpes.

WOLFARTH C., BRISSAUD C. & PONTON C. (2018a). Transcrire et normer un corpus scolaire : pour quelles analyses ? In C. BRISSAUD, M. DREYFUS & B. KERVYN, Éds., *Repenser l'écriture et son évaluation au primaire et au secondaire*, volume 36 de collection Diptyque, p. 121–145. Presses universitaires de Namur. HAL : hal-01883221.

WOLFARTH C., PONTON C. & BRISSAUD C. (2018b). Gestion de la morphographie verbale en production d'écrits : que peut nous apprendre un corpus longitudinal ? *Repères. Recherches en didactique du français langue maternelle*, (57), 209–226. Number : 57 Publisher : Éditions de l'École normale supérieure de Lyon, DOI : 10.4000/reperes.1576.

WOLFARTH C., PONTON C. & TOTEREAU C. (2017). Apports du TAL à la constitution et à l'exploitation d'un corpus scolaire au travers du développement d'un outil d'annotation orthographique. *Corpus*, **Spécificités et contraintes des grands corpus de textes scolaires : problèmes de transcription, d'annotation et de traitement**(16), 185–214. Bases, corpus et langage - UMR 6039, DOI : 10.4000/corpus.2796, HAL : hal-01878701.

# A Annexes

## A.1 Consigne RésolCo

Raconte une histoire dans laquelle tu insèreras séparément et dans l'ordre donné les trois phrases suivantes :
P1 - Elle habitait dans cette maison depuis longtemps.
P2 - Il se retourna en entendant ce grand bruit.
P3 - Depuis cette aventure, les enfants ne sortent plus la nuit. (découpez et collez les bandelettes dans votre texte) (Ho-Dac *et al.*, 2019, p. 1)

## A.2 Consignes Scolinter

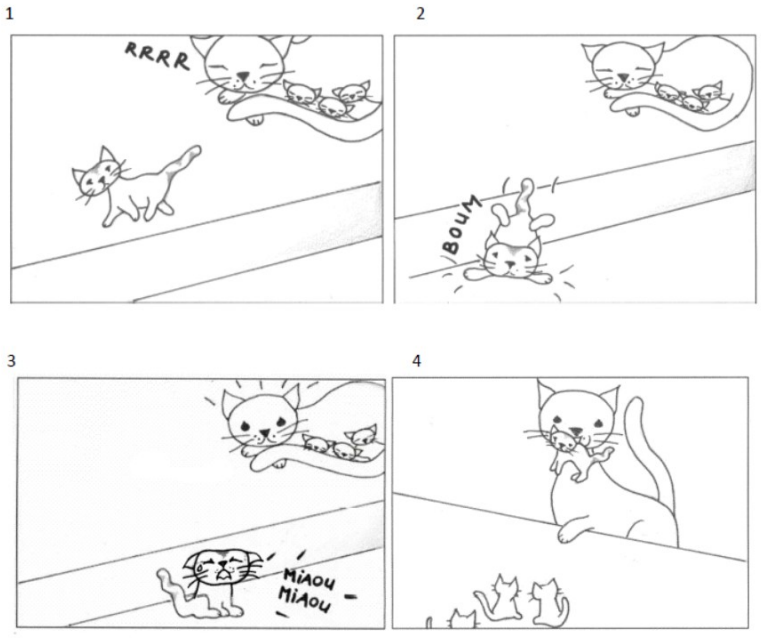


FIGURE 2 – Images présentées aux élèves lors de la production écrite en CP

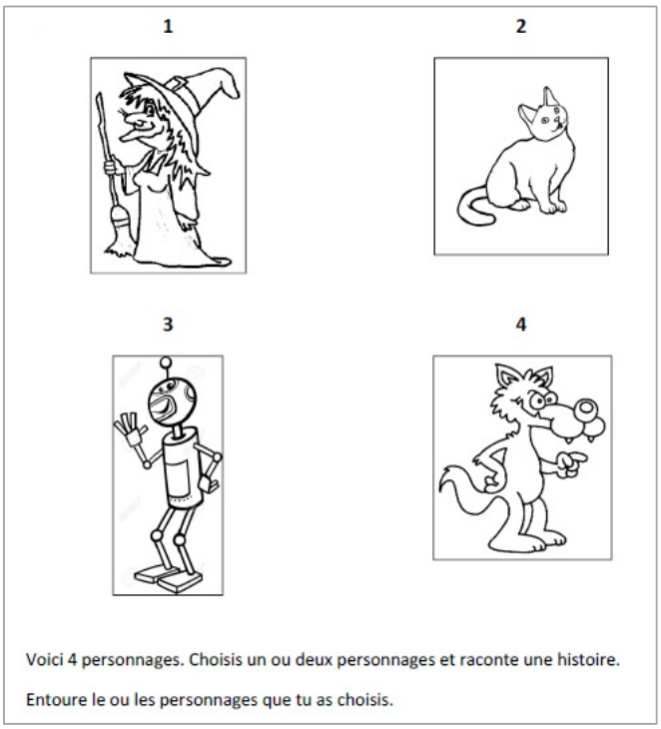


FIGURE 3 – Images présentées aux élèves lors de la production écrite en CE1, CE2, CM1, CM2

## A.3 Jeu d'étiquettes

Chaque référent issu de la consigne est identifié par une étiquette univoque : **cat** pour le chat, **witch** pour la sorcière, **wolf** pour le loup et **robot**. Les personnages externes sont indiqués par le biais de l'étiquette **extN**, où N correspond à l'ordre de parution dans le texte des personnages externes (**ext1**, **ext2**, **ext3**...). Si plusieurs référents issus de la consigne sont présents, ils sont numérotés à partir du deuxième référent évoqué (ex. si deux chats sont présents dans le texte, ils seront indiqués par les étiquettes **cat** et **cat1**).

## A.4 Distribution des mentions par type - comparaison entre français et italien.

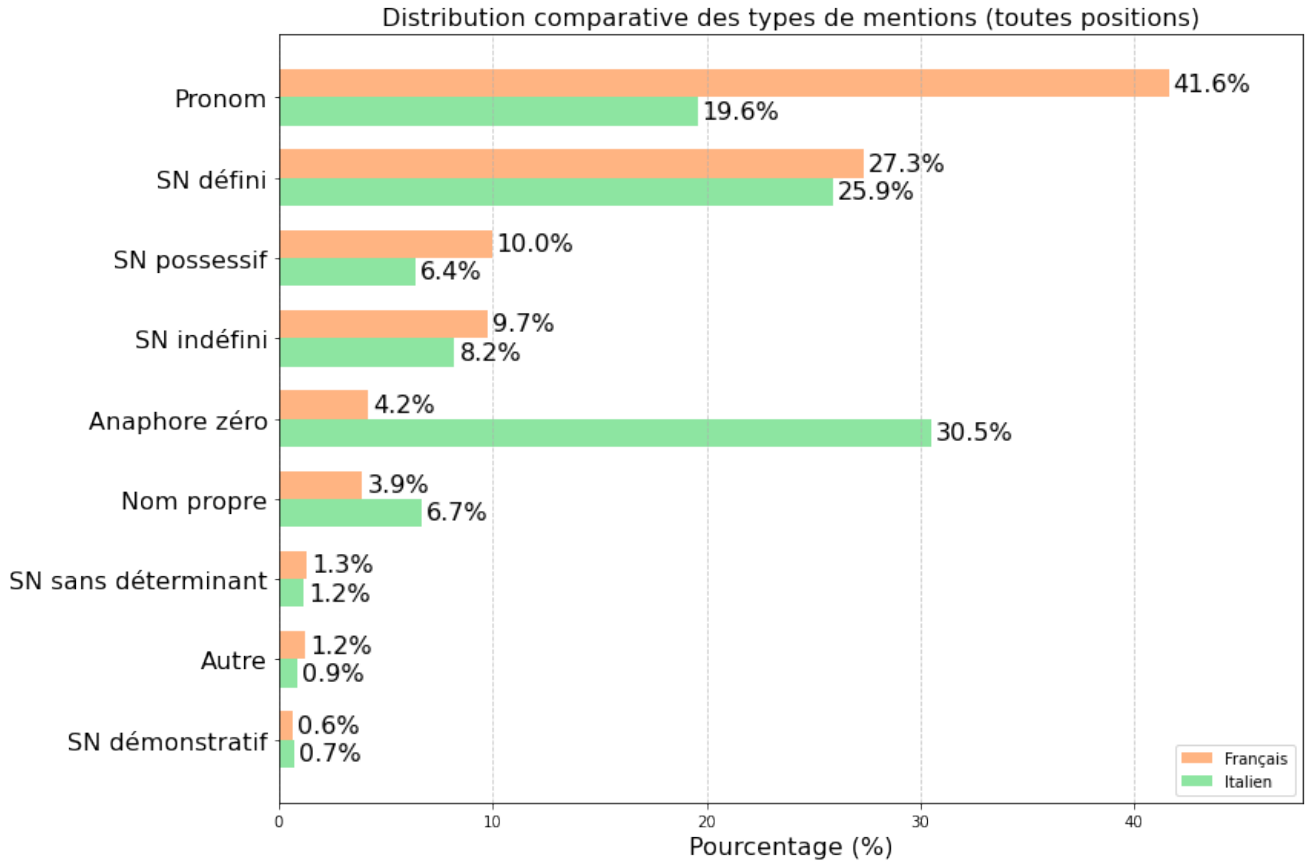


FIGURE 4 – Comparaison entre type de mentions dans le corpus annoté français et italien - toutes mentions confondues

## A.5 Analyse qualitative

### A.5.1 Double marquage du référent

Dans ce texte le référent ext1 a été indiqué par deux mentions d'affilé "lui le lapin" dans la phrase 5.

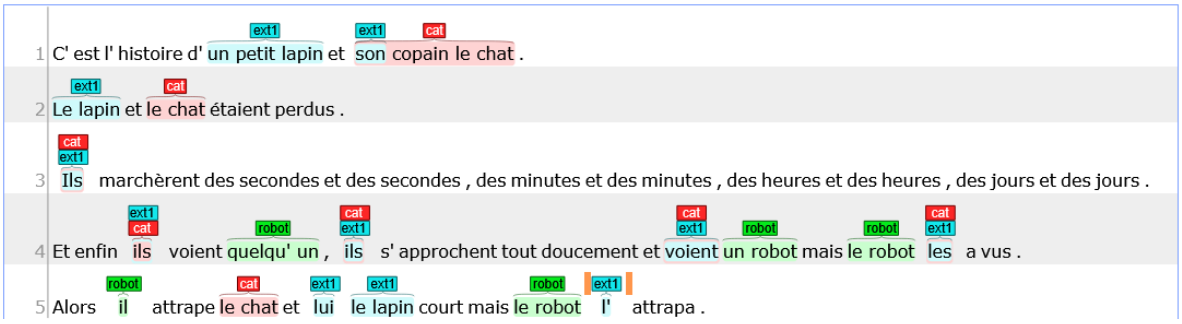


FIGURE 5 – Texte de niveau CE1, élève 2977

Dans ce texte, le référent chat a été marqué à travers un pronom suivi d'un syntagme nominal (P 3).

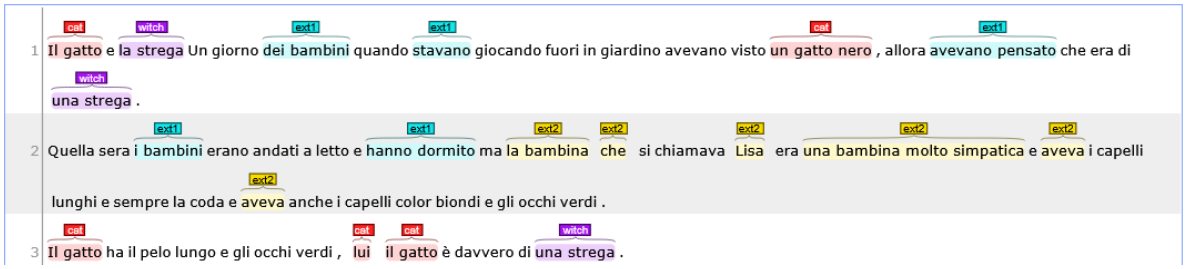


FIGURE 6 – Extrait du texte italien de niveau CE2, élève 1676

### A.5.2 Absence d'un référent explicite

Dans ce texte, le référent ext1 a été annoté bien qu'il ne semble pas être rattaché à un référent explicité à travers des mentions "pleines" dans le texte.

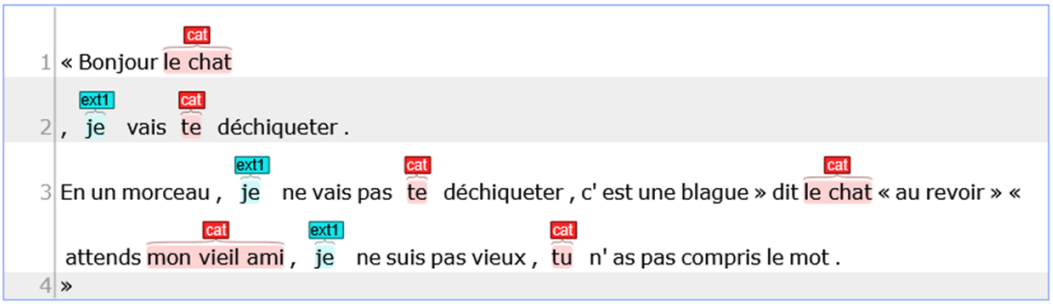


FIGURE 7 – Texte de niveau CE1, élève 2977

### A.5.3 Gestion du référent à prévalence pronominale

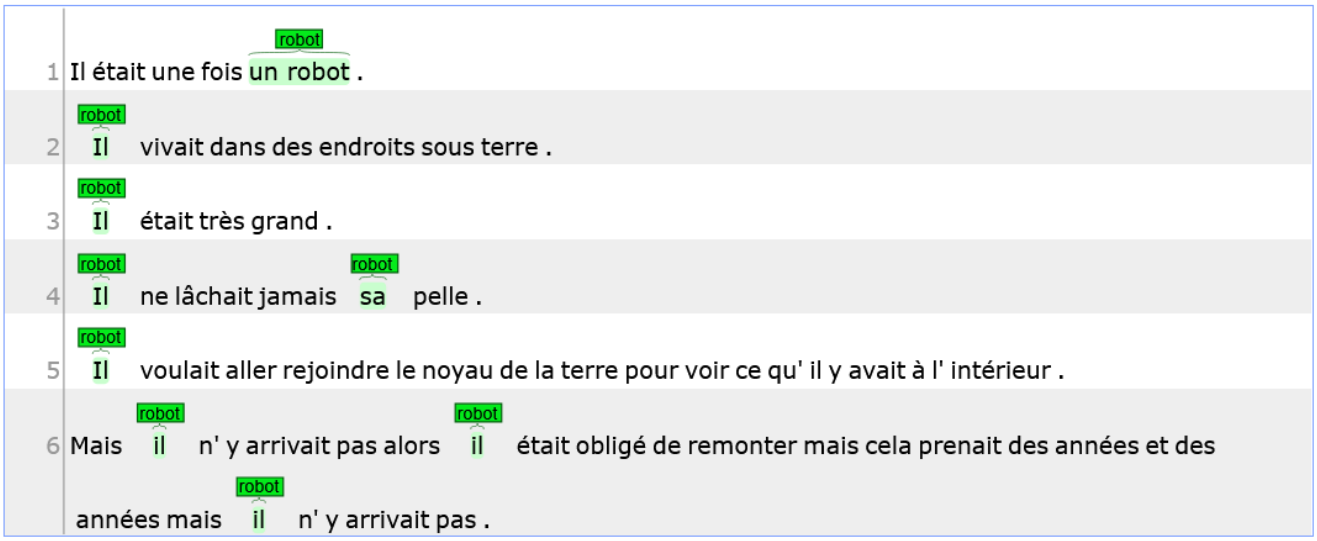


FIGURE 8 – Texte de niveau CE1, élève 3045

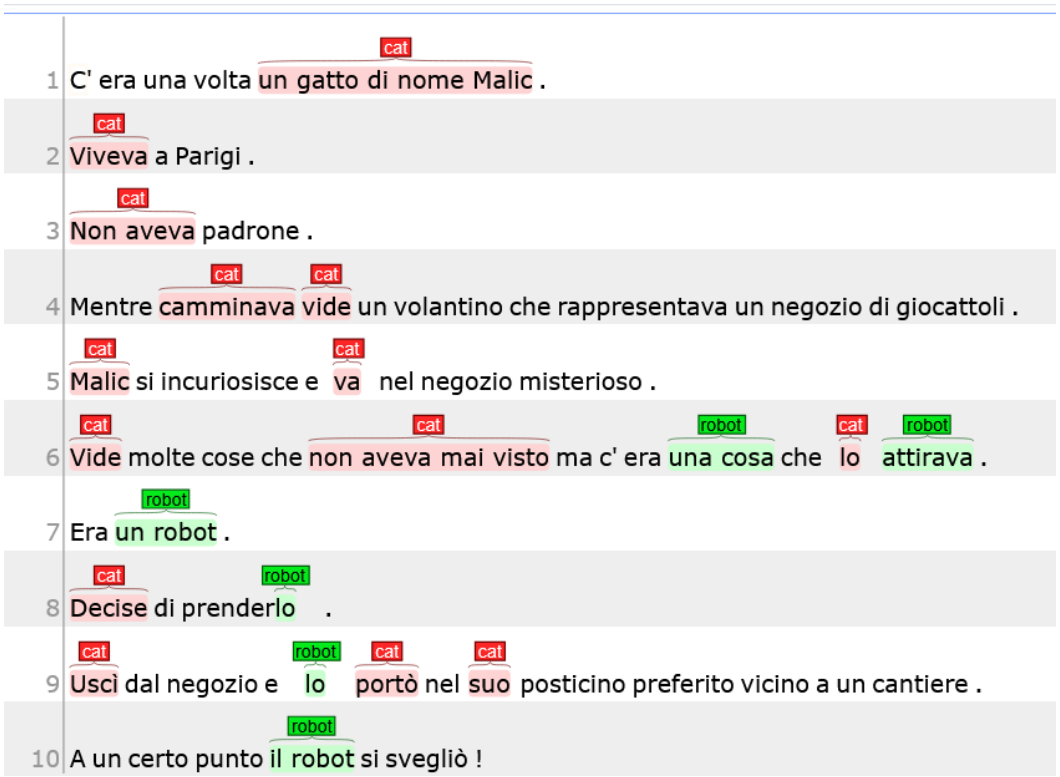


FIGURE 9 – Extrait du texte italien de niveau CE1, élève 386

### A.5.4 Gestion du référent à prévalence nominale

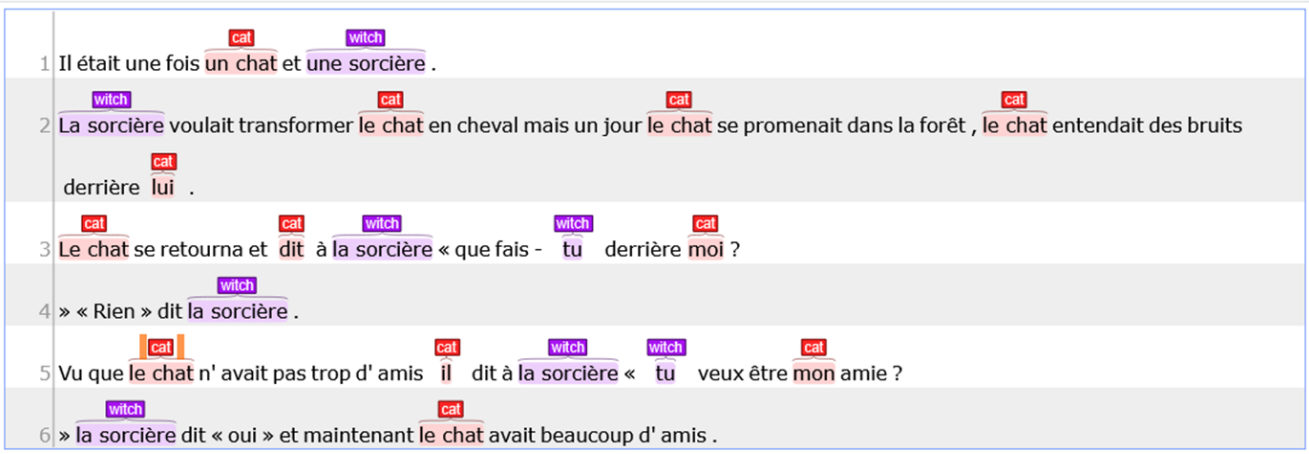


FIGURE 10 – Texte de niveau CE1, élève 1830

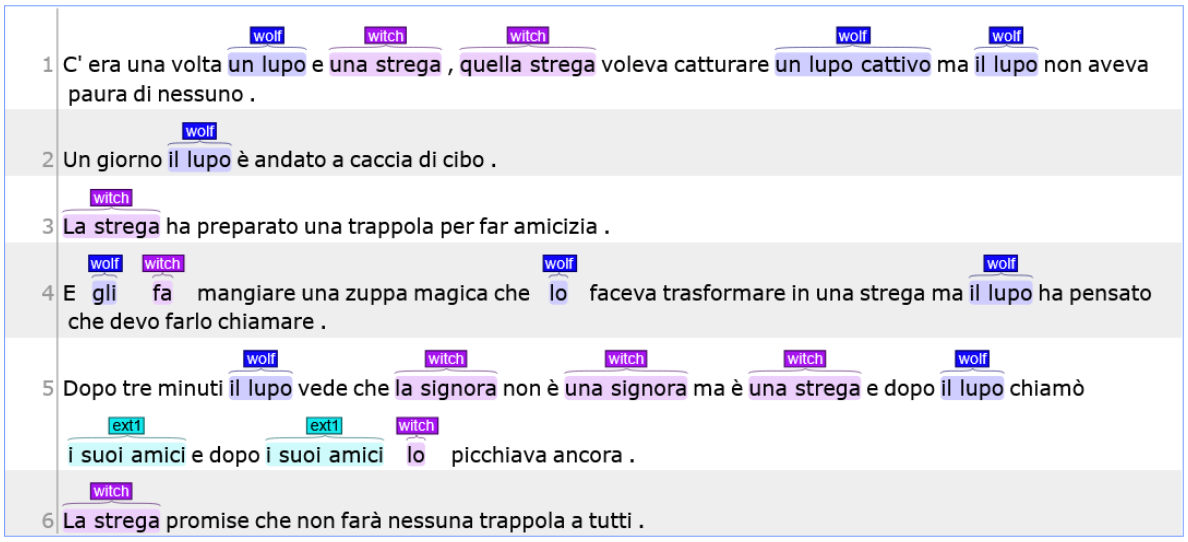


FIGURE 11 – Texte italien de niveau CE1, élève 1470