# Projet de recherche

Master 1 LITL
2024/2025

# Caractérisation des contextes de maintien d'un personnage dans une chaîne de référence.

Geley Mathilde

Sous la direction de Mme Lydia-Mai Ho-Dac et Mme Josette Rebeyrolle.

# Table des matières

**Remerciements**

Je souhaite commencer par remercier mes directrices de mémoire, Lydia-Mai Ho-Dac et Josette Rebeyrolle, pour leurs conseils ainsi que pour leur disponibilité lors de nos échanges et rencontres.

Je souhaite également remercier les professeurs du master LITL pour la qualité de leurs enseignements et leur accompagnement tout au long de cette année scolaire.

Je remercie chaleureusement la promotion LITL 2024-2025 pour la richesse de nos échanges et le soutien que j'ai pu y trouver.

Enfin, j'adresse toute ma gratitude à mes proches et ma famille pour leur compréhension et leurs encouragements inconditionnels. Je remercie en particulier Betty, Georges, Camille et Claire, sans qui cette année n'aurait pas été la même.

## *Introduction*

Le travail de rédaction d'une histoire est un exercice complexe qui mobilise des compétences linguistiques précises. Il implique de présenter des personnages, de les ancrer dans un contexte narratif et de les faire interagir de manière cohérente tout au long du récit. La continuité référentielle constitue un enjeu fondamental pour assurer la bonne compréhension d'un texte. D'un point de vue linguistique, de nombreuses questions se posent quant à la manière dont un référent est introduit, maintenu ou réintroduit au fil du texte. En effet, le scripteur ou le locuteur doit choisir parmi divers moyens linguistiques permettant d'assurer cette continuité référentielle. Les référents peuvent être rappelés sous une variété de formes et on peut notamment distinguer les syntagmes nominaux ainsi que les pronoms. Dans l'exemple :

1) Il était une fois une femme qui vivait dans une maison étrange. Elle habitait dans cette maison depuis longtemps. [EC-CM2-2016-MQRVX-D1-R4-V1]

Nous pouvons identifier le syntagme nominal u*ne femme*, le pronom relatif *qui* et le pronom personnel *Elle.*

De nombreux travaux se sont penchés sur les différentes formes que peuvent prendre les expressions référentielles (Charolles, 2002), ainsi que sur les contraintes qui orientent leur choix en contexte (Ariel, 1990). Ce mémoire a pour ambition de s'inscrire dans ces questionnements. Il a pour objectif d'analyser les contextes de maintien de référents humains dans des textes narratifs d'élèves de cycle 3, de cycle 4 ainsi que d'étudiant.es de Master 2 afin de comparer l'usage des pronoms et des formes nominales dans la construction des chaînes de référence.

Plus précisément, notre principal objectif de recherche est de caractériser le lien entre la distance qui sépare deux maillons d'une chaîne référentielle et la forme de l'expression référentielle qui est employée. Nous souhaitons analyser si la forme d'un maillon varie selon la distance qui le sépare du maillon précédent. Pour ce travail, nous avons choisi d'observer en particulier les pronoms personnels et relatifs ainsi que les syntagmes nominaux.

Pour affiner notre travail, deux variables entrent en jeu et viennent nourrir nos questionnements :

• Le niveau scolaire : nous allons nous demander si la variation de la forme selon la distance évolue avec le niveau scolaire d'un élève.

• Le référent : nous souhaitons observer si le référent a une influence dans la variation de la forme selon la distance. Pour préciser, nous souhaitons également prendre en compte le type de référence (collective ou non). Nous avons l'ambition d'analyser si la variation de la forme selon la distance évolue en fonction du référent puis du statut collectif ou individuel de ce dernier.

Il s'agit ainsi d'explorer les éventuelles régularités ou tendances dans le choix des formes référentielles et en particulier pour les formes pronominales et nominales.

Puisque ce mémoire cherche à observer les contextes de maintien de référents, la notion de chaîne de référence (CR dans la suite de ce mémoire) est intéressante à développer. En effet, les CR permettent d'assurer la cohésion textuelle en permettant de maintenir le sens et en assurant la reprise des référents tout au long du discours (Corblin, 1995).

Pour être en mesure de proposer des observations sur l'évolution des compétences référentielles d'élèves que l'on puisse généraliser, l'étude de corpus est primordiale. L'analyse de textes rédigés par des élèves et des étudiant.es doit se baser sur des données larges et de qualité afin de mettre en évidence des phénomènes récurrents (Elalouf, 2011). Les outils informatiques sont aujourd'hui nécessaires à cette étude (Mélanie-Becquet & Landragin, 2014 ; Schnedecker, 2021, 71), qu'ils se basent sur une annotation manuelle, automatique ou hybride. Nous souhaitons illustrer, à travers ce mémoire, l'apport des outils automatiques pour l'étude de la référence dans les textes d'élèves. En effet, notre script d'extraction est basé sur des

annotations de la référence déjà existantes dans le corpus RésolCo. Nous souhaitons nous inscrire dans une perspective analytique et montrer comment des approches outillées peuvent enrichir l'étude des mécanismes référentiels chez les élèves. La semi-automatisation de l'extraction des maillons référentiels permet un traitement systématique du corpus et ouvre la voie à des analyses quantitatives difficiles à envisager manuellement.

L'étude de productions écrites scolaires est essentielle pour améliorer la compréhension de la maîtrise de la continuité référentielle des élèves. En observant les évolutions entre différents cycles scolaires, cette recherche pourrait contribuer à une meilleure compréhension des mécanismes d'acquisition de la construction d'une chaîne référentielle. *In fine*, ce mémoire pourrait proposer un regard sur la gestion de la chaîne de référence dans les productions d'élèves et l'évolution de la maîtrise de cette compétence.

Ce mémoire s'organise en cinq étapes. La première partie est consacrée à un état de l'art. Cette partie présente des travaux théoriques portant sur la référence, les chaînes référentielles et la diversité grammaticale des maillons. Pour finir cette partie, nous abordons le sujet de l'acquisition de la compétence référentielle. La deuxième partie présente le corpus qui nous intéresse en précisant son origine, sa collecte ainsi que la manière dont les données sont structurées. La troisième partie décrit le processus mis en place pour extraire et traiter les chaînes de référence. Cette partie est suivie d'une analyse statistique des phénomènes observés, présentée dans la quatrième partie. Une première analyse complémentaire est proposée dans la cinquième partie. Enfin, la conclusion propose une synthèse des principaux résultats obtenus et évoque plusieurs pistes de réflexion pour prolonger cette étude.

## *Partie 1. Continuité référentielle : concepts-clés*

Pour construire un texte cohérent, le scripteur doit introduire un référent et assurer sa continuité. Oberle (2019 : 2), définit ce phénomène ainsi :

> *« Une expression référentielle, ou mention, est un segment de texte qui renvoie à une entité extralinguistique : le référent. »*

Notre travail s'appuie sur l'analyse de l'ensemble des expressions linguistiques qui renvoient à un même référent dans un texte, autrement dit, « un même référent qui apparaît sous des formes différentes » (Combettes, 2021).

### 1.1. La chaîne de référence : définition

Les ensembles d'expressions référentielles se nomment des chaînes de référence. Pour les analyser, il est primordial d'identifier les expressions qui les composent et les liens qui les unissent. Nous commencerons par définir les notions d'anaphore et de coréférence afin de mieux comprendre les dynamiques qui sous-tendent les chaînes de référence.

#### 1.1.1. L'anaphore

L'anaphore est un concept défini par Milner (1982 : 18),

> *« il y a relation d'anaphore entre deux unités A et B quand l'interprétation de B dépend crucialement de l'existence de A, au point que l'on peut dire que l'unité B n'est interprétable que dans la mesure où elle reprend entièrement ou partiellement A. »*

Cette définition permet d'affirmer qu'une expression anaphorique n'est pas autonome et a besoin d'être interprétée grâce au contexte. Dans leur ouvrage, Riegel et al. (1994) partagent cette définition et précisent que l'anaphore « implique le renvoi à un élément antérieur du texte » (Riegel et al., 1994 : 1029). En cela, elle s'oppose au terme de cataphore :

> *« cataphore, qui désigne le renvoi à un élément postérieur dans le texte » (Riegel et al., 1994 : 1029).*

Ces citations permettent de comprendre que l'anaphore désigne les renvois à un élément introduit antérieurement lorsque la cataphore désigne les renvois à un élément introduit de manière postérieure dans le texte. Dans notre mémoire, nous parlerons davantage d'*anaphore* mais les phénomènes de *cataphore* sont également pris en considération puisque nous travaillons à l'intérieur de chaînes de référence.

Dans la phrase 2), nous pouvons observer l'utilisation du pronom personnel *elle*. Ce pronom *elle* ne peut être interprété seul et il est impossible d'identifier correctement le référent dont il est question.

2) Elle habitait dans cette maison depuis longtemps.

Riegel et al. (1994) expliquent que le pronom de rang 3 (Riegel et al., 1994 : 368)

> *« fonctionne souvent, en effet, comme un anaphorique, ce qui explique qu'il soit le seul à varier en genre et en nombre en fonction des caractéristiques de son antécédent ».*

Grâce à cette variation, nous pouvons analyser le pronom car il marque le genre (ici féminin) et le nombre (ici singulier) et formuler des prédictions sur le référent désigné. Toutefois, cette inférence ne signifie pas pour autant que nous puissions avoir accès au référent sans faire appel à davantage de contexte.

L'exemple 3) est tiré d'une copie d'élève du corpus RésolCo, que nous présenterons dans la partie 2.1.2. Le corpus RésolCo. Dans l'extrait, nous pouvons retrouver la phrase 2), insérée dans un début de récit.

3) Il était une fois une fille du nom de Sarah qui avait un petit frère Maxence. Ils vivaient dans une maison qui était à proximité des bois, Sarah adorait cette maison. Elle habitait dans cette maison depuis longtemps. [CO-3e-2018-FSBJC6-D1-R6-V1]

La présence du nom propre *Sarah* comme instanciation du syntagme nominal (que nous nommerons simplement nom propre par la suite), permet d'identifier le référent du pronom *elle*. Grâce au contexte, le pronom peut être interprété. Cet exemple illustre une expression anaphorique dans laquelle le nom propre *Sarah* est repris par le pronom *elle*. Ces deux éléments sont « co-présents et accessibles dans le même environnement textuel » (Schnedecker, 2021 : 13).

Il est possible de rencontrer différents « procédés anaphoriques » (Riegel et al., 1994 : 1030) et notamment :

- L'anaphore coréférente, que l'on peut observer dans la phrase 4).

4) Il était une fois une fille du nom de Sarah. Elle vivait dans une maison qui était à proximité de les bois. [CO-3e-2018-FSBJC6-D1-R6-V2]

Dans cet exemple, *Sarah* et *elle* renvoient exactement au même référent.

- L'anaphore non coréférente, comme c'est le cas dans la phrase 5).

5) Cette vieille bâtisse demeurait dans sa famille depuis des générations. A peine en dépassait-on le seuil que l'on se trouvait assailli par l'ambiance lourde qui y régnait. [CO-3e-2018-FSBJC6-D1-R6-V2]

Dans cet exemple, l'anaphore est faite entre le tout et une partie, plus précisément entre la *vieille bâtisse* et *le seuil* qui est une partie de la demeure. C'est donc une anaphore associative, non coréférente.

Afin de mieux comprendre ce que signifient les nuances entre anaphore coréférente et anaphore non coréférente, il est important de présenter une définition de la coréférence.

### 1.1.2. La coréférence

La notion de coréférence est essentielle dans l'analyse du discours, notamment pour comprendre comment les entités sont identifiées, reprises et reliées entre elles au fil d'un énoncé. Milner (1982) en propose une définition, reprise par Schnedecker (2021 : 34) :

> *« Il y a relation de coréférence entre deux unités référentielles A et B quand elles se trouvent avoir la même référence - ce qui peut arriver sans que l'interprétation de l'une soit affectée par l'interprétation de l'autre » (Milner, 1982 : 32)*

Autrement dit, et comme le précise Corblin (1985), on peut parler de coréférence lorsque deux termes que l'on peut interpréter « de manière indépendante désignent en fait, dans un texte, le même individu » (Corblin, 1985 : 126). La différence principale entre la coréférence et l'anaphore réside donc dans la manière d'interpréter la référence.

L'exemple 6) montre une relation de coréférence.

6) John et ses deux fils séjournaient dans une vieille bâtisse appartenant à une femme veuve. Elle habitait dans cette maison depuis longtemps. Tout y semblait figé dans le temps, un temps lointain et regorgeant de mystères. John et les garçons avaient passé la soirée au coin du feu à écouter leur hôte leur conter l'histoire effrayante de l'ancienne propriétaire de la maison. [UN-M2-2018-TUTJ2-D1-R4-V1-]

Dans cette phrase, les deux expressions *une femme veuve* et *leur hôte* peuvent être interprétées de manière indépendante et renvoient au même référent.

Nous avons vu que les anaphores ne sont pas forcément coréférentes. De la même manière, la coréférence n'est pas forcément anaphorique si elle repose sur des expressions référentielles dites autonomes, que l'on peut interpréter indépendamment.

Dans le cas de la coréférence non stricte, le référent n'est pas exactement identique pour les deux expressions référentielles. Des difficultés apparaissent parfois lors de l'identification du référent. On peut parler d'ambiguïté référentielle lorsque le choix entre plusieurs candidats potentiels (Delaborde & Landragin, 2019) est difficile à faire. Lorsque le référent n'est pas clairement identifiable car il n'est pas désigné de manière précise, on peut évoquer la notion de flou référentiel (Landragin, 2007).

Les deux notions que sont l'anaphore et la coréférence sont étroitement liées. Tous les chercheurs n'utilisent pas les termes de la même manière et il est parfois possible de trouver un emploi où les deux notions sont « interchangeables », d'autres où la « coréférence est un sous-type d'anaphore » et « certains évitent totalement d'utiliser cette notion » (Ogrodniczuk et al., 2015 : 23, cité par Schnedecker, C., 2021 : 50). Toutefois, ces deux notions reposent sur une relation entre deux expressions qui figurent au sein d'un même énoncé. Dans ce mémoire, nous choisissons de prendre en compte l'ensemble des expressions référentielles, qu'elles soient liées par anaphore ou par coréférence. C'est pourquoi la notion de chaîne de référence nous semble particulièrement pertinente.

### 1.1.3. Définition de la chaîne de référence

Ce mémoire s'appuie sur la notion de chaîne de référence, définie par Corblin (1995 : 15) ainsi :

> *« On appelle donc chaîne de référence une suite d'expressions d'un texte entre lesquelles l'interprétation établit une identité de référence »*

Autrement dit, lorsqu'un ensemble d'expressions présentes dans un texte renvoie à une même entité, il forme une chaîne de référence. Des définitions plus récentes dans la littérature, comme celle proposée par Landragin (2021 : 7), s'appuient notamment sur la définition de Corblin (1995) et semblent converger.

> *« Une chaîne de référence (Chastain, 1975 ; Corblin, 1995 ; Schnedecker, 1997) regroupe l'ensemble des expressions référentielles qui réfèrent au même référent, c'est-à-dire au même individu, objet concret ou abstrait appartenant au monde extralinguistique. »*

Notre travail se concentre sur des référents qui renvoient à des individus, bien que les référents non humains puissent faire l'objet d'un travail de prolongement, comme nous l'évoquerons dans la partie Perspectives.

La définition de Schnedecker (2019 : 3) permet de préciser les liens référentiels attendus dans la *chaîne de référence* qui

> *« suppos[e] une forme d'identité référentielle entre les référents évoqués par les expressions référentielles, qui ne repose, ni nécessairement ni exclusivement, sur la relation d'anaphore »*

La chaîne de référence peut donc accepter que les liens entre les expressions soient anaphoriques ou coréférentiels. Cette chaîne peut parfois être nommée *chaîne de coréférence,* comme le souligne Schnedecker (2019 : 4)

> *« Notons au passage que, à la différence de la notion d'anaphore, les termes métalinguistiques visant à appréhender celle de chaîne de référence restent encore très fluctuants dans la plupart des travaux qu'ils soient d'obédience linguistique ou taliste : on parle ainsi de cohesive chains, de coreferential chains, de coréférences au pluriel. »*

Dans ce mémoire, nous choisissons d'utiliser les termes *chaîne de référence* ou *CR*.

Les expressions référentielles ou *mentions* (Salles, 2015) composant la chaîne sont nommées *maillons*. Les éléments en gras dans l'exemple 7) sont les maillons de la chaîne de référence renvoyant au référent *Il.*

7) Une nuit **Pierre** se réveilla et **il** avait faim. **Il** descendit donc pour aller à la cuisine. **Il** se retourna en entendant ce grand bruit.

La chaîne de référence est composée de 4 maillons : un nom propre, *Pierre* et trois pronoms personnels *Il*.

#### 1.1.4. L'intérêt de la notion de chaîne de référence

La chaîne de référence suppose, au minimum, trois expressions référentielles (Schnedecker & Landragin, 2014). Cette conception permet de dépasser la vision proposée par l'anaphore et la coréférence. En dessous de trois expressions, ces deux notions suffisent pour décrire les phénomènes de référence. Or, il peut être intéressant de prendre en compte les chaînes de référence dans leur ensemble, plutôt que d'observer des relations référentielles binaires comme c'est le cas lors de l'analyse de l'anaphore et de la coréférence. Comme le souligne Corblin (1995, cité par Schnedecker, 2021 : 50), la notion de chaîne

> *« permet de dépasser les contextes de simple succession de deux termes auxquels se limite le plus souvent le linguiste qui sort du domaine phrastique ».*

Ainsi, l'étude des CR offre une vision plus globale et dynamique de la référence, en tenant compte de l'ensemble des liens qui sous-tendent la narration.

Étudier les chaînes de référence, c'est mettre en lumière les mécanismes par lesquels les scripteurs structurent l'organisation de l'information. Les chaînes de référence permettent de suivre et d'analyser les différentes formes sous lesquelles un référent peut apparaître.

### 1.2. La diversité des maillons

Charolles (1988) propose une définition des maillons pouvant composer la chaîne de référence. Il affirme que

> *« seules peuvent appartenir (donner lieu à) une chaîne des expressions employées référentiellement, c'est-à-dire toutes et rien que les expressions nominales (ou pronominales) permettant d'identifier un individu (un objet de discours) quelle que soit sa forme d'existence (personne humaine, événement, entité abstraite) » (Charolles, 1988 : 8, cité par Schnedecker & Landragin, 2014).*

Les expressions référentielles s'opposent aux « expressions prédicatives qui leur attribuent des propriétés » (Charolles, 2002 : 23). Le prédicat attribue une propriété à un référent, cette propriété n'est pas introduite comme une entité. Dans :

8) Jeanne lui dit en étant très désolée : « Je te demande de m'excuser Paul ma mère ne fait plus le ménage, ni la cuisine, ni rien du tout depuis que mon père nous a quittées pour rejoindre l'au-delà (..) » [CO-4e-2018-LSPJJRC-D1-R22-V1]

l'élève scripteur réfère à un personnage, *Jeanne*, et lui attribue la propriété d'être *désolée*. Le participe passé employé comme adjectif qualificatif n'est donc pas un maillon de la chaîne de référence. La chaîne de référence de cet exemple est la suivante : *Jeanne – Je – m' – ma – mon – nous*. Elle est donc composée de six maillons.

La suite de notre travail présentera les formes des expressions référentielles qui nous intéressent particulièrement dans ce mémoire.

### 1.2.1. Les pronoms personnels et relatifs

La « catégorie syntaxique spécifique » constituée par les pronoms a des caractéristiques particulières (Zribi-Hertz & Godard, 2021). Tout d'abord, les pronoms peuvent être simples (je, lui) ou complexes (moi-même). Certains peuvent varier en personne, en genre ou en nombre tandis que d'autres sont invariables. Ils peuvent désigner des référents humains, des référents animés ou inanimés.

Les pronoms « constituent une classe extrêmement diverse » (Charolles, 2002 : 184). Cependant, dans le cadre de ce mémoire, nous avons choisi de nous concentrer uniquement sur les pronoms personnels et les pronoms relatifs.

Les pronoms personnels permettent de « référer de manière définie » (Charolles, 2002 : 183), de faire référence à des entités particulières bien qu'ils ne comportent pas de tête lexicale. C'est-à-dire qu'ils ne s'appuient pas sur un nom pour désigner leur référent. Les pronoms supposent que le référent est déjà connu et que les indices permettent de les interpréter.

Les référents des pronoms contenus dans un dialogue ne peuvent être identifiés qu'à partir de la situation d'énonciation (Landragin, 2021bis) :

> *« Elles ne décrivent pas leur référent en exprimant des propriétés, mais indiquent la procédure à suivre pour le trouver dans la situation. Ainsi je désigne 'la personne qui dit je' »*

Dans les écrits, la procédure peut être plus complexe car le moment de l'écriture ne coïncide pas avec le moment de la lecture, la situation d'énonciation ne regroupe donc pas le « locuteur et l'interlocuteur » (Landragin, 2021 bis). Son interprétation peut être réalisée à l'aide d'une indication précise, d'un nom propre ou d'une description. Dans l'exemple 8), les pronoms personnels *Je* ou *m'* peuvent être interprétés car ils sont introduits par un verbe de parole nécessaire à la reconstitution de la situation d'énonciation indiquant quel personnage prend la parole et par le nom propre *Jeanne.*

D'après Charolles (2021), les pronoms relatifs sont toujours anaphoriques. Le référent correspond au nom antécédent qu'ils complètent. La phrase

9) Il était une fois une femme qui s'appela Emse.

comporte un pronom relatif *qui* dont l'antécédent est *une femme*. Le référent du maillon *qui* est donc *une femme*, ils font donc partie de la même chaîne de référence et comportent un lien anaphorique.

Charolles (2021) évoque le fait que

> *« les pronoms relatifs (auquel, lequel, qui) ne sont pas sensibles au caractère actif ou saillant de leur référent, mais peuvent le mettre en valeur. »*

Le caractère actif et saillant d'un référent sera abordé dans la partie 1.3. Les théories cognitives. Ce qui retient notre attention ici, c'est le fait qu'un pronom relatif puisse mettre en valeur un référent.

Dans les exemples :

10) Il y a une mère qui sortait souvent le chien. [CO-3e-2016-VTAC305-D1-R3]

11) Il était une fois une femme qui s'appela Emse. [CO-6e-2016-PJPR1-D1-R9]

Le pronom relatif *qui* suit immédiatement les expressions référentielles *une mère* et *une femme.* Cet enchaînement permet, selon Charolles (2021), de mettre en valeur les référents. Cette utilisation du pronom relatif peut être importante pour l'analyse que nous souhaitons réaliser dans ce mémoire, puisque aucun token ne sépare les deux mentions et qu'il nous semble important de prendre en considération les cas dans lesquels deux expressions référentielles se suivent immédiatement.

### 1.2.2. Le syntagme nominal défini

Les syntagmes nominaux définis sont introduits par un déterminant et permettent de désigner un référent identifiable (Tasmowski & Laca, 2021). Les autrices précisent que

> *« Trois déterminants peuvent introduire un syntagme nominal défini : l'article défini, le déterminant démonstratif et le déterminant possessif »*

Ces trois déterminants présentent un référent qui peut être identifié par le lecteur ou l'interlocuteur, mais leurs propriétés varient. Un déterminant démonstratif « ajoute une forme de pointage » quand le déterminant possessif ajoute « une relation avec une autre entité » (Tasmowski & Laca, 2021). Charolles distingue clairement les trois types de SN définis (2002). Dans ce mémoire, nous choisissons de nous approcher de la perspective de Charolles (2002) et de présenter les SN démonstratifs et les SN possessifs dans des sections dédiées. Cette partie détaille les SN introduits par un article défini.

Un SN défini donne des informations sur la nature de son référent (Charolles, 2002 : 75). Il peut être complet et ne désigner qu'un seul référent (ou un ensemble de référents) défini. Les SN définis complets se distinguent par leur *autonomie référentielle*, le fait qu'ils puissent être interprétés sans faire appel à la situation d'énonciation car ils

> *« comportent en [eux]-mêmes des indications sur le contexte » (Charolles, 2002 : 85).*

Nous n'avons pas trouvé d'occurrence de SN défini complet dans notre corpus, mais nous pouvons prendre pour exemple la phrase 12), empruntée à Charolles (2002 : 75). Dans cet exemple, le référent est clairement identifiable et ne laisse que peu de place à l'ambiguïté.

12) La championne de France 1999 de planche à voile

Ce SN inclut des indications « spécifiant le contexte dans lequel [il est valide], indépendamment de la situation d'énonciation. »

A l'inverse, un SN défini incomplet

> *« conti[ent] des N qui peuvent s'appliquer à un nombre indéfini d'êtres ou de groupes d'êtres particuliers » (Charolles, 2002 : 76)*

C'est par exemple le cas pour la phrase 13). Les SN incomplets ont besoin d'un contexte particulier pour pouvoir être interprétés correctement et s'inscrire dans la mémoire des lecteurs. L'utilisation d'un SN défini permet d'attirer l'attention des lecteurs sur une entité qui est supposée identifiable.

13) La petite fille étonnée sortit de sa petite maison dans la forêt et aperçut des enfants. [CO-6e-2016-PJPR1-D1-R11-V1]

Le syntagme nominal défini *La petite fille étonnée* peut être désigné comme incomplet car un contexte est nécessaire pour l'interpréter. Il n'est pas autonome.

### 1.2.3. Le syntagme nominal démonstratif

Les expressions démonstratives sont souvent acceptables dans les mêmes contextes que les SN définis présentés plus haut, voire même,

> *« dans de nombreux contextes, ces deux formes sont substituables. » (Charolles, 2002 : 105)*

Cette possible substitution peut être observée dans les phrases 14) et 15).

14) Elle vivait désormais seule avec son majordome. Cet homme avait un regard glaçant et laissait l'impression que la mort n'était pas loin. [UN-M2-2021-UCL-D1-R43]

15) Elle vivait désormais seule avec son majordome. L'homme avait un regard glaçant et laissait l'impression que la mort n'était pas loin.

L'expression démonstrative *Cet homme* pourrait être remplacée par *L'homme* sans perdre la référence au personnage du *majordome*. La différence entre les deux formes référentielles est parfois difficile à identifier. Charolles (2002 : 106) explique que

> « *la différence entre définis et démonstratifs ne saurait se réduire à une opposition entre, d'un côté, les descriptions définies qui seraient indépendantes du contexte d'énonciation et, de l'autre, les descriptions démonstratives qui en seraient dépendantes.* »

Il est donc complexe de différencier les emplois d'un SN défini et d'un SN démonstratif. La différence entre les deux ne peut résider uniquement dans ces oppositions car les expressions démonstratives et définies peuvent s'appuyer sur le contexte de manière similaire et permettent de désigner un référent défini. Cependant, les SN démonstratifs s'appuient parfois sur la notion de « caractéristiques matérielles » (Charolles, 2002 : 106) et doivent être interprétés à la lumière de la situation d'énonciation dans laquelle ils sont produits :

> *« le démonstratif exploite les caractéristiques matérielles de la situation dans laquelle il est utilisé pendant que le défini exploite la conception que s'en font les participants à la communication. » (Charolles, 2002 : 107)*

En d'autres termes, le démonstratif renvoie à un référent dans la situation concrète d'énonciation, tandis que le défini s'appuie sur la représentation mentale que les interlocuteurs ont déjà construite.

### 1.2.4. Le syntagme nominal possessif

Le SN possessif est introduit par un déterminant possessif, qui s'accorde en genre et en nombre avec le nom.

Un syntagme nominal possessif réfère de manière particulière. En effet, Charolles (2021) explique que souvent dans cette expression « seul le déterminant est anaphorique ». C'est notamment le cas pour la phrase 16) et la phrase 17).

16) Lia était passionnée par le dessin. Depuis son plus jeune âge , elle adorait se balader en forêt avec ses parents et faire des croquis d'arbres et de fleurs. [UN-M2-2021-UCL-D1-R52]

17) Un jour, une fille du nom de Jeanne invite son ami Paul chez elle. [CO-4e-2018-LSPJJRC-D1-R22-V1]

Dans ces syntagmes nominaux possessifs, seuls les déterminants *son* et *ses* appartiennent à la chaîne de référence *Elle.* En revanche, les SN possessifs *son ami Paul* et *ses parents* sont des expressions référentielles permettant d'introduire de nouveaux personnages. Cet emploi du possessif comme introduction d'une nouvelle chaîne permet d'établir des liens entre les référents.

### 1.2.5. Le syntagme nominal indéfini

Le SN indéfini est introduit par un déterminant indéfini (comme *un, plusieurs, des,…*). Il présente le référent comme n'étant pas « immédiatement identifiable » (Tasmowski & Laca, 2021).

Les syntagmes nominaux indéfinis sont décrits par Charolles (2002) comme des expressions « autonomes référentiellement » car elles s'appuient d'abord sur les connaissances lexicales des lecteurs et n'ont pas besoin d'un contexte particulier. C'est pour cela que les SN indéfinis sont souvent utilisés pour introduire de nouveaux référents qui ne sont pas encore activés pour les destinataires. Nous pouvons le

remarquer dans la phrase 18) qui contient un SN indéfini expansé (*une vieille femme nommée Geraldine*) et qui constitue la première phrase du récit ou dans la phrase 19) qui contient le SN indéfini *une vieille sorcière*.

18) Une vieille femme nommée Geraldine habitait dans une vieille maison qui appartenait auparavant à sa grand-tante. [CO-4e-2018-LSPJJRD-D1-R17-V1]

19) Mais ce qu'ils ignoraient c'est qu'il y avait une vieille sorcière qui y habitait. [EC-CM2-2016-SGLEA-D1-R22-V1]

Ces exemples illustrent l'autonomie référentielle d'un syntagme nominal car *une vieille femme nommée Geraldine* et *une vieille sorcière* sont les premiers maillons des chaînes de référence dont ils font partie et ne s'appuient pas sur un contexte particulier pour être interprétés.

### 1.2.6. Les noms propres comme instanciation du syntagme nominal

Dans le cadre de notre analyse, nous avons choisi d'extraire les noms propres comme instanciation des syntagmes nominaux (SN). Les noms propres sont des référents uniques qui désignent de manière précise des entités spécifiques, comme des individus, des lieux ou des objets. Charolles (2002) précise cependant que les destinataires doivent être conscients que le nom propre utilisé réfère à une entité particulière. Ils permettent de concentrer l'attention du destinataire sur des référents pour pouvoir les réintroduire plus tard.

20) Ce week-end, Isabelle était si contente, ses cousins venaient lui rendre visite. [CO-3e-2016-VTAC305-D1-R19-V1]

Le référent désigné par le nom propre *Isabelle* est un référent unique qui désigne un individu. Ce référent est davantage mis en avant que le référent désigné par *ses cousins*. Les deux syntagmes sont en position sujet pourtant la présence du nom propre peut laisser présager que c'est bien *Isabelle* qui sera le personnage central du récit.

Il peut arriver qu'un nom propre et un syntagme nominal défini puissent être utilisés dans les mêmes contextes. Le référent *Elle* est désigné par un syntagme nominal défini dans la phrase 21) ou par un nom propre, comme dans la phrase 22) (*Caïly*). Les exemples 23) et 24) proposent d'inverser les expressions référentielles.

21) La fillette qu'elle était avait un sacré caractère. [CO-6e-2016-VTAC603-D1-R11]

22) Max était terrorisé alors que Caïly continuait d'avancer [CO-6e-2016-VTAC603-D1-R11]

23) Caïly avait un sacré caractère.

24) Max était terrorisé alors que la fillette continuait d'avancer.

Le sens des phrases est conservé, ce qui pourrait être un indice que les deux formes sont employées de manière équivalente.

Cependant,

> *«les Np désignent un particulier dans tous ses états, dans toutes les phases de son existence, sous toutes ses apparences, dans ce qui constitue son essence individuelle, alors que les descriptions définies n'exploitent jamais que des traits accidentels.» (Charolles 2002 : 55).*

Autrement dit, *Caïly* pourrait être présentée comme un personnage discret, dont le caractère est effacé. Elle pourrait être présentée comme *femme* car être une *fillette* n'est pas un trait immuable et inhérent. Cependant, il est nécessaire de l'introduire en tant qu'individu pour être en mesure de lui attribuer des traits. Searle, cité par Charolles (2002 : 57), résume les noms propres comme

> *« des outils commodes nous permettant de désigner des êtres particuliers sans avoir à nous engager et à nous mettre d'accord avec nos interlocuteurs sur ce qui en fait précisément des êtres singuliers. »*

Selon Charolles (2021), un nom propre, par sa spécificité et son poids sémantique, renforce la saillance d'un référent, là où un pronom la diminue. Désigner une entité par un nom propre place l'entité au centre et permet ensuite de la désigner par un pronom par la suite.

#### 1.2.7. Les indices référentiels

Pour finir, nous souhaitons élargir nos observations à des expressions référentielles qui ne sont pas évoquées par Charolles mais qui nous semblent pertinentes car elles permettent également de référer et doivent être prises en compte dans les chaînes de référence. Ces expressions sont présentées par Landragin (2011), et résultent d'un travail d'analyse de textes dans l'objectif de construire « un corpus de référence annoté finement en phénomènes référentiels et coréférentiels » (Landragin, 2011 : 61).

Selon Landragin, tous les éléments référentiels ne présentent pas le même degré de visibilité linguistique. Certains sont explicitement marqués comme les syntagmes nominaux ou les pronoms, évoqués plus tôt, et que l'auteur nomme « maillons forts », tandis que d'autres restent implicites. Landragin (2011) parle d'« indices » ou de maillons « faibles » pour les maillons invisibles comme le sujet non exprimé (ou sujet zéro). Il considère toutefois le sujet zéro comme un élément référentiel à part entière. L'exemple 25) permet de mieux comprendre comment il fonctionne.

25) Emse sortit et chercha son mari mais rien. [CO-6e-2016-PJPR1-D1-R9-V1]

Le verbe *sortit* a pour sujet *Emse*, tout comme le verbe *chercha.* Cependant, l'élève scripteur n'a pas réécrit le nom propre. Le sujet du second verbe est donc sous-entendu, implicite. Landragin (2011) explique que l'absence du sujet ne signifie pas que cet élément n'est pas inscrit dans la mémoire du lecteur, qu'il n'est pas saillant comme nous le détaillerons dans la prochaine partie.

D'autres éléments, tels que les accords verbaux ou participiaux, peuvent également jouer un rôle dans l'identification des référents, en apportant des indices de genre ou de nombre. Toutefois, Landragin (2011) explique ne pas les intégrer à son analyse, considérant que ces phénomènes relèvent d'un « aspect grammaticalisé et nécessaire » (Landragin, 2011 : 8). Nous choisissons d'adopter ce point de vue pour la suite du travail.

### 1.3. Les théories cognitives

Afin de mieux comprendre l'organisation de la référence, plusieurs auteurs ont proposé des théories pour tenter de qualifier les contraintes qui sous-tendent les choix effectués pour sélectionner les « catégories grammaticales des expressions référentielles » (Schnedecker, 2021 : 122) qui composent la chaîne.

#### 1.3.1. Le choix de l'expression référentielle selon Ariel

Dans les années 1980-1990, des théories dites cognitives apparaissent. Ces théories sont centrées sur l'étude du conditionnement du choix d'une expression référentielle. Elles s'appuient sur la notion de statut cognitif et présentent notamment l'idée que

> *« l'usage des différentes expressions référentielles dépend de / signale le statut cognitif supposé qu'occupe un référent dans la mémoire ou le focus d'attention du récepteur/lecteur. » (Schnedecker, 2021 : 125)*

Pour résumer, les formes référentielles ne sont pas neutres, elles marquent un lien avec le statut cognitif supposé du lecteur ou de l'interlocuteur. Le locuteur doit veiller à envisager ce que le destinataire connaît ou ce qu'il a actuellement en mémoire. Les expressions référentielles sont donc classées selon « leur degré d'accessibilité » (Schnedecker, 2021 : 126). Les théories cognitives suggèrent par exemple qu'un nom propre

peut être utilisé lorsque son référent est supposé être encore inconnu et peut donc être attendu en première mention. Au contraire, le pronom personnel, comme évoqué plus haut, désigne un référent accessible (Schnedecker, 2021). Ariel (1990, cité par Schnedecker, 2021 : 73) considère que les « anaphores zéro », comme les sujets zéros décrits plus haut, indiquent une accessibilité élevée.

Pour évaluer le degré d'accessibilité d'un référent, sa présence dans la mémoire ou l'attention du destinataire, Ariel (1990, 28-29, citée par Schnedecker, 2021 : 127) propose d'étudier quatre éléments :

> *« la distance entre l'antécédent et l'anaphore,*
>
> *la compétition i.e. le nombre de « compétiteurs » au rôle d'antécédent,*
>
> *la saillance : l'antécédent étant un référent saillant, surtout s'il s'agit d'un sujet ou d'un non-sujet (sic),*
>
> *l'unité : selon que l'antécédent se situe dans ou dehors du même cadre, monde, point de vue, segment ou paragraphe que l'anaphore. »*

Ces quatre critères permettent d'évaluer l'accessibilité d'un référent et de formuler un certain nombre de prédictions quant à la forme attendue pour l'introduire ou le maintenir dans le récit.

Toujours selon Ariel (1990), le lien entre l'accessibilité et la forme d'une expression référentielle ne dépend pas du hasard. Il dépend de son degré d'informativité, c'est-à-dire de sa quantité d'informations lexicales transmises et de sa rigidité. Une forme informative, comme un SN défini, est attendue pour désigner un référent moins accessible. Au contraire, une forme moins informative, comme un pronom, est attendue pour désigner un référent accessible. Une expression rigide réfère de manière moins ambiguë car elle désigne un référent unique et stable.

### 1.3.2. Le critère de la distance

Dans le cadre de ce mémoire, nous choisissons de nous concentrer sur un critère de la théorie d'Ariel afin de pouvoir observer sa réalisation dans les copies d'élèves et d'étudiant.es : la distance. Cette sélection est motivée par la volonté d'étudier un critère central. De plus, ce critère permet d'observer des indices concrets de la variation de la construction de la référence dans les copies d'élèves.

D'après Ariel (1990), plus la distance entre deux maillons est grande, moins le référent est accessible. A l'inverse, plus les maillons sont rapprochés, plus le référent est accessible. Ariel (1990) démontre que le pronom est davantage utilisé lorsque son antécédent est situé dans la même phrase ou dans la phrase précédente. Cette configuration permet d'affirmer que le pronom est plus présent lorsque son antécédent est proche et donc considéré comme plus accessible au regard du critère de la distance. A l'inverse, le pronom est moins présent lorsque l'antécédent se situe dans un paragraphe différent ou que les deux éléments sont séparés par plus d'une phrase. L'antécédent du pronom est donc moins accessible selon le critère de la distance.

Ce critère, qui peut sembler simple à observer, soulève cependant de nombreux questionnements relatifs :

> *« i) aux modalités de calcul : faut-il appréhender la distance en nombre de syllabes (Lust, 1981), de mots (Gernsbacher, 1990 ; Kibrik, 2011), de SN (Boudreau & Kittredge, 2005), de phrases (Givón, 1983), de nœuds rapportés à la structure rhétorique des textes (Kibrik, 2011) ? ii) aux termes du calcul : faut-il entamer le décompte à partir du premier mot de la mention ? de la tête du SN (Oberlé, 2016) qui permet de prendre en compte les SN enchâssés ? » (Schnedecker, 2019 : 8)*

Les choix méthodologiques en lien avec le décompte de la distance entre deux maillons doivent faire l'objet d'une réflexion. L'exemple 26) nous permet d'illustrer différents questionnements.

26) Ma fille Aurore logeait dans une belle maison, à la campagne. Elle habitait dans cette maison depuis longtemps. J'adorais aller la voir. [CO-3e-2016-VTAC305-D1-R15]

La chaîne de référence que nous étudions est composée de 3 maillons : le groupe nominal possessif *Ma fille Aurore* et les pronoms personnels *Elle* et *la*. Pour mesurer les distances entre ces éléments, il est important de définir précisément le point de départ : faut-il commencer à l'intérieur du groupe nominal ou à la fin de ce maillon ? Les distances en mots ou en syllabes seront impactées par ce choix et sont résumées dans le tableau 1.

| | Point de départ | En syllabes | En mots | En GN | En phrases |
|---|---|---|---|---|---|
| **Ma fille Aurore – Elle** | Fin GN1 | 12 | 8 | 2 | 0 |
| | Début GN1 | 17 | 11 | 3 | 0 |
| **Elle – la** | Fin GN2 | 17 | 9 | 1 | 0 |
| | Début GN2 | 17 | 9 | 1 | 0 |

*Tableau 1: Calcul de la distance selon différentes modalités*

Nous pouvons observer une variation dans le décompte de la distance dès que le maillon est composé de plusieurs mots. De plus, nous pouvons remarquer que le degré de variation du calcul dépend de l'unité choisie : certaines unités semblent lisser les mesures.

Cet exemple nous conduit à nous rapprocher des approches de Gernsbacher (1990) et de Kibrik (2011), citées par Schnedecker (2021). En effet, nous estimons que le décompte en mots est le plus pertinent pour notre travail. Toutefois, nous apportons une nuance et choisissons de compter en tokens, afin de permettre une analyse plus précise, incluant non seulement les mots, mais aussi la ponctuation et les symboles.

### 1.3.3. La forme du premier maillon de la chaîne

Le premier maillon d'une chaîne de référence est déterminant, car il introduit le référent dans le discours et initie la structure de la chaîne. Charolles (1987) souligne l'importance de l'ordre des mentions qui composent une CR. Selon cet auteur, la première mention dans une chaîne de référence a de fortes chances d'être un syntagme nominal indéfini ou défini, un nom propre ou une expression nominale démonstrative (dans un usage déictique). La phrase 27) introduit le référent *Elle* par un SN indéfini comportant un nom propre.

27) Une vieille femme nommée Geraldine habitait dans une vieille maison qui appartenait auparavant à sa grand-tante.[CO-4e-2018-LSPJJRD-D1-R17-V1]

Le pronom est moins susceptible d'apparaître en première mention car il suppose un référent accessible au préalable. Cependant, pour des raisons stylistiques, il arrive que certains auteurs fassent ce choix (Philippe, 1998 : 52).

28) Elle habitait dans cette maison depuis longtemps. En vérité, elle y était même née. Comme sa mère et sa grand-mère. Cette vieille bâtisse demeurait dans sa famille depuis des générations. A peine en dépassait-on le seuil que l'on se trouvait assailli par l'ambiance lourde qui y régnait. Non, Nina n'aimait définitivement pas cette maison. [UN-M2-2021-UCL-D1-R6-V1-5]

29) Elle habitait dans cette maison depuis longtemps. Cette maison, elle l'avait choisie avec son mari lors de leur première année de mariage. [UN-M2-2021-UCL-D1-R50-V1]

Dans la phrase 28), le pronom *Elle* introduit le référent *Elle* et montre un choix stylistique. Le pronom est employé dans un cadre de cataphore. Le nom propre *Nina* qui permet d'interpréter le pronom arrive après plusieurs reprises pronominales (*Elle, elle*) et un déterminant possessif (*sa*). Dans la phrase 29), l'usage du pronom *Elle* mène à un flou référentiel. Cette phrase est la première du récit, le pronom n'a pas d'antécédent clair et le lecteur construit sa propre hypothèse. L'interprétation ne se réalise pas dans une cataphore car la chaîne référentielle *Elle* est composée des maillons suivants : *Elle – elle – son – leur – elle – leurs parents*. Ses maillons sont principalement pronominaux. Le SN possessif référant à *Elle* ne permet pas d'interpréter précisément le premier pronom.

Par la suite du récit, les expressions référentielles peuvent prendre diverses formes selon les besoins imposés par le contexte narratif et les contraintes pesant sur le choix de la forme référentielle.

### 1.3.4. Le principe de redénomination

Lorsqu'un référent n'est plus actif, il est nécessaire de le réintroduire afin de pouvoir être compris par le destinataire. Dans cette perspective, Charolles (2021) évoque le principe de redénomination : il s'agit de rappeler le nom ou une partie du nom propre afin de clarifier la référence ou de remettre au centre une entité. L'exemple 30) nous permet d'observer une redénomination par le biais du nom propre *Christine*.

30) Il était une fois une très belle femme qui s'appelait **Christine** qui décida d'acheter une maison à une vieille femme. Elle habitait dans cette maison depuis longtemps. Mais d'après la vieille femme, sa maison est hantée. C'est pour ça qu'elle la vend. Mais **Christine** ne croyait pas à ça. **Elle** décida de s'installer dans cette maison avec son mari et ses enfants. [CO-3e-2016-VTAC305-D1-R10-V1]

La redénomination permet de remettre *Christine* au centre de l'action car le personnage n'est pas suffisamment saillant face à *la vieille dame*. En effet, *Christine* est introduit en même temps que *la vieille dame* et il est nécessaire d'éviter les ambiguïtés dans le choix de l'antécédent en clarifiant à l'aide d'un nom.

La redénomination constitue un moyen efficace pour remettre en lumière un référent devenu inactif dans le discours. Toutefois, ce procédé n'est envisageable que pour des entités pouvant recevoir un nom propre. Lorsqu'une telle nomination n'est pas possible, le locuteur doit recourir à d'autres formes, comme un défini ou un déterminant démonstratif, afin de réactiver le référent concerné.

Les définitions proposées par Charolles (2021) dans la Grande Grammaire du Français ne s'opposent pas aux théories cognitives développées par Ariel (1990). Au contraire, elles s'inscrivent dans la même logique, dans la mesure où des critères comme la saillance, la compétition entre référents ou encore la distance (discursive ou cognitive) sont justement ceux qui permettent de déterminer les statuts cognitifs d'un référent.

Après avoir présenté les fondements théoriques de la référence, nous nous intéressons à présent à la place de la référence dans le contexte scolaire, au cœur de notre travail.

## 1.4. L'acquisition et l'évolution des compétences référentielles

Cette partie se penche sur l'acquisition de la référence. Tout d'abord nous explorons comment l'acquisition des compétences référentielles est abordée dans la littérature. Dans un second temps, nous présentons la prise en compte de la référence dans les programmes du cycle 3 et du cycle 4.

### 1.4.1. L'acquisition des compétences référentielles

L'étude de Salazar Orvig et al. (2004 : 80)

> *« a permis de saisir l'émergence des pronoms de 3° personne et l'augmentation de leur importance à partir de 2 ans et demi »*

mettant ainsi en évidence les prémices du développement des compétences référentielles chez l'enfant. Dans la continuité de cette perspective développementale, cette partie présente plusieurs travaux qui se sont intéressés à l'évolution de ces compétences au fil de la scolarité. Charolles (1988 bis), par exemple, explore les défis de la gestion des risques de confusion entre des personnages. Il a été demandé à des élèves de rédiger des textes narratifs à partir de planches dessinées. Dans la synthèse des résultats, il explique que

> *« du CM1 à la 6e le nombre d'emplois prêtant à confusion dans les textes obtenus diminue sensiblement »*

Plus précisément, Charolles (1988 bis : 82) propose un tableau dans lequel sont indiqués les pourcentages de cas confus, par niveau scolaire. Les résultats mettent en lumière que les CM1 présentent un taux de 8,25% d'emplois fautifs « relativement au nombre total des formes de reprise et de désignation » contre 4 % pour les CM2 et seulement 3 % en 6e (Charolles, 1988 bis : 82). Ces résultats doivent cependant être nuancés car les élèves de CM2 et 6e ont souvent recours à des stratégies « d'évitements », c'est-à-dire que les élèves choisissent de ne pas évoquer une partie du récit présent dans la planche et qui est plus complexe que le reste de l'histoire. De plus,

> *« le recours au nom propre, qui constitue une stratégie habile pour résoudre les problèmes rédactionnels rencontrés dans cette épreuve, est plutôt le fait des élèves du CM2 et de 6e » (Charolles, 1988 : 86)*

Plus précisément, les CM1 ont recours à 6,7 % de noms propres, les CM2 15,2 % et les 6e 16,5 %.

Ces observations semblent indiquer une évolution des compétences référentielles des élèves.

Dans une étude de 1978, Charolles indique également que des élèves de CE2, CM1 et CM2 « semblent maîtriser assez bien » (Charolles, 1978 : 16) les procédures référentielles telles que les « définitivisations » (utilisation d'un déterminant défini pour rappeler un substantif). De plus, les élèves semblent correctement déduire des informations implicites à partir de l'énoncé.

Charolles (1978) précise cependant que les pronominalisations restent complexes pour les élèves. Il repère des emplois anaphoriques sans référent clairement introduit ou des ambiguïtés référentielles. Il précise que l'interprétation d'une expression référentielle peut être « compromise par l'éloignement » entre deux mentions.

Une étude de Pepin (2009) s'intéresse à la coréférence dans la narration d'élèves du secondaire (environ 14 ans) au Québec. Elle indique que la gestion de la coréférence est complexe et propose une classification des défis rencontrés par les élèves, qu'elle nomme des « défauts ».

Nous pouvons notamment retrouver (Pepin, 2009 : 5) :

- *antécédents absents,*
- *omission de la coréférenciation (répétition d'un terme qui aurait dû être substitué),*
- *mauvais choix de déterminant,*
- *substituts imprécis.*

D'autres catégories sont mentionnées, dont le défaut le plus fréquent dans le corpus de l'étude : l'antécédent trop éloigné (28 % des défauts repérés).

Ces travaux mettent en lumière une progression de la maîtrise référentielle : si les élèves développent des stratégies, certaines difficultés persistent. La partie suivante présente les programmes scolaires évoquant les compétences référentielles.

### 1.4.2. La référence dans les programmes scolaires

Dans le cadre de ce mémoire, nous avons choisi de nous intéresser particulièrement au cycle 3 et au cycle 4.

Les programmes nationaux de français du cycle 3 concernent les élèves de CM1, CM2 et 6e. Ils introduisent la notion de référence anaphorique comme un levier pour la compréhension d'un texte littéraire. Ainsi, la compétence « Comprendre un texte littéraire et se l'approprier » du BOEN 25 du 22 juin 2023, met en valeur l'importance de la « vigilance quant aux reprises nominales et pronominales ». De plus, il est demandé de « prendre en compte les normes de l'écrit pour formuler, transcrire et réviser » ce qui inclut les compétences « Respecter la cohérence et la cohésion : syntaxe, énonciation, éléments sémantiques qui assurent l'unité du texte » et « Utiliser les connecteurs logiques, temporels, les reprises anaphoriques, les temps verbaux pour éviter des dysfonctionnements ».

Cette maîtrise se poursuit « en se complexifiant » au cycle 4, qui concerne les classes de cinquième, quatrième et troisième, par « l'étude du lexique, de la syntaxe et de la cohérence textuelle ». Ainsi, d'après le BOEN 31 du 30 juillet 2020 la compétence « construire les notions permettant l'analyse et l'élaboration des textes et des discours » propose aux enseignants et enseignantes de travailler cette notion à l'aide d'

> *« exercices de variation et de substitution : repérage des substituts nominaux et pronoms de reprise ; procédés de désignation et de caractérisation, rôle des déterminants ; transfert de ces notions dans l'expression écrite ou orale. ».*

Il apparaît donc clairement que les compétences des élèves doivent évoluer entre le cycle 3 et le cycle 4 et que la notion de reprise est importante tout au long de la scolarité. Cette notion est un élément clé de l'apprentissage, mobilisée pour la lecture, l'écriture et la maîtrise de la langue.

Pour conclure, dans cette première partie nous avons défini les phénomènes référentiels qui seront au cœur de notre étude. Nous avons montré que le choix de la forme de l'expression référentielle est sous-tendu par un ensemble de contraintes présentées par les théories dites cognitives. Nous avons particulièrement mis en avant le critère de la distance entre deux mentions d'une même chaîne de référence. Pour finir, nous avons présenté comment les compétences référentielles sont abordées dans les programmes scolaires et la littérature.

Au terme de cet état de l'art, il nous semble possible de dire que si la littérature décrit de nombreux aspects de la gestion référentielle, en particulier dans des corpus scolaires, la relation précise entre la distance séparant deux maillons d'une chaîne de référence et le choix de la forme référentielle reste à explorer de manière systématique. Notre travail pose la question suivante : dans des textes narratifs scolaires, comment la distance entre deux maillons d'une chaîne de référence est-elle liée au choix de la forme employée ? En particulier entre deux « maillons forts » : les syntagmes nominaux et les pronoms. Y aurait-il un lien entre le niveau scolaire, la distance inter-maillons et la forme employée ? Comment le référent influence-t-il ces choix ? Ces questions visent à éclairer les régularités ou tendances observables dans la gestion des chaînes de référence et à caractériser l'évolution de cette compétence dans la scolarité. La partie suivante décrit le corpus composé de travaux d'élèves et d'étudiant.es sur lequel nous basons notre travail : le corpus RésolCo ( CLLE et l'université Toulouse – Jean Jaurès).

## *Partie 2. Présentation des données*

Ce chapitre se concentre sur la présentation des données sur lesquelles ce travail s'appuie. Dans un premier temps, il s'agira de décrire la constitution du corpus. Dans un second temps, nous présenterons le processus d'annotation des données. Pour finir, nous expliquerons quel traitement est appliqué aux données pour produire les fichiers qui serviront de support pour notre travail.

### 2.1. La constitution des données

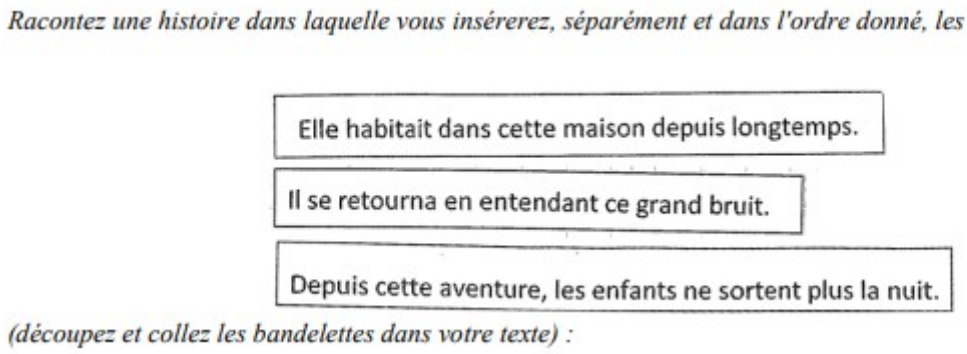
Les données de ce mémoire ont été constituées préalablement dans le cadre d'un projet. Nous souhaitons présenter ce projet afin de situer le contexte de la collecte et les objectifs poursuivis.

#### 2.1.1. Le projet E-CALM

Le projet E-CALM réunit quatre laboratoires en linguistique et en didactique : CIRCEFT (Paris 8), CLLE (Toulouse 2), CLESTHIA (Paris 3) et LIDILEM (Grenoble). Il a pour objectif de constituer un « grand corpus d'écrits scolaires et universitaires » pour permettre un nouveau regard des enseignants et enseignantes sur les écrits des élèves. Le projet E-CALM s'inscrit dans la volonté de rendre possible une exploration outillée (Ponton et al., 2022) de travaux scolaires et universitaires. Le travail initial s'est déroulé entre janvier 2018 et décembre 2022. Il est composé de quatre sous-corpus accessibles en ligne.

#### 2.1.2. Le corpus RésolCo

Ce mémoire se basera sur l'analyse d'une partie du corpus RésolCo, constitué par le laboratoire CLLE de l'université de Toulouse - Jean Jaurès. Ce corpus est construit afin d'étudier la maîtrise de la cohérence et de la continuité référentielle à différents niveaux scolaires. Pour permettre cette étude, une consigne unique est donnée aux participants et cherche à provoquer la résolution de problèmes de référence (Garcia-Debanc et al. 2017). L'illustration 1 correspond à la consigne retenue pour la construction du corpus.


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

Elle habitait dans cette maison depuis longtemps.

Il se retourna en entendant ce grand bruit.

Depuis cette aventure, les enfants ne sortent plus la nuit.

*(découpez et collez les bandelettes dans votre texte) :*

*Illustration 1: Consigne donnée aux élèves pour la constitution du corpus RésolCo (Garcia-Debanc et al. 2017)*

Cette consigne propose aux élèves de rédiger un texte narratif fictionnel. Les élèves doivent insérer dans leur récit des pronoms personnels (*il* et *elle*), un groupe nominal défini (*les enfants*) et des groupes nominaux composés d'un démonstratif et d'un nom (*cette maison, ce grand bruit*, *cette aventure*). Ces éléments doivent être introduits et faire l'objet de reprises afin de produire un texte cohérent. L'utilisation d'étiquettes à découper permet de s'assurer que l'élève ne modifie pas les phrases de la consigne.

Les élèves et les étudiant.es qui participent à l'étude fréquentent des établissements d'Occitanie, d'Île-de-France (Ho-Dac et al., 2020) et de Belgique. Le choix des établissements permet d'obtenir des

profils variés (rural, urbain, REP). Pour permettre de repérer facilement certaines informations sur les conditions de collectes, chaque copie porte un identifiant unique. L'illustration 2 permet de comprendre la composition de l'identifiant.

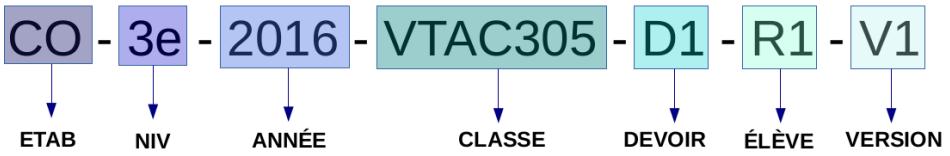



*Illustration 2: Composition des identifiants des fichiers - RésolCo*

Le premier code représente l'établissement et permet de montrer si la copie provient d'une école (EC), d'un collège (CO) ou de l'université (UN). Les éléments qui suivent permettent d'avoir des informations sur l'année de la rédaction, l'identifiant de la classe et celui du devoir, l'identifiant de l'élève précédé d'une lettre correspondant au corpus (ici R pour RésolCo) et la version du devoir. Dans notre travail, cet identifiant accompagne systématiquement les exemples entre deux crochets.

### 2.2. Le processus d'annotation de la continuité référentielle

Nous allons détailler dans cette partie les processus qui ont permis d'annoter la continuité référentielle dans les copies du corpus ainsi que les critères retenus.

Les copies sont collectées puis numérisées, transcrites manuellement de la manière la plus fidèle possible en prenant en compte les éventuelles marques de l'élève (rature, illustration) et anonymisées. Des métadonnées sont ajoutées, permettant de conserver des informations sur la collecte des copies. Des annotateurs procèdent à la normalisation orthographique en repérant les mots mal orthographiés et en proposant une correction. Ces étapes sont essentielles pour l'annotation et l'automatisation de l'exploration de la continuité référentielle.

Les personnes annotant la continuité référentielle dans le corpus RésolCo ne prennent en compte que les référents humains présents dans l'ensemble des copies du corpus, c'est-à-dire ceux liés à la consigne d'écriture présentée plus haut (*il*, *elle*, *les enfants*). Cette annotation permet de se concentrer sur les référents jouant un rôle de premier plan dans le récit et de créer des annotations comparables entre textes de même niveau et à travers le corpus. Chaque copie a été annotée manuellement par au moins une personne francophone. Pour les cas d'ambiguïté ou de doute, un *gold standard* a été rédigé. Le guide d'annotation RésolCo présente les différents éléments à annoter pour identifier les maillons. Autrement dit, le guide permet de maîtriser la cohérence et la comparabilité des annotations entre différents annotateurs et différents textes. L'exemple 31) nous permet de montrer diverses expressions référentielles qui peuvent être annotées.

31) Il était une fois **une petite princesse qui** s'appelait **Iris**. **Elle** habitait dans cette maison depuis longtemps Un jour dans le bus pour aller au collège **elle** a rencontré un petit prince qui s'appelait Tony dans le bus Tony a entendu xxx et là il se retourna en entendant ce grand bruit Tony parle à **Iris** dans le bus mais **Iris** ne sait pas xxx au garçon comme c'est le 1er garçon qu'**elle** a rencontré de **sa** vie. Depuis cette aventure, *les enfants* se sortent plus la nuit Et enfin **Iris** et Tony se marièrent ***ils*** vécurent heureux pour toujours. [CO-6e-2018-VTAC605-D1-R1]

Les maillons qui sont pris en compte par les annotateurs selon le guide d'annotation concernent *Elle* (en gras dans l'exemple), *Il* (soulignés dans l'exemple) et *Les* Enfants (en italique dans l'exemple). Les formes de maillons annotées sont les suivantes :

- Syntagmes nominaux (SN), qui incluent les groupes nominaux, les adjectifs et les groupes prépositionnels. Dans cet exemple : *une petite princesse* , *au garçon, un petit prince, le premier garçon*.

- Possessifs. Dans cet exemple : *de sa vie*.
- Pronoms, sauf les pronoms réfléchis. Dans cet exemple : *elle, ils*, *il*, *qui, qu'*.
- Noms propres (NP). Dans cet exemple : Iris, *Tony*.
- Sujet zéro, dans le cas où le sujet n'est pas explicite, le groupe verbal est inclus dans le maillon. Par exemple : *Elle s'est enfuie et ne s'est jamais remise de ce drame*.(Federzoni et al., 2011)
- Référents multiples, doivent comporter le trait « groupe ». Pour qu'un référent soit considéré comme multiple, il doit contenir le référent et une autre identité. Par exemple : *toute la famille.* (Federzoni, S., et al., 2011) ou dans notre exemple, *ils* qui désigne *Elle* et *Il.*

Le guide précise également que les maillons contenus dans un discours direct doivent être pris en compte, les auteurs considèrent en effet que le passage au discours direct ne constitue pas une rupture de la chaîne, tout comme pour les maillons apparaissant dans un titre.

### 2.3. Le traitement des fichiers

Les fichiers que nous utilisons sont accessibles dans divers formats, mais notre préférence va au format conllu. Ce format a été produit automatiquement grâce à Stanza, une bibliothèque Python développée par le Stanford NLP Group. Le résultat final se présente sous la forme d'un tableau de 10 colonnes séparées par des tabulations. L'illustration 3 est un extrait d'une copie au format conllu.

```
# text = Une jour, une fille du nom de Jeanne invite son ami Paul chez elle.
# sent_id = CO-4e-2018-LSPJJRC-D1-R22-V1-0
1	Une	un	DET	_	Definite=Ind|Gender=Fem|Number=Sing|PronType=Art	2	det	_	start_char=0|end_char=3|ner=O|mod=0|orig=0|coref=0|instr=0
2	jour	jour	NOUN	_	Gender=Masc|Number=Sing 11	obl:mod _	start_char=4|end_char=8|ner=O|mod=0|orig=0|coref=0|instr=0
3	,	,	PUNCT		2	punct	start char=8|end char=9|ner=O|mod=0|orig=0|coref=0|instr=0
```

*Illustration 3: Présentation du format conllu*

Chaque token présente les informations suivantes, détaillées par le *readme* du gitlab du corpus annoté :

- ID : l'index du mot. Chaque phrase commence par 1.
- Form : Forme du mot ou de la ponctuation.
- Lemma : Lemme ou racine de la forme du mot.
- UPOS : Étiquette de catégorie grammaticale universelle.
- XPOS : Nos données ne sont pas concernées par cette étiquette.
- Feats : Liste des caractéristiques morphologiques issues de l'inventaire des caractéristiques universelles ou d'une extension définie pour une langue spécifique ; un tiret bas (_) si non disponible.
- Head : La tête du mot (ID ou 0) en lien avec le « deprel »
- Deprel : Relation de dépendance universelle.
- DEPS : Nos données ne sont pas concernées par cette étiquette.
- MISC : Les annotations RésolCo, et notamment celles qui concernent la référence. Notre corpus contient ici
    - mod=0|del|subst|add|na indique si une correction a été apportée : suppression (del), ajout (add) ou substitution (subst)
    - orig=0|xxx|na indique si le token a été correctement orthographié (orig=0) ou non. Si ce n'est pas le cas, xxx donne la forme originale.
    - coref=0|(xxx)|(xxx|xxx|xxx) indique si le token a été annoté comme une expression référentielle qui réfère à l'un des personnages principaux du récit. Si le token n'a pas été annoté comme inclus dans une mention on trouvera *coref=0*. Les parenthèses permettent de montrer la position

du token dans une mention composée de plusieurs tokens. Si le token correspond à une mention composée d'un seul token : *coref=(xxx)*. Si le token correspond au début d'une mention multi-token : *coref=(xxx.*. Si le token correspond à la fin d'une mention multi-token : *coref=xxx)*. Pour finir, si le token est inclus dans une mention collective, l'indication xxx est suivie de la chaîne "Coll" .

- instr=0|1| indique si la phrase fait partie de celles de la consigne (1) ou non (0)

Prenons l'exemple 32) afin de présenter le format conllu.

32) Il vit un chien dans le jardin il courut par-dessus la clôture. [CO-3e-2016-VTAC305-D1-R1-V1-5]

Nous pourrons alors trouver en sortie l'analyse suivante pour les premiers tokens.

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 1 | Il | lui | PRON | _ | Emph=No\|Gender=Masc\|Number=Sing\|Person=3\|PronType=Prs | 2 | nsubj | _ | start_char=406\|end_char=408\|ner=O\|mod=0\|orig=Il\|coref=(II)\|instr=0 |
| 2 | vit | vivre | VERB | _ | Mood=Ind\|Number=Sing\|Person=3\|Tense=Pres\|VerbForm=Fin | 0 | root | _ | start_char=409\|end_char=412\|ner=O\|mod=0\|orig=0\|coref=0\|instr=0 |

*Tableau 2: Extrait du fichier CO-3e-2016-VTAC305-D1-R1-V1.conllu*

Nous pouvons voir que le premier mot, *il* a pour lemme *lui*, que c'est un pronom personnel masculin et singulier de 3$^{e}$ personne. Cet élément est lié au second dont il est le *nsubj*. Il est identifié comme élément coréférent de *il* dans une expression composée d'un seul token. Le second token, à l'inverse, n'est pas identifié comme faisant référence à un personnage (coref=0). Nous savons toutefois que c'est un verbe dont le lemme est *vivre*, qu'il est conjugué au présent de l'indicatif, à la troisième personne du singulier. Il est la racine de la phrase (*root*).

Les différentes colonnes permettent d'obtenir des informations primordiales pour notre travail. Par exemple, chaque ligne comporte des informations permettant d'identifier et de préciser les classes grammaticales (également POS dans la suite de ce mémoire). Dans le tableau 2, on peut retrouver l'étiquette PRON pour le pronom *il* et VERB pour le verbe *vit*.

Les copies contiennent également des métadonnées, présentées notamment sous forme de commentaires comme l'illustre l'exemple 33).

33) # sent_id = CO-4e-2018-LSPJJRC-D1-R22-V1-0

Sur cet exemple, nous pouvons reconnaître l'identifiant présenté dans la partie 2.1.2. Le corpus RésolCo. Les informations contenues dans cet identifiant nous permettent d'obtenir des précisions sur les conditions de collecte de la copie. Le dernier élément, ici *0,* indique le rang de la phrase associée. Dans notre cas, la phrase qui suit est la première de la copie.

### 2.4. Aperçu quantitatif des données

A ce jour, 964 copies ont été annotées. Le tableau 3 permet d'avoir un aperçu des données qui composent le corpus RésolCo (Ho-Dac L.-M., et al., 2020). On peut voir le détail du nombre de copies, de phrases et de tokens pour chaque classe, du CE2 au Master 2.

| | Copies | Phrases | Token |
|---|---|---|---|
| CE2 | 258 | 2062 | 31483 |
| CM1 | 65 | 535 | 7710 |
| CM2 | 204 | 2416 | 36872 |
| 6e | 114 | 1442 | 22191 |
| 5e | 4 | 70 | 1196 |
| 4e | 47 | 864 | 13754 |
| 3e | 68 | 1463 | 24683 |
| 2nde | 15 | 141 | 2432 |
| Master 2 | 189 | 4523 | 83828 |
| Total | 964 | 13516 | 224149 |

*Tableau 3: Aperçu quantitatif de la ressource RésolCo*

Cependant, comme évoqué dans la partie 1.4. L'acquisition et l'évolution des compétences référentielles, dans le cadre de ce mémoire nous avons choisi de nous intéresser particulièrement au cycle 3 et au cycle 4. Nous choisissons également d'inclure des copies d'étudiant.es de Master (M2 dans le corpus). Ce choix s'inscrit dans une volonté de comparer des phénomènes référentiels tout au long de la scolarité. En incluant le niveau Master, nous souhaitons observer comment les compétences référentielles évoluent avec l'âge et l'exposition à des pratiques langagières plus complexes. En effet, les étudiant.es de Master ayant participé à l'élaboration de ce corpus suivent des cursus tels que « Master Professeurs des Écoles, Master Métiers de l'écriture et Master Sciences du Langage » (Garcia Debanc et al., 2017). L'écriture occupe une place importante dans ces parcours et nous pouvons supposer que les étudiant.es ont acquis une maîtrise solide des compétences référentielles. Les copies de Master nous permettront de porter un regard sur les évolutions entre les cycles et de les mettre en perspective avec les compétences de scripteurs que l'on peut qualifier d'experts.

Nous ne retenons pour cette étude qu'une partie du corpus RésolCo, que nous présentons dans le tableau 4.

| | Copies | Tokens | Maillons annotés |
|---|---|---|---|
| Cycle 3 | 238 | 41266 | 6845 |
| Cycle 4 | 119 | 39633 | 5373 |
| Master | 189 | 83828 | 3229 |
| Total | 546 | 164727 | 15447 |

*Tableau 4: Composition du corpus de cette étude*

Il apparaît que le corpus du cycle 3 est plus fourni que le corpus du cycle 4 et du Master. En effet, le corpus du cycle 3 comporte 119 copies de plus que le corpus du cycle 4, soit un corpus 50% plus fourni. De plus, l'extraction présentée dans la partie Partie 3. Extraction des chaînes de référence repère environ deux fois plus de maillons pour le cycle 3 (6845) que pour le Master (3229).

Nous faisons cependant le choix de conserver l'ensemble des copies des corpus, estimant que cette disparité ne constitue pas un obstacle pour l'analyse que nous souhaitons mener puisque les tests statistiques envisagés permettent de prendre en compte des corpus dont la taille est inégale.

## *Partie 3. Extraction des chaînes de référence*

Cette partie se concentre sur trois aspects essentiels. Tout d'abord, la présentation de la sortie attendue. Ensuite, la description du gold standard, sa constitution et sa pertinence. Pour finir, la description et l'évaluation de la sortie. Cette sous-partie proposera une présentation quantitative ainsi que l'analyse des résultats générés par l'extraction automatique, incluant une évaluation critique de leur fiabilité et des erreurs remarquées.

### 3.1. Choix méthodologiques

Les données utilisées pour l'analyse doivent être extraites spécifiquement pour répondre aux objectifs de l'étude. Cette partie détaille la méthode choisie et les caractéristiques des extractions attendues.

#### 3.1.1. Méthode d'extraction des maillons

Notre extraction des maillons se base sur les fichiers conllu. Ce format permet d'avoir accès aux annotations accompagnant le corpus ReSolCo, comme décrit dans la partie 2.3. Le traitement des fichiers .

Dans un premier temps, nous avons choisi d'extraire tous les termes comportant une indication de coréférence dans la colonne misc, à l'aide d'une expression régulière indiquant *Elle, Il* ou *Les Enfants.*

Dans un second temps, nous avons ajouté de nouvelles contraintes afin de regrouper correctement les expressions référentielles composées de plusieurs mots. Pour cela, nous avons eu recourt à un système d'automate accompagné d'expressions régulières afin de baser l'extraction sur les parenthèses. Cette étape permet d'identifier si un maillon est seul (s'il est contenu entre deux parenthèses), s'il est le premier maillon d'une expression composée (seule la parenthèse ouvrante est présente), s'il est contenu dans une expression (aucune parenthèse n'est présente) ou s'il termine l'expression en court (seule la parenthèse fermante est présente). Ces contraintes permettent de séparer deux maillons consécutifs correctement. Cette étape est apparue nécessaire lors de l'analyse des extractions du gold (3.3.1. Script python appliqué aux copies annotées : analyse des erreurs).

Prenons la phrase :

> 34) Personne au village ne connaissait le nom de la femme qui habitait dans cette maison, ni le nom de ses enfants. [CO-3e-2016-VTAC305-D1-R16-V1]

La chaîne de référence *Elle* est composée du SN défini *la femme*, du pronom relatif *qui et* du déterminant possessif *ses.* La chaîne de référence *Il* est composée de la mention collective *ses enfants*, tout comme la chaîne *Les Enfants.* Le tableau 5 correspond aux annotations de cette phrase.

| | |
|---|---|
| Personne | \|coref=0\|instr=0 |
| au | \|coref=0\|instr=0 |
| à | _ |
| le | _ |
| village | coref=0\|instr=0 |
| ne | coref=0\|instr=0 |
| connaissait | \|coref=0\|instr=0 |
| le | \|coref=0\|instr=0 |
| nom | \|coref=0\|instr=0 |
| de | \|coref=0\|instr=0 |
| la | \|coref=(Elle\|instr=0 |
| femme | \|coref=Elle)\|instr=0 |
| qui | \|coref=(Elle)\|instr=0 |
| habitait | \|coref=0\|instr=0 |
| dans | \|coref=0\|instr=0 |
| cette | \|coref=0\|instr=0 |
| maison | \|coref=0\|instr=0 |
| , | \|coref=0\|instr=0 |
| ni | \|coref=0\|instr=0 |
| le | \|coref=0\|instr=0 |
| nom | \|coref=0\|instr=0 |
| de | \|coref=0\|instr=0 |
| ses | \|coref=(Elle),(IlColl,(lesEnfants\|instr=0 |
| enfants | \|coref=IlColl),lesEnfants)\|instr=0 |

*Tableau 5: Extrait de l'annotation de la copie CO-3e-2016-VTAC305-D1-R16-V*

Ce tableau met en lumière les différentes annotations qui peuvent exister dans notre corpus. Le maillon *la femme* est composé de plusieurs mots référents, qui sont indiqués par les parenthèses : le début *(Elle* et la fin *Elle).* Le maillon *qui* est unique, c'est pourquoi sa référence est notée *(Elle)*. Le cas du SN possessif est plus complexe. Le déterminant *ses* est à la fois un maillon unique *(Elle)* et le début d'un maillon composé pour deux référents, l'un collectif et l'autre non : *(IlColl* et *(lesEnfants*. Le nom *enfants* marque la fin de ce maillon *IlColl)* et *lesEnfants).*

### 3.1.2. Méthode d'extraction d'informations

Pour la constitution d'un corpus permettant d'observer certains phénomènes décrits plus haut, nous avons choisi d'extraire les pronoms ainsi que les syntagmes nominaux ayant été identifiés comme maillons d'une chaîne de référence. Ces formes représentent les éléments principalement impliqués dans les phénomènes de reprise référentielle, c'est pourquoi nous choisissons de restreindre notre sélection pour ce mémoire.

La classe grammaticale de chaque maillon constitue une variable centrale pour notre analyse. Nous proposons un classement en trois catégories : pronom, syntagme nominal et autre. Ce classement repose sur l'étiquette morphosyntaxique attribuée par l'outil d'annotation Stanza. Les maillons référents étiquetés comme pronoms sont classés dans la catégorie *pronom*. Ce choix méthodologique vise à garantir une analyse statistique fine sans pour autant la surcharger et à renforcer la comparabilité des résultats. Ce classement n'exclut pas d'approfondir l'analyse des pronoms relatifs séparément des pronoms personnels dans un travail ultérieur. Dans le cadre de ce mémoire de Master 1, ce classement nous semble correspondre aux objectifs fixés, centrés sur la relation entre la distance inter-maillons et la forme référentielle employée.

Les maillons comportant un nom commun ou un nom propre sont regroupés sous la catégorie *syntagme nominal*. Les déterminants (article indéfini, possessifs, démonstratifs,…) sont inclus dans les SN si le nom qu'ils introduisent fait partie de la même chaîne. Ainsi, dans l'exemple 17), *son ami Paul* doit être extrait comme SN *Il* et le déterminant *son,* appartenant à la CR *Elle,* est étiqueté *autre*.

Enfin, les autres types de maillons, tels que les sujets zéro ou les déterminants dont le nom ne fait pas partie de la même CR, sont également classés dans la catégorie *autre*. Notre extraction inclut les maillons *autre* afin de pouvoir les quantifier, bien que notre analyse ne porte pas sur cette catégorie.

Pour chaque maillon identifié, le script python permet de repérer le référent et son caractère collectif. Pour notre analyse, nous avons choisi de regrouper les maillons collectifs et non collectifs dans la même chaîne référentielle (*Il, Elle* ou *Les Enfants*) et de préciser si l'expression référentielle est collective (*oui* ou *non*). Ce choix nous permet de suivre l'évolution d'un référent, tout en gardant la possibilité d'analyser les effets de la référence collective, sans multiplier les chaînes de référence.

Le script python permet ensuite d'extraire le niveau scolaire de l'élève, grâce à une expression régulière appliquée aux lignes de commentaire. Cette étape est construite à partir des métadonnées présentes dans la copie, décrites dans la partie 2.1.2. Le corpus RésolCo :

- *cycle 3* pour les élèves de CM1, CM2 et 6e ;
- *cycle 4* pour les élèves de 5e, 4e et 3e ;
- *master* pour les étudiant.es (dans notre corpus uniquement de M2).

Le script permet de calculer la distance (en tokens) qui sépare le maillon extrait du maillon précédent. Concrètement, chaque token du corpus est numéroté de manière séquentielle. Pour chaque paire de maillons consécutifs, on identifie le numéro du premier maillon (ou du premier mot si l'expression référentielle est un syntagme) ainsi que l'index du maillon précédent (ou le dernier mot de l'expression référentielle, le cas échéant). La distance est alors obtenue en soustrayant les deux nombres correspondant aux rangs des expressions référentielles. Afin de compter uniquement les tokens situés entre les expressions, nous excluons les mots des maillons eux-mêmes. Prenons l'exemple suivant :

> 35) Mais les enfants avaient toujours peur et le petit garçon réentendit des petits bruits inqiuétants puis il se retourna en entendant ce grand bruit. [CO-4e-2018-LSPJJRD-D1-R9-V1]

Dans la copie conllu, la numérotation des tokens recommence à chaque phrase. Le script python permet de numéroter les tokens sur l'ensemble de la copie. Cette indexation permet de compter des distances inter-phrastiques. Pour la phrase 35) nous retrouvons la numérotation suivante :

| Mais | les | enfants | avaient | toujours | peur | et | le |
|---|---|---|---|---|---|---|---|
| 66 | 67 | 68 | 69 | 70 | 71 | 72 | 73 |

| petit | garçon | réentendit | des | petits | bruits | inqiuétants | puis | il |
|---|---|---|---|---|---|---|---|---|
| 74 | 75 | 76 | 77 | 78 | 79 | 80 | 81 | 82 |

*Tableau 6: Exemple d'indexation de tokens*

Pour calculer la distance entre les deux premiers maillons, le script doit prendre en compte l'index du dernier token du maillon *les enfants* (68) et l'index du premier token du maillon suivant : *le petit garçon* (73). Nous trouverons alors 73 - 68 = 5 tokens. On soustrait ensuite 1 pour exclure la jonction entre les deux mentions. Autrement dit, on cherche à savoir combien de mots se trouvent entre les deux mentions. Dans notre exemple, il y a bien 4 tokens entre les deux mentions : *avaient toujours peur et* (69,70,71,72). Pour la distance entre *le petit garçon* (75) et *il* (82), le script fonctionne de la même manière : 82 - 75 - 1 = 6 tokens. Nous retrouvons bien 6 tokens : *réentendit des petits bruits inqiyétants puis* (76, 77, 78, 79, 80, 81).

Lorsque le maillon analysé est le premier de la chaîne de référence, aucune distance n'est calculée car il n'existe pas de maillon antérieur à prendre en compte pour cette mesure. Ce cas est signalé par la mention *NA*. Les premiers maillons des chaînes pourront faire l'objet d'une analyse ultérieurement, c'est pourquoi nous faisons le choix de les garder dans nos extractions.

Enfin, afin de pouvoir proposer un travail complémentaire, nous proposons d'extraire le POS du m-1. Pour cet affichage, nous choisissons de nous baser sur les étiquettes POS Stanza uniquement. Nous détaillerons ce travail dans la partie Partie 5. Analyse des maillons précédents (m-1).

Finalement notre script python a pour objectif d'extraire les maillons de chaque référent de la consigne, qu'ils soient collectifs ou non, tout en tenant compte des expressions composées de plusieurs mots. Les maillons faisant partie de plusieurs chaînes apparaissent plusieurs fois. Dans la phrase

36) D'accord on regarde mais après on rentre directement ? [CO-4e-2018-LSPJJRC-D1-R22]

Le pronom *on* est annoté comme maillon des chaînes *ElleColl*, *IlColl* et *LesEnfants.* Il devra donc apparaître trois fois. Il sera une fois dans chaque chaîne et portera l'étiquette *oui (*collectif) pour les CR *Elle* et *Il*.

### 3.1.3. Description de la sortie attendue

A partir des éléments présentés plus haut, nous sommes en mesure de nous attendre à une extraction similaire au tableau 7.

| Copie | Forme | Référent | Collectif | POS | Distance en token | Cycle | POS du m-1 |
|---|---|---|---|---|---|---|---|
| CO-4e-2018-LSPJJRC-D1-R22 | elle | Elle | non | PRON | 4 | 4 | DET |

*Tableau 7: Présentation de la sortie attendue*

La première colonne comporte l'identifiant de la copie. Ensuite, nous pouvons retrouver la forme du maillon, ce qui permet une vérification de la qualité de l'extraction plus aisée. Cette évaluation pourra porter sur la composition du maillon ou l'étiquette qui lui a été attribuée dans la colonne POS. La troisième colonne permet de distinguer les référents (*Elle*, *Il*, *Les enfants).* La quatrième colonne précise si le référent est collectif (oui) ou s'il ne l'est pas (non). La cinquième colonne montre si le maillon est un pronom (PRON), un syntagme nominal (SN) ou *autre*. La septième colonne présente la distance en token entre le maillon et le m-1. La huitième colonne indique le niveau scolaire. Finalement, nous pouvons retrouver le POS du maillon précédent.

## 3.2. Constitution du gold standard

Afin d'être en mesure d'évaluer l'extraction automatique des maillons composant les chaînes de référence sélectionnées, nous avons commencé par annoter manuellement deux copies. Cette étape nous a permis de nous approprier les données et de vérifier la cohérence entre nos observations, les annotations existantes et le guide d'annotation. Nous avons identifié les maillons composant la chaîne de référence *Elle* qui inclut les mentions *ElleColl*. Nous avons travaillé avec une copie du cycle 3 (CO-6e-2016-PJPR1-D1-R9-V1) et une copie du cycle 4 (CO-4e-2018-LSPJJRC-D1-R22). Nous avons choisi de nous concentrer sur des copies de ces cycles dans un souci de cohérence méthodologique et de faisabilité. Les copies de niveau Master servent principalement d'élément de comparaison dans la mesure où nous pouvons supposer que les étudiant.es présentent une maîtrise plus avancée des expressions référentielles. Le gold permet d'évaluer la qualité de nos extractions et d'analyser les erreurs. Dans cet objectif, l'annotation de phénomènes référentiels peut-être moins stabilisés des cycles 3 et 4 nous semble pertinent.

La première étape consiste à repérer les maillons qui nous intéressent pour la suite du travail : les pronoms et les SN faisant partie d'une CR. La seconde étape consiste à compter le nombre de tokens séparant le maillon identifié du maillon m-1, peu importe la forme du m-1.

L'annotation manuelle a permis d'identifier 57 maillons appartenant à la chaîne de référence visée. Le tableau 8 présente l'extraction attendue dans la copie du cycle 3.

| Forme | Référent | Collectif | POS | Distance en token | Cycle |
|---|---|---|---|---|---|
| Une femme | Elle | Non | SN | NA | 3 |
| qui | Elle | Non | PRON | 0 | 3 |
| Emse | Elle | Non | SN | 2 | 3 |
| Elle | Elle | Non | PRON | 1 | 3 |
| Emse | Elle | Non | SN | 29 | 3 |
| elle | Elle | Non | PRON | 7 | 3 |
| Leur mère | Elle | Non | SN | 13 | 3 |

*Tableau 8: Extraction manuelle des maillons de la CR Elle - cycle 3*

Nous pouvons observer 7 maillons pour la chaîne *Elle* dont le POS correspond aux critères retenus pour notre étude. Le tableau 9 indique les maillons identifiés dans la copie du cycle 3 ainsi que dans la copie du cycle 4.

| | Cycle 3 | Cycle 4 | Total |
|---|---|---|---|
| SN | 4 | 18 | 22 |
| PRON | 3 | 32 | 35 |
| Total | 7 | 50 | 57 |

*Tableau 9: Répartition des formes référentielles selon le cycle*

En tout 57 maillons ont été annotés manuellement. On observe que les pronoms sont plus nombreux que les SN dans les deux copies. Contrairement à la copie du cycle 3, des référents collectifs apparaissent dans la copie du cycle 4.

Ces données serviront de base de comparaison pour évaluer les performances de l'extraction automatique, présentée dans la section suivante. Une concordance forte entre notre annotation et l'extraction est attendue afin de valider la robustesse du script Python avant son application à l'ensemble du corpus.

### 3.3. Extraction automatique des chaînes de référence

Dans cette partie, nous commencerons par appliquer le script python aux copies qui constituent le gold. Cette étape nous permet de présenter les résultats obtenus, dans la perspective d'évaluer la fiabilité des scripts développés.

#### 3.3.1. Script python appliqué aux copies annotées : analyse des erreurs

Le script python, présenté dans la section 3.1. Choix méthodologiques, permet d'extraire les maillons et d'en afficher la forme, le pos, le type (collectif ou non), le niveau scolaire de l'élève rédacteur et la distance en token qui sépare le maillon du m-1.

L'extraction des maillons pour les copies annotées nous semble satisfaisante car elle obtient un fscore total de 0,94. Cependant, il peut être intéressant de se pencher sur les erreurs d'extractions. Deux cas de figure nous intéressent : les faux positifs et les faux négatifs.

**Analyse des faux positifs**

Les faux positifs se retrouvent dans l'extraction des maillons de la copie du cycle 3 et dans la copie du cycle 4. Plusieurs cas sont repérés :

- Une erreur de segmentation due à un article partitif : Le premier maillon de la copie de cycle 4 *une fille du nom de Jeanne* est mal interprété par le script python. Ce maillon est extrait sous la forme *nom de Jeanne*. Cette erreur peut être expliquée par la présence de l'article contracté *du*, qui apparaît dans le fichier conllu sous la forme *de + le.* L'article est annoté comme coréférent mais *de* et *le* ne le sont pas. Le script considère donc que *une fille du* est le premier maillon et que *nom de Jeanne* est le second. Ce type d'erreur a 85 occurrences sur l'ensemble des maillons (pronom, SN ou autre) de la ressource.
- Une erreur de segmentation : Lorsque deux maillons d'une chaîne de référence se suivent directement, le script python les compte comme une expression référentielle unique. C'est notamment le cas pour le premier maillon de la copie du cycle 3 : *une femme qui.* L'annotation sépare *une femme* du pronom *qui* mais l'extraction les regroupe et attribue l'étiquette *SN.* Cette erreur d'extraction a un impact sur les calculs de la distance. Cette erreur concerne 2 occurrences dans notre corpus. Il est cependant complexe de quantifier le nombre exact d'occurrences car cette observation résulte d'une analyse manuelle.
- Une annotation ambiguë : Le groupe *prise de panique elle* extrait de la copie du cycle 4 pose un problème d'analyse. Dans le corpus, l'outil Stanza identifie *prise* comme un nom, suivi d'une préposition (*de*) et d'un nom (*panique*), analysant ainsi le groupe comme un syntagme nominal. Ce groupe peut être interprété comme une expression figée. Cependant, *prise de panique* ne devrait pas faire partie de notre extraction car cette expression n'est pas une expression référentielle, mais une construction attributive. Elle attribue la propriété d'être *prise de panique* au référent *Elle*. Cette erreur n'est observable que dans la mesure où nous avons procédé à une analyse manuelle de chaque maillon. Il est donc difficile de quantifier le nombre exact d'occurrences dans notre corpus.
- Une erreur de segmentation et d'annotation. L'expression *Calme-toi Jeanne* est tout d'abord annotée de manière erronée car *Calme-toi* est compté comme un seul token et porte l'étiquette nom propre. Ce token est également annoté intégralement comme maillon de la chaîne de référence quand seul le pronom *toi* serait attendu. De plus, le script python segmente mal les maillons qui se suivent immédiatement, ce qui mène à prendre également en compte le nom propre *Jeanne* dans le même maillon. Cette erreur mène à un mauvais décompte des tokens entre *Jeanne* et le m-1, *toi*.

**Analyse des faux négatifs**

Un cas de faux négatif est également repéré dans la copie du cycle 4. Autrement dit, un maillon que nous avions identifié n'est pas extrait de la copie.

- Une occurrence de *calme-toi* n'est pas extraite car elle est annotée comme un seul token et porte l'étiquette interjection (*INTJ)*. Le script n'extrait pas les interjections, expliquant ainsi l'absence de l'expression référentielle *toi.*

### 3.3.2. Évaluation de la fiabilité de l'extraction

Pour garantir une extraction fiable, nous faisons le choix d'ajouter des contraintes à notre script. Notre objectif est de prendre en compte le repérage les maillons uniques qui suivent directement une expression référentielle afin de mieux segmenter les extractions, comme présenté dans la partie 3.1.1. Méthode d'extraction des maillons. Après ce travail, nous arrivons à extraire le premier maillon sous la forme *une fille nom de Jeanne.* Les maillons *une femme* et *qui*, qui étaient extraits comme un seul maillon sont à présent séparés, ce qui permet de bien évaluer leur distance (0 token).

Finalement, seules les erreurs liées à une annotation erronée persistent. Cependant, seule une observation manuelle et systématique pour l'ensemble du corpus pourrait mener à une quantification précise.

Nous choisissons donc d'appliquer le script en gardant à l'esprit que certaines extractions peuvent être imparfaites.

### 3.4. Analyse des extractions pour l'ensemble du corpus

Le tableau 10 montre le format de l'extraction comprenant les informations nécessaires à nos premières analyses.

| Fichier | Référent | POS | Forme | Collectif | Cycle | Distance |
|---|---|---|---|---|---|---|
| CO-4e-2018-LSPJJRC-D1-R22-V1.conllu | Elle | SN | une fille nom de Jeanne | non | 4 | NA |
| CO-4e-2018-LSPJJRC-D1-R22-V1.conllu | Elle | pronom | elle | non | 4 | 3 |
| CO-4e-2018-LSPJJRC-D1-R22-V1.conllu | Elle | pronom | Elle | non | 4 | 1 |

*Tableau 10: Format de l'extraction effectuée par le script python*

Le script python appliqué à notre corpus permet d'extraire un total de 15 447 expressions référentielles : 7237 pronoms, 5058 SN et 3152 autres maillons. Les maillons *autre* sont principalement des déterminants possessifs et des verbes marquant un sujet zéro. Quelques erreurs d'étiquettes proposées par Stanza peuvent aussi être identifiées. Par exemple :

37) Maïmouna ne voulait pas rentrer mais Katie très curieuse voulait rentrer, comme Maïmouna avait peur elle ne voulait pas rester toute seule Katie dit « allez maïmouna n'aie pas peur » [EC-CM1-2018-SEZB-D1-R12-V1]

Les premières occurrences de *Maïmouna* portent l'étiquette PROPN, nom propre. Cependant, *maïmouna* contenu dans le dialogue n'a pas de classe grammaticale associée (X).

Parmi les expressions référentielles extraites, nous sélectionnons 12 295 mentions qui correspondent à nos critères (pronom ou SN), soit environ 79,6 % des extractions totales.

Cependant, nous avons observé que 408 mentions comportent une valeur négative en tant que distance. Parmi ces mentions, 370 sont des pronoms ou des SN. Cette valeur est observée dans deux cas :

- un oubli de parenthèse fermante dans la colonne misc du fichier conllu. La parenthèse non fermée est mal prise en compte par le script python et le décompte est fait jusqu'au prochain maillon de la CR comportant une parenthèse fermante, créant parfois une boucle. C'est le cas pour l'exemple :

38) Eve et Adriana sont toutes les deux âgées de 9 ans. [CO-6e-2017-VTAC602-D1-R9-V1]

qui est annoté ainsi :

| | |
|---|---|
| Eve | coref=(lesEnfants\|instr= |
| et | coref=lesEnfants\|instr= |
| Adriana | coref=lesEnfants)\|instr= |
| sont | coref=0\|instr=0 |
| toutes | coref=(lesEnfants\|instr= |
| les | coref=lesEnfants\|instr= |
| deux | coref=lesEnfants\|instr= |
| âgées | coref=lesEnfants\|instr= |
| de | coref=lesEnfants\|instr= |
| 9 | coref=0\|instr=0 |
| ans | coref=0\|instr=0 |

*Tableau 11: Exemple d'annotation menant à une erreur : CO-6e-2017-VTAC602-D1-R9-V1*

Nous pouvons voir que la première expression *Eve et Adriana* est correctement annotée. Cependant, l'expression référentielle suivante commence avec *toutes* et inclut les maillons *les deux* âgées de mais les parenthèses ne sont pas fermées et le token suivant n'est pas annoté comme faisant partie de la CR.

- Un mauvais traitement du script python : certains maillons collectifs composés de plusieurs mots mènent à une distance négative. Ce problème survient dans le cas où les annotations sont imbriquées C'est le cas pour l'exemple suivant.

| | |
|---|---|
| les | coref=(ElleColl,(IlColl,(lesEnfants\|instr=1 |
| enfants | coref=ElleColl),IlColl),lesEnfants)\|instr=1 |

*Tableau 12: Exemple d'annotation complexe*

Le script calcule correctement la distance pour la CR *Les Enfants,* mais repère -1 pour les CR *Il* et *Elle.* Malgré nos efforts, nous ne sommes pas parvenu à corriger cette erreur, liée à la gestion des parenthèses imbriquées et à la présence de référents à la fois collectifs et non collectifs. Nos propositions compromettant la bonne identification de maillons composés ou séparant les CR collectives des non collectives, nous choisissons de supprimer les cas problématiques. Conserver ce script nous permet d'identifier les maillons concernés pour un approfondissement éventuel dans le cadre d'un travail ultérieur.

A la lumière de nos observations, nous choisissons de retirer de notre analyse les mentions dont la distance est négative, soit 3,32 % des maillons de la ressource et 3 % des maillons pronom ou SN. Cette décision est liée à une limite de nos compétences techniques et à la nécessité de garantir des résultats fiables. De plus, nous estimons que le corpus reste suffisamment étendu pour l'analyse. Nous proposerons donc un travail d'analyse sur un total de 11926 expressions référentielles. La suite du chapitre présente la répartition quantitative des extractions.

### 3.4.1. Répartition des POS des expressions référentielles, par cycle

Les expressions référentielles sont réparties de la sorte : 7139 sont des pronoms et 4787 sont identifiées comme SN. Le tableau 13 précise la répartition quantitative de chaque classe grammaticale par cycle.

| | Cycle 3 | Cycle 4 | Master | Total |
|---|---|---|---|---|
| Pronom | 3293 | 2477 | 1369 | 7139 |
| SN | 2136 | 1610 | 1041 | 4787 |
| Total | 5429 | 4087 | 2410 | 11926 |

*Tableau 13: Répartition des pronoms et SN selon le niveau*

A première vue, le tableau pourrait suggérer que les POS des expressions référentielles varient grandement selon les niveaux car le cycle 3 contient 1924 pronoms de plus que les étudiant.es de Master 2. Cependant, le graphique 1 permet d'observer la répartition, en pourcentage, des pronoms et des SN par niveau.

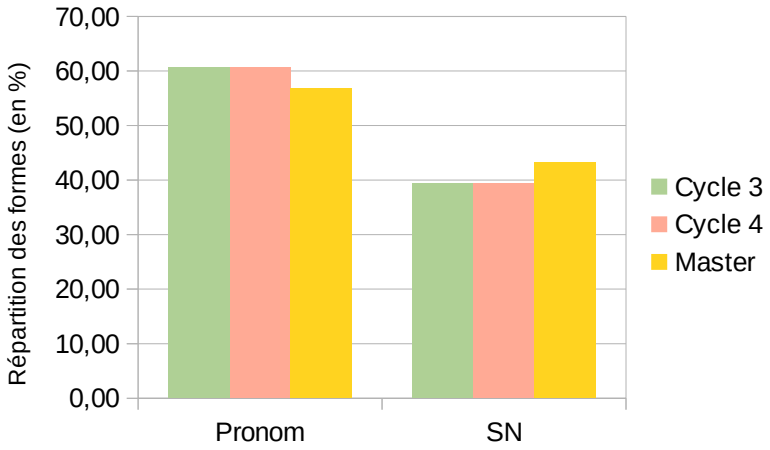



*Figure 1: Répartition des pronoms et SN pour chaque cycle (en %)*

Le diagramme fait ressortir une relative homogénéité dans les usages des formes référentielles. Les trois cycles étudiés présentent une variation relativement faible dans la répartition des POS des expressions référentielles. Les SN représentent 39,94% des expressions référentielles extraites pour le cycle 3. Cette distribution se retrouve pour le cycle 4 puisque 39,39% des expressions extraites sont des SN. Cette classe est légèrement plus représentée pour le Master : 43,20% Les pronoms sont donc majoritairement employés dans l'ensemble des copies du corpus, bien que la distribution entre les deux catégories grammaticales soit plus équilibrée au Master.

### 3.4.2. Répartition des POS des expressions référentielles par référent

Nous avons ensuite choisi d'observer la répartition des mentions selon les référents. Le diagramme 2 présente la répartition en pourcentage des référents *Elle, Il* et *Les Enfants,* par cycle. Cette répartition prend en compte les mentions collectives et non collectives de manière non différenciée. La différence quantitative entre ces deux traits sera présentée dans la partie 3.4.3. Répartition des POS des expressions référentielles selon le trait « Collectif ».

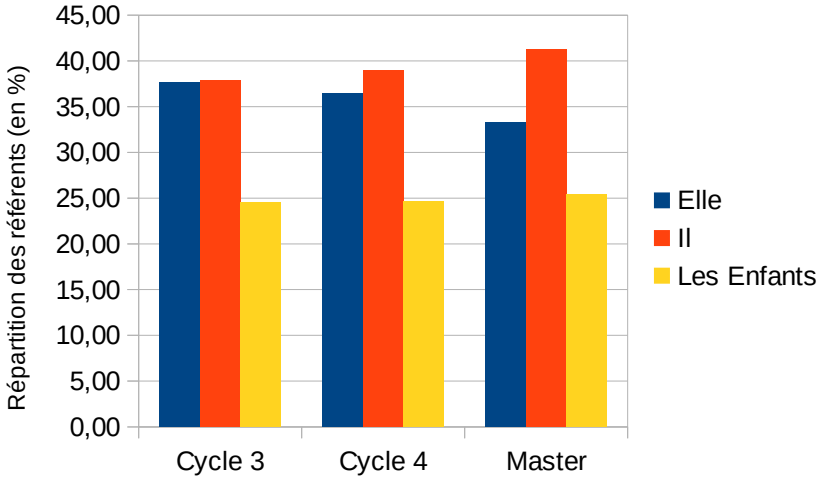



*Figure 2: Répartition des référents par cycle (en %)*

Ce diagramme met en évidence plusieurs points. Tout d'abord, la CR *Les Enfants* est la moins employée dans l'ensemble des niveaux scolaires, avec une fréquence homogène : 24,48 % au cycle 3, 24,64 % au cycle 4 et 25,44 au Master.

Ensuite, les copies du cycle 3 ont une fréquence équivalente entre les référents *Il* (37,83 %) et *Elle* (37,69 %). Au cycle 4, une légère majorité de mentions fait partie de la CR *Il* 38,90 % contre 36,46 % pour la CR *Elle)*. Cette différence est davantage marquée au Master où la CR *Il* représente 41,29 % des reprises.

Le tableau 14 nous permet de porter un regard plus précis sur le nombre de maillons extraits par cycle et par référent.

| | Cycle 3 | Cycle 4 | Master | Total |
|---|---|---|---|---|
| Elle | 2046 | 1490 | 802 | 4338 |
| Il | 2054 | 1590 | 995 | 4639 |
| Les Enfants | 1329 | 1007 | 613 | 2949 |

*Tableau 14: Répartition quantitative des maillons par référent et par cycle*

Le tableau indique que la chaîne de référence *Il* est la plus présente dans les extractions. La CR *Elle* est moins représentée avec 301 maillons d'écart. La CR *Les Enfants* (2949 maillons) représente environ 1,5 fois moins de maillons que les CR *Elle* et *Il* (4338 et 4639).

### 3.4.3. Répartition des POS des expressions référentielles selon le trait « Collectif »

Les mentions peuvent être annotées comme *collectives* lorsqu'elles font partie d'une mention faisant référence à au moins une autre entité que celle du référent principal.

Le tableau 15 précise le nombre d'occurrences de maillons collectifs et non collectifs extraits par cycle.

| | Cycle 3 | Cycle 4 | Master | Total |
|---|---|---|---|---|
| Collectif | 1072 | 1021 | 482 | 2575 |
| Non collectif | 4357 | 3066 | 1928 | 9351 |

*Tableau 15: Répartition des formes selon le trait "Collectif"*

Sur ce tableau nous remarquons une nette majorité de mentions non collectives, et ce pour l'ensemble des niveaux scolaires. Le diagramme 3 présente la répartition, en pourcentage, des expressions référentielles selon le statut collectif ou non collectif du référent et selon les niveaux scolaires.

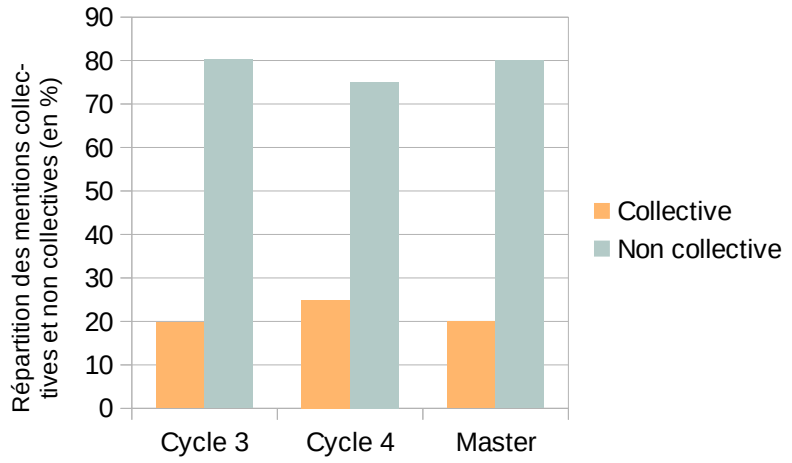



*Figure 3: Répartition des mentions collectives et non collectives selon les niveaux (en %)*

Ce graphique indique que les élèves du cycle 3 emploient des maillons non collectifs dans 80,25 % des cas. Cette distribution se retrouve également au cycle 4 (75,02%) et pour les étudiant.es de Master (80%). Par ailleurs, les expressions collectives (2575 au total) sont environ trois fois moins nombreuses que les expressions non collectives (9351).

### 3.4.4. Répartition des POS pour le premier maillon

Nous avons également observé la répartition des POS des expressions référentielles qui constituent le premier maillon d'une chaîne de référence. Cette étude a permis de repérer 218 pronoms utilisés comme premier maillon contre 937 syntagmes nominaux. La figure 4 présente la répartition, en pourcentage, de chaque POS selon les niveaux scolaires.

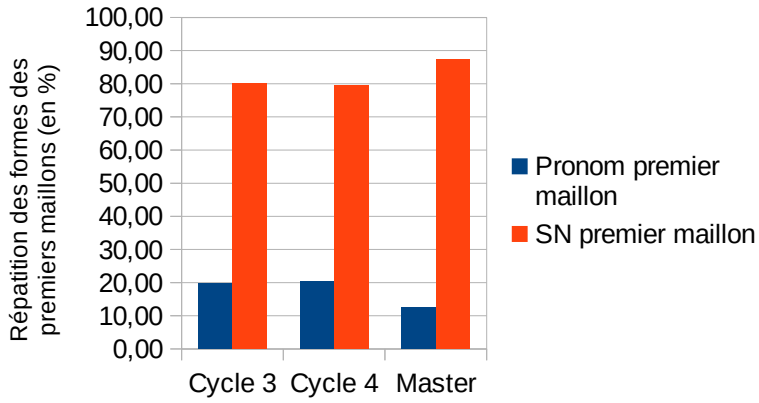



*Figure 4: Répartition des formes des premiers maillons, par niveau (en %)*

Nous pouvons remarquer que les cycles 3 et 4 présentent une répartition similaire : 19,75% des premiers maillons sont des pronoms au cycle 3 contre 20,43% au cycle 4. En revanche, les étudiant.es de master utilisent moins fréquemment le pronom comme premier maillon puisque seuls 12,64% des premiers maillons sont des pronoms pour ce niveau.

Cette répartition peut être l'indice d'une évolution de la maîtrise référentielle. En effet, l'usage de syntagmes nominaux comme premiers maillons peut témoigner de la volonté de faire appel à des expressions référentielles plus explicites, à l'inverse des pronoms qui nécessitent une interprétation basée sur le contexte narratif, comme évoqué dans la partie 1.3.3. La forme du premier maillon de la chaîne. La figure met en lumière une évolution entre le cycle 3 et le Master.

## *Partie 4. Analyse statistique*

### 4.1. Approche statistique

Cette partie vise à vérifier l'existence d'un lien statistique entre la distance entre deux maillons et la forme référentielle. Dans cet objectif, nous procéderons à des analyses statistiques, à l'aide du logiciel Rstudio. Tout d'abord, nous introduirons les variables sur lesquelles l'analyse portera. Ensuite, nous testerons le lien entre distance et forme avant d'affiner l'analyse à l'aide de 3 variables.

#### 4.1.1. Présentation des variables

L'étude de la liaison statistique commence par l'observation des variables. Les deux variables principales sont la distance et le POS de l'expression référentielle (SN ou pronom) :

- La distance est une variable quantitative discrète, ce qui signifie que ses valeurs ne peuvent être qu'entières. Dans notre corpus la distance s'étend de 0 tokens à 666. La moyenne est de 14,29 tokens et la médiane est de 8. Cet écart est un indicateur de valeurs extrêmes et d'une moyenne influencée par des valeurs élevées. De plus, l'écart-type de 23,99 tokens indique une grande variabilité des distances. L'histogramme logarithmique 5 permet d'observer la distribution des valeurs de la variable *distance*.

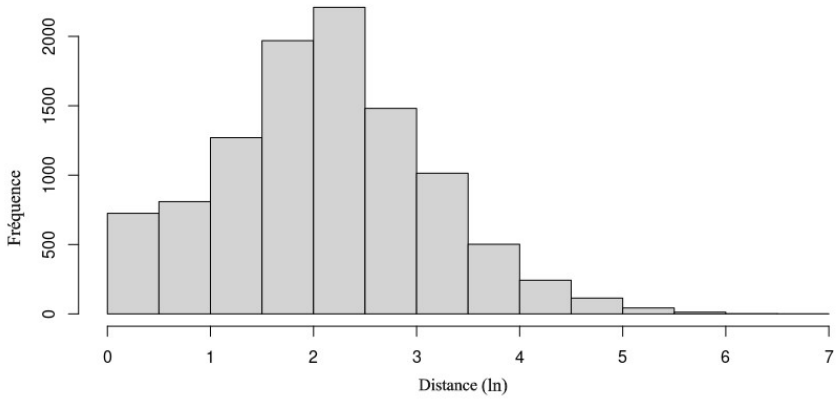



*Figure 5: Répartition des valeurs de la distance*

La transformation logarithmique des distances (ln) permet de visualiser la distribution clairement, contrairement à un histogramme linéaire qui écrase la visualisation des valeurs fréquentes (ici proches de 0). De plus, le logarithme permet de comparer des ordres de grandeurs. Nous pouvons remarquer une asymétrie ainsi qu'un étalement des valeurs à droite. Ces observations et la boîte à moustaches 6 nous permettent d'affirmer que la distribution n'est pas normale : de nombreuses valeurs extrêmes sont repérées (*outliers*) : ce sont les points visibles au-dessus de la boîte.

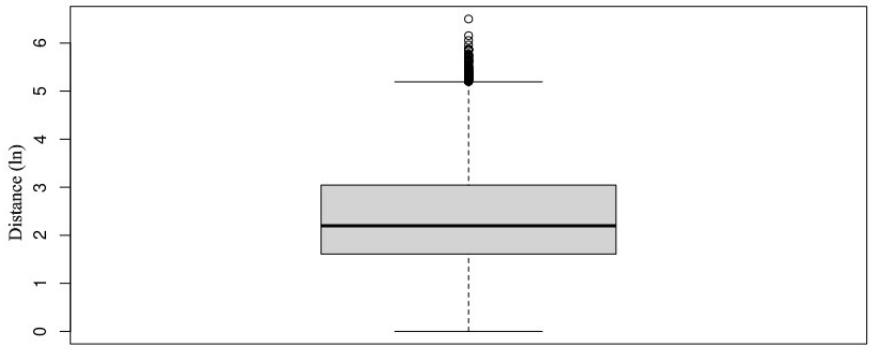



*Figure 6: Boîte à moustaches de la variable distance*

L'axe y représente des intervalles de distance après transformation logarithmique (ln).

- La forme choisie (ici nommée *pos*), est une variable qualitative nominale qui présente deux valeurs : *pron* ou *SN.* En tout, cette variable comporte 7139 pronoms et 4787 SN. La distribution est donc non équitable.

Après l'étude du lien entre la distance et la forme référentielle, d'autres variables peuvent permettre d'observer d'éventuelles variations et seront prises en compte :

- Le niveau scolaire (également nommé *cycle* dans notre travail) est une variable qualitative présentant trois valeurs : *Master*, *3* et *4*. En tout, cette variable comporte 5429 maillons extraits de copies du cycle 3, 4087 maillons du cycle 4 et 2410 de Master. La répartition est donc inéquitable
- Le référent est une variable qualitative nominale qui présente trois valeurs : *Il, Elle* ou *Les Enfants.* En tout, cette variable comporte 4338 références à *Elle,* 4639 références à *Il* et 2949 références *Les Enfants*. La distribution est donc non équitable, notamment pour *Les Enfants* qui est largement sous-représenté.
- La référence collective est une variable qualitative nominale qui présente deux valeurs : *oui* ou *non*. En tout, 9351 mentions sont non collectives et 2575 sont collectives. Cette répartition est déséquilibrée.

Cette présentation nous permet de mieux comprendre nos données et de choisir les méthodes d'analyses statistiques appropriées pour chaque test. En effet, compte tenu des natures des variables, des distributions anormales et des effectifs déséquilibrés, il apparaît nécessaire d'avoir recours à des tests statistiques non paramétriques pour effectuer les analyses comparatives qui suivent.

### 4.1.2. Étude de la liaison statistique distance - POS

Nous commençons par analyser si la forme d'un maillon d'une chaîne de référence varie selon la distance qui le sépare du maillon précédent. Pour proposer une réponse, nous commençons par croiser la variable numérique (la distance) avec la variable qualitative (le pos). Cette étape nous permet d'observer la distance moyenne pour chaque pos. Le graphique 7 illustre ces résultats.

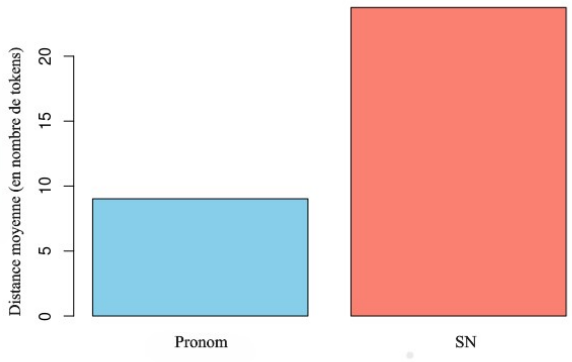



*Figure 7: Distance moyenne par POS (en nombre de tokens)*

Le graphique permet d'observer que la distance est plus de deux fois plus étendue pour les SN (23,75) que pour les pronoms (9,02). Cependant, ces résultats doivent être nuancés car de nombreux *outliers*

sont présents et peuvent influencer la moyenne, comme nous pouvons l'observer sur la boîte à moustaches suivante (figure 8).

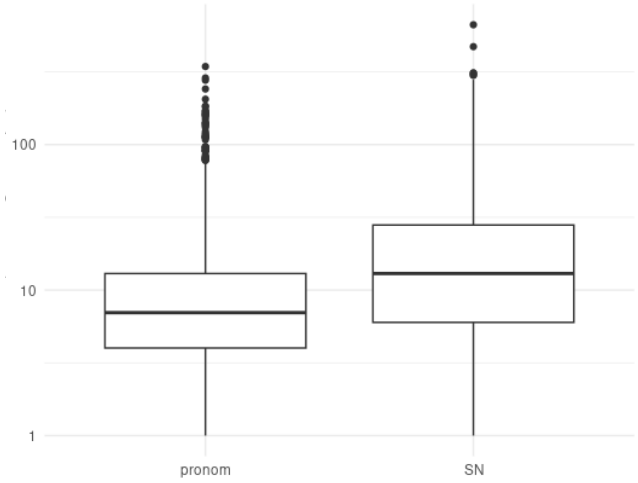



*Figure 8: Distribution logarithmique de la distance par POS*

Pour faciliter la lecture de la boîte à moustaches, nous avons choisi d'avoir recours à une transformation logarithmique des valeurs de la distance. Les distances indiquées sont toutefois bien les valeurs réelles (avec la valeur extrême de 666 tokens). Cette boîte à moustaches montre que les médianes sont plus élevées pour les individus qui ont une étiquette *SN* que pour ceux qui ont une étiquette pronom. Cependant, les points situés en haut des moustaches du graphique représentent des individus dont les valeurs sont extrêmes. Lorsque l'on observe certains cas de référence très éloignée, nous pouvons repérer

- Des reprises dans un dialogue, comme c'est le cas du pronom *on* dans la phrase 36). Ce pronom est employé après 287 tokens sans référence à la chaîne *Les Enfants*. Cependant, le référent est accessible grâce au contexte : *Les Enfants* correspond à *Elle* et *Il* qui dialoguent.
- Une erreur dans l'annotation, similaire aux cas générant des boucles : une parenthèse mal fermée ne permet pas un décompte correct.
- Un usage particulier de SN. En effet, la valeur la plus élevée des SN correspond à une reprise du référent *Les Enfants*. Dans cette copie de Master, le SN défini *les enfants* est employé de manière générique et ne désigne pas des personnages récurrents dans le récit. La copie évoque *les enfants* au début et à la fin de son récit, par les phrases suivantes qui sont respectivement les phrases 5 et 40 du récit et sont séparées par 666 tokens.

39) L'histoire que je vais vous raconter se passe il y a de nombreuses années, quand les enfants pouvaient encore sortir la nuit. [UN-M2-2021-UCL-D1-R47-V1]

40) Depuis cette aventure, les enfants ne sortent plus la nuit. [UN-M2-2021-UCL-D1-R47-V1]

Les cas extrêmes nous conduisent à nuancer nos propos. Il est donc important de vérifier nos observations en effectuant un test statistique.

Pour vérifier que le lien entre la distance et le POS soit statistiquement significatif, nous choisissons le test t de Wilcoxon-Mann-Whitney. Ce test est approprié car il est adapté aux valeurs qualitatives binaires et aux valeurs quantitatives dont la distribution est non normale. La valeur-p permet de vérifier dans quelle mesure il est possible de généraliser nos résultats. Sa valeur-seuil est fixée à 0,05.

Puisque la valeur retournée par le test de Wilcoxon est inférieure au seuil ($p < 2.2e\text{-}16$), il est donc possible d'affirmer que la liaison est significative. Le pronom est séparé du maillon précédent par une

distance plus courte que celle séparant le syntagme nominal du maillon qui le précède. Cette observation va dans le sens des théories présentées dans la partie 1.3. Les théories cognitives, car la nature même du pronom suppose un référent actif. Au contraire, le SN peut être employé de manière plus autonome ou s'inscrire dans le cadre de la redénomination et donc permettre de clarifier la référence dans un contexte plus éloigné ( 1.3.4. Le principe de redénomination). L'exemple 41) tiré d'une copie du cycle 3 illustre un usage illustrant nos observations.

41) Un soir de fête dans une maison bourgeoise du sud de la France, la propriétaire de la maison s'appelait Alice, elle voulait rassembler ses amis chez elle pour fêter sa première place à un concours de mode...Elle habitait dans cette maison depuis longtemps. Quand les premiers invités arrivèrent, le volume de la musique augmentait de plus en plus. Alice baissa le volume quelques minutes après que tous les invités soient présents. [CO-6e-2016-PJPR1-D1-R22-V1]

La chaîne de référence *Elle* est composée de 6 maillons correspondant à notre étude : *la propriétaire de la maison*, *Alice*, *elle, elle, Elle, Alice.* Nous n'affichons pas les maillons *ses* et *sa*. Le tableau suivant est l'analyse générée par notre script.

| | | | |
|---|---|---|---|
| SN | la propriétaire de la maison | non | NA |
| SN | Alice | non | 2 |
| pronom | elle | non | 1 |
| pronom | elle | non | 2 |
| pronom | Elle | non | 8 |
| SN | Alice | non | 24 |

*Tableau 16: Résumé de la chaîne Elle dans un extrait de la copie CO-6e-2016-PJPR1-D1-R22-V1*

Nous pouvons voir que les pronoms présentent des distances relativement courtes (1, 2 et 8 tokens) tandis que les SN ont des distances plus variables et étendues (2 et 24). Le premier nom propre employé permet de préciser le SN défini *la propriétaire*. Par la suite, les pronoms indiquent un référent accessible : proche, introduit par un SN et un nom propre, en position sujet. Le second nom propre semble permettre d'éviter une référence ambiguë après l'introduction de nouveaux personnages, *les premiers invités*.

## 4.2. Analyse des sous-groupes

Afin de poursuivre le travail, nous souhaitons observer la liaison entre distance et POS en fonction des 3 variables présentées plus haut. Pour cela, nous allons constituer des sous-groupes afin de tester si la liaison est similaire et significative pour l'ensemble des groupes. Trois variables seront observées : le niveau scolaire, le référent et le statut de mention collective.

### 4.2.1. La liaison distance - POS selon le niveau scolaire

Pour commencer, nous choisissons d'observer la liaison selon le niveau scolaire. Le tableau 17 présente la moyenne de la distance séparant un maillon de celui qui le précède, par catégorie grammaticale des expressions référentielles et par niveau scolaire.

| Niveau | Forme | Distance moyenne (en tokens) |
|---|---|---|
| Cycle 3 | Pronom | 8 |
| | SN | 16,83 |
| Cycle 4 | Pronom | 9,42 |
| | SN | 24,01 |
| Master | Pronom | 10,69 |
| | SN | 35,88 |

*Tableau 17: Moyenne de la distance par forme et par niveau scolaire*

Il apparaît que la distance séparant le pronom du maillon précédent est plus réduite que la distance séparant le SN du maillon précédent pour l'ensemble des niveaux scolaires. L'étude du tableau permet de mettre en lumière trois points :

- La distance moyenne des pronoms augmente de 2,69 tokens entre le cycle 3 (8) et le Master (10,69).
- La distance moyenne des SN augmente de manière plus marquée, puisqu'elle augmente de 19,05 tokens (16,83 au cycle 3 contre 35,88 pour le Master).
- La différence entre pronom et SN s'accentue avec le niveau scolaire. L'écart entre SN et pronom passe de 8,83 tokens au cycle 3 (16,83 – 8) à 25,19 tokens au Master (35,88 – 10,69).

Nous pouvons conclure que la distance moyenne augmente avec le niveau scolaire et ce, pour chaque POS. Cette progression est particulièrement marquée pour les SN.

Nous choisissons de représenter ces valeurs dans la boîte à moustaches 9 afin d'observer les médianes et la présence d'éventuelles valeurs extrêmes pouvant influencer les valeurs des moyennes. Les boîtes représentent les niveaux scolaires : rouge pour le cycle 3, vert pour le cycle 4 et bleu pour le Master.

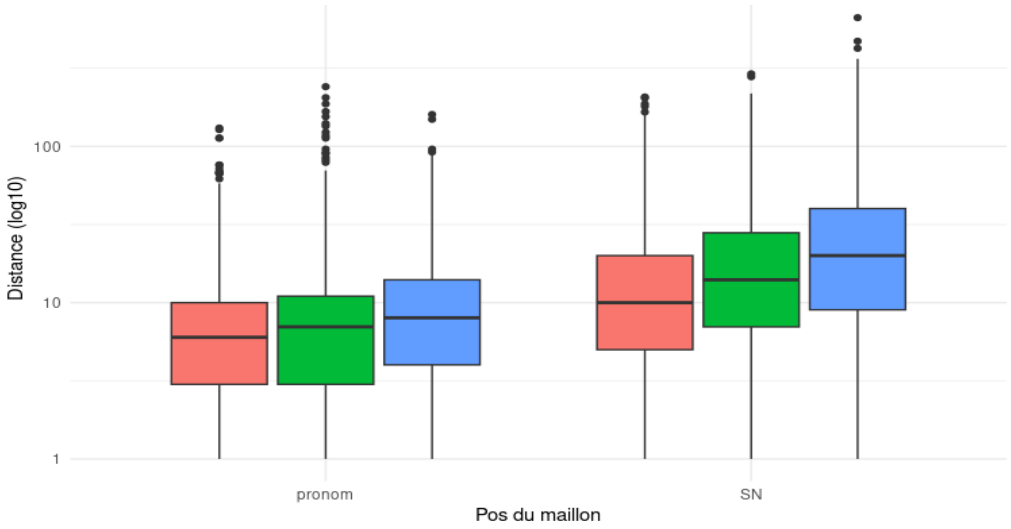



*Figure 9: Boîte à moustaches présentant la distance selon la forme par niveau scolaire*

On remarque que la distribution de la distance séparant les pronoms du m-1 est relativement homogène pour tous les niveaux scolaires observés puisque les valeurs des médianes sont rapprochées pour les 3 niveaux, bien que légèrement en hausse. La boîte à moustaches montre bien que de nombreux *outliers* existent pour les pronoms, notamment au cycle 4.

La distribution de la distance séparant les SN du m-1 semble évoluer de manière plus marquée et régulière. Les étudiant.es de Master (bleu) semblent avoir une distance médiane bien plus élevée que les élèves de cycle 3 (rouge) et de cycle 4 (vert). Les trois niveaux présentent des valeurs extrêmes. Nous pouvons remarquer que les boîtes des SN sont un peu plus étendues que celles des pronoms, témoignant d'écarts à la médiane légèrement plus marqués pour les SN.

Sur les deux représentations, nous pouvons observer que la distance augmente selon le niveau scolaire. En particulier pour les syntagmes nominaux, pour lesquels la distance évolue progressivement du cycle 3 au Master.

Afin de vérifier que nos observations puissent être généralisées, nous choisissons d'effectuer le test t de Wilcoxon-Mann-Whitney dans chaque sous-groupe. Pour faire cela, nous avons créé 3 groupes : les valeurs de distance selon le POS pour le cycle 3, les valeurs pour le cycle 4 et pour le Master. Pour chaque groupe, nous avons repris le même test. Nous avons pu observer que les p-values étaient systématiquement en dessous du seuil de 0,05 ($p < 2.2e-16$). Ces résultats nous permettent de dire que le lien est significatif.

Nous pouvons conclure que la distance est plus courte entre un pronom et la mention précédente qu'entre un SN et la mention précédente pour l'ensemble des niveaux. Pour finir, il apparaît que le niveau scolaire influence la distance moyenne entre deux maillons d'une même chaîne de référence car nous observons une hausse de la distance moyenne et médiane entre chaque cycle et en particulier pour les SN.

### 4.2.2. La liaison distance - POS selon le référent

Cette seconde partie se concentre sur la liaison entre la distance et le POS, selon le référent (*Il*, *Elle*, *Les Enfants*). Le tableau 18 présente la distance moyenne entre le maillon étudié et le précédent, par forme d'expression référentielle et par référent.

| Référent | Forme | Distance moyenne (en tokens) |
|---|---|---|
| Elle | Pronom | 8,28 |
| | SN | 27,5 |
| Il | Pronom | 8,82 |
| | SN | 19,43 |
| Les Enfants | Pronom | 10,79 |
| | SN | 24,6 |

*Tableau 18: Moyenne de la distance, par forme et par référent*

Sur ce tableau, il apparaît que la distance moyenne entre un maillon et le m-1 est toujours plus faible pour les pronoms que pour les SN, quel que soit le référent. En particulier pour la CR du référent *Elle*, les pronoms sont séparés du m-1 par 8,28 tokens en moyenne, tandis que les SN sont séparés du m-1 par 27,5 tokens en moyenne. Cet écart est le plus marqué (19,22 token de différence). Une tendance similaire est observée pour la CR du référent *Il* : 8,82 tokens pour les pronoms contre 19,43 pour les SN (10,61 tokens d'écart). Cette différence est également observée dans la CR du référent *Les Enfants* : 10,79 tokens pour les pronoms contre 24,6 pour les SN (13,81 tokens d'écart).

De plus, le référent semble avoir un lien avec la distance moyenne puisque des variations sont observées pour chaque POS selon les référents :

- Pour les pronoms, 2,51 tokens d'écart sont remarqués entre la CR *Elle* et la CR *Les Enfants,* qui présentent les valeurs les plus éloignées.
- Pour les SN, 8,07 tokens d'écart sont observés entre les CR *Il* et *Elle.*

Bien que légères, les différences peuvent indiquer une influence du référent dans les choix référentiels, de manière plus marquée pour les SN.

La boîte à moustaches 10 permet de représenter cette répartition et d'observer la tendance générale plus aisément.

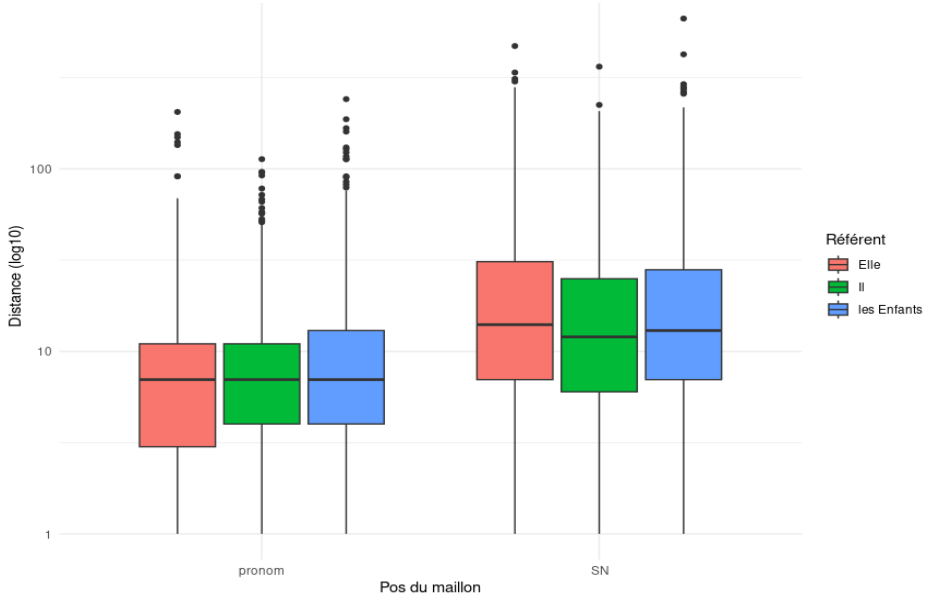



*Figure 10: Distribution de la distance selon le pos, par référent*

Nous pouvons observer que la médiane de la distance entre un pronom et le m-1 est systématiquement plus basse que celle observée entre SN et le m-1, pour l'ensemble des référents. Les valeurs des pronoms sont très proches pour l'ensemble des référents. Pour les SN, la CR *Elle* a les valeurs les plus élevées. De plus, la dispersion autour de la médiane est généralement plus forte pour les SN, ce qui suggère une plus grande variabilité de la distance. Cette variabilité est observable également pour les pronoms de la CR *Elle.*

La présence d'un nombre important de valeurs extrêmes, notamment pour les pronoms des trois référents, nous pousse à nuancer les observations faites sur les moyennes. De même, plusieurs valeurs particulièrement élevées influencent la distance à l'intérieur de la CR *Les Enfants*. Pour expliquer ces valeurs, nous avons vu l'emploi générique d'un SN défini désignant *les enfants* en tant que classe d'individus (40)). Un autre cas a retenu notre attention :

42) Charles était un grand-père très apprécié de **ses petits-enfants** mais très sévère. Un des trois garçons finit par expliquer qu'ils s'étaient donnés rendez-vous à minuit précise en bas de la plage avec des filles de la villa voisine. Pour impressionner celles-ci, ils avaient eu l'idée de faire un feu d'artifice sur la plage. Malheureusement, ils n'arrivèrent jamais à faire fonctionner le feu d'artifice qui explosa finalement quelques minutes occasionnant ainsi les cris des enfants qui avaient oublié le feu d'artifice et qui furent donc effrayés par cette explosion soudaine. Charles soupira, plus de peur que de mal. Il ramena les enfants à la maison, les força à s'excuser auprès de leur grand-mère qui était terrifiée à l'idée que ces petits-enfants adorés avaient pu se retrouver seuls à cette heure de la nuit. Depuis cette aventure, **les enfants** ne sortent plus la nuit. Cela servit de leçon aux plus jeunes des petits-enfants qui n'avaient jamais vu leur grand-mère aussi paniquée que cette nuit-là. [UN-M2-2021-UCL-D1-R46-V1]

La chaîne de référence *Les Enfants* comprend les maillons désignant l'ensemble des petits-enfants. Dans l'exemple 42) seules deux expressions sont concernées : *ses petits-enfants*, *les enfants*. Ces deux expressions sont les seules dans cet extrait qui réfèrent à tous les *petits-enfants* et qui n'incluent pas les *filles de la villa*

*voisine. Ces petits-enfants adorés* désigne uniquement les personnages dont le récit est raconté et non à l'intégralité des *petits-enfants*.

Cette chaîne est étendue car *Les Enfants* est à nouveau une généralisation et permet de référer à des personnages qui ne font pas partie, à proprement parler, du récit. Cet emploi montre une compétence experte de la référence. La continuité référentielle ne repose pas uniquement sur la présence nécessaire des référents dans le récit mais également sur la représentation mentale du groupe auquel ils appartiennent.

Pour dépasser la simple observation, nous choisissons de recourir au test t de Wilcoxon-Mann-Whitney. Nous souhaitons vérifier la significativité de chaque valeur, c'est pourquoi nous avons effectué le test séparément pour chaque référent. Les p-value retournées sont toutes inférieures au seuil de 0,05 ($p < 2.2e-16$) ce qui signifie que les différences observées entre les distances des pronoms et les syntagmes nominaux sont statistiquement significatives pour chaque référent.

### 4.2.3. La liaison distance - POS selon la mention collective

Dans cette partie, nous étudions le lien entre distance et catégorie grammaticale d'une expression référentielle selon le type de maillon (collectif ou non). Le tableau 19 présente les distances moyennes entre un maillon et son m-1 selon la forme référentielle (pronom ou SN) et le type de maillon (collectif ou non collectif).

| Type de maillon | Forme | Distance moyenne (en tokens) |
|---|---|---|
| Non collectif | Pronom | 8,77 |
| | SN | 24,41 |
| Collectif | Pronom | 9,75 |
| | SN | 20,51 |

*Tableau 19: Distance moyenne par forme et type de maillon*

Ce tableau montre que les pronoms ont des distances moyennes plus réduites que celles observées pour les SN, peu importe le type de maillon. Pour les expressions non collectives, les pronoms ont une distance moyenne de 8,77 contre 24,41 pour les SN. Concernant les expressions collectives, les pronoms ont une distance moyenne de 9,75 contre 20,51 pour les SN. Nous pouvons également constater que la distance entre les maillons pronoms non collectifs et les maillons pronoms collectifs évolue très peu (0,98 tokens). La distance des SN varie davantage, bien que la différence reste relativement faible (3,9 tokens). Cette différence pourrait suggérer que le caractère collectif d'un maillon réduit légèrement la distance moyenne entre le maillon SN et le m-1, mais que son influence est moindre.

La figure 11 illustre la distribution des distances pour chaque type de maillon.

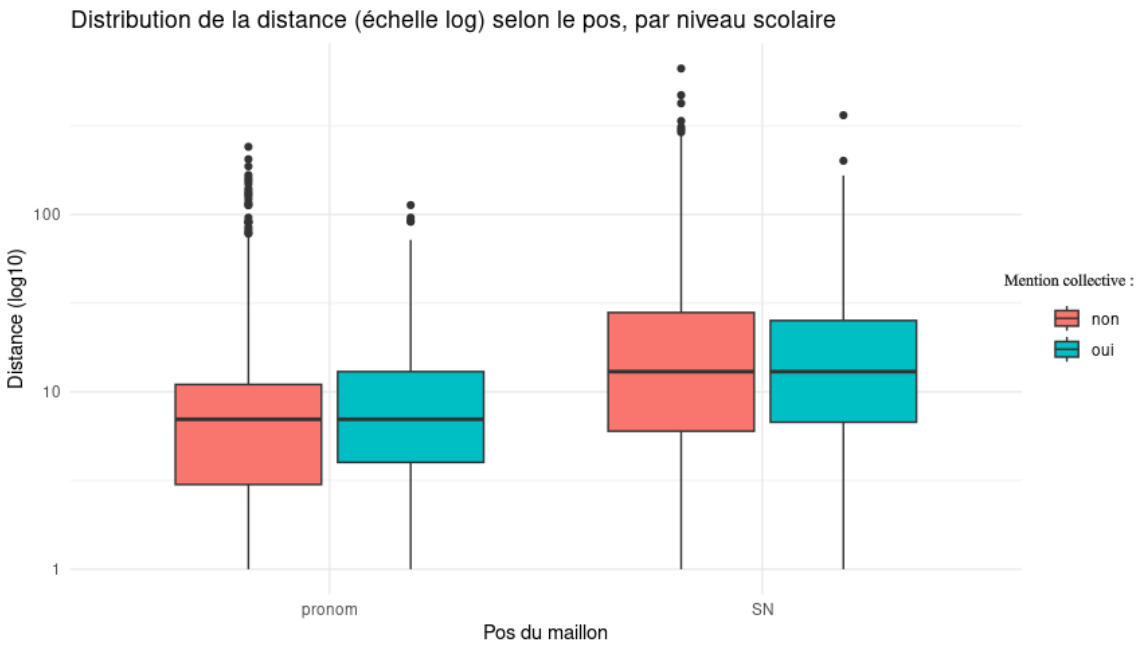



*Figure 11: Distribution de la distance selon les formes référentielles et le type de maillon*

La tendance observée dans le tableau est confirmée par la boîte à moustaches : les médianes des distances sont toujours plus élevées pour les SN que celles des pronoms. De plus, nous pouvons remarquer que la variabilité de la distance pour les pronoms et les SN collectifs est moins grande que pour les autres maillons, car leurs boîtes sont plus resserrées. Les points visibles au dessus des boîtes sont des valeurs extrêmes qui témoignent de cas atypiques qui permettent de nuancer la régularité observée. Les mentions collectives semblent présenter moins de valeurs atypiques que les maillons collectifs.

La médiane des mentions collectives est relativement proche de celle des mentions non collectives. Autrement dit, les pronoms collectifs ou non collectifs ont des médianes très proches et les SN collectifs et non collectifs ont des médianes identiques. Cette observation pourrait suggérer que le statut collectif est peu lié à la distance entre le maillon et son m-1 mais peut avoir un lien avec la variabilité des distances.

Afin de vérifier nos observations, nous utilisons le test t de Wilcoxon-Mann-Whitney pour les deux types de maillons séparément. Les deux valeurs retournées sont inférieures au seuil de 0,05 ($p < 2.2e\text{-}16$ ) ce qui nous permet d'affirmer que le lien observé entre distance, POS du maillon et type de maillon (collectif ou non) est statistiquement significatif.

### 4.3. Analyse croisée des variables

La partie précédente proposait d'observer si la distance influence le choix de la forme de l'expression référentielle. Pour cela, nous avons d'abord étudié séparément les liens entre la distance et le POS d'un maillon. Ensuite, nous avons examiné les effets de différentes variables sur cette liaison, une à une. Dans cette partie, nous souhaitons présenter une synthèse des liens pour l'ensemble des variables.

Afin de visualiser l'ensemble des tendances présentes dans le corpus, nous avons produit un tableau récapitulatif des distances moyennes selon les 4 variables qualitatives présentées plus haut : le POS, le niveau scolaire, le référent et le type de maillon. Le tableau 20 montre la moyenne calculée pour chaque regroupement. Nous avons choisi de trier les moyennes dans l'ordre croissant, afin de pouvoir observer les regroupements qui présentent la distance moyenne la plus basse ou la plus élevée plus aisément. Ce classement permet également d'observer les moyennes proches et de mettre en lumière des caractéristiques qui sont liés à une distance similaire.

| POS | Collectif | Référent | Cycle | Moyenne de distance |
|---|---|---|---|---|
| pronom | non | Elle | 3 | 6,75 |
| pronom | non | Il | 3 | 7,01 |
| pronom | non | Elle | 4 | 8,29 |
| pronom | non | Il | 4 | 8,36 |
| pronom | oui | Il | 3 | 8,54 |
| pronom | oui | les Enfants | 3 | 8,84 |
| pronom | oui | Elle | 3 | 8,89 |
| pronom | oui | Elle | 4 | 10,06 |
| pronom | oui | Il | 4 | 10,2 |
| pronom | non | les Enfants | Master | 10,23 |
| pronom | non | Elle | Master | 10,29 |
| pronom | oui | Elle | Master | 10,54 |
| pronom | non | les Enfants | 3 | 10,66 |
| pronom | non | Il | Master | 10,8 |
| pronom | oui | les Enfants | 4 | 10,85 |
| pronom | non | les Enfants | 4 | 11,35 |
| pronom | oui | Il | Master | 11,72 |
| SN | non | Il | 3 | 14,08 |
| pronom | oui | les Enfants | Master | 14,36 |
| SN | oui | Il | 3 | 14,37 |
| SN | oui | Elle | 3 | 15,75 |
| SN | non | les Enfants | 3 | 17,01 |
| SN | oui | les Enfants | 3 | 18,44 |
| SN | oui | Elle | 4 | 18,86 |
| SN | oui | Il | 4 | 19,46 |
| SN | non | Elle | 3 | 19,71 |
| SN | non | Il | 4 | 20,04 |
| SN | non | les Enfants | 4 | 24,9 |
| SN | oui | Elle | Master | 26,04 |
| SN | oui | les Enfants | 4 | 26,23 |
| SN | non | Il | Master | 27,12 |
| SN | oui | Il | Master | 27,52 |
| SN | non | Elle | 4 | 29,51 |
| SN | non | les Enfants | Master | 38,59 |
| SN | oui | les Enfants | Master | 44 |
| SN | non | Elle | Master | 46,28 |

*Tableau 20: Distance moyenne pour chaque regroupement*

Nous pouvons remarquer que les pronoms non collectifs du cycle 3 présentent les distances moyennes les plus courtes (6,75 pour *Elle* et 7,01 pour *Il*). Cette distance augmente légèrement au cycle 4 (8,29 pour *Elle* et 8,36 pour *Il*), ce qui suggère une relative stabilité. Au contraire, ce sont les SN qui présentent les valeurs les plus élevées, en particulier au Master : les SN collectifs de la CR *Les Enfants* (44) et les SN non collectif de la CR *Elle (*46,28*)* sont les expressions référentielles qui présentent la distance la plus étendue et qui représentent le double des valeurs des mêmes maillons pour le cycle 4.

Ce tableau met en lumière des éléments essentiels pour notre étude mais sa lecture est complexe en raison du nombre de variables croisées. C'est pourquoi le graphique 12 illustre les liens entre le POS d'une mention (pronom ou SN), le référent (*Elle, Il* ou *Les Enfants*), le type du maillon (collectif ou non), le niveau scolaire et la distance entre un maillon et le m-1. Ce graphique est mis à l'échelle logarithmique, ce qui

explique quelques différences entre les moyennes du tableau et la place du point sur le graphique. Cette mise à l'échelle est utile pour représenter toutes les distances sans écraser les données.

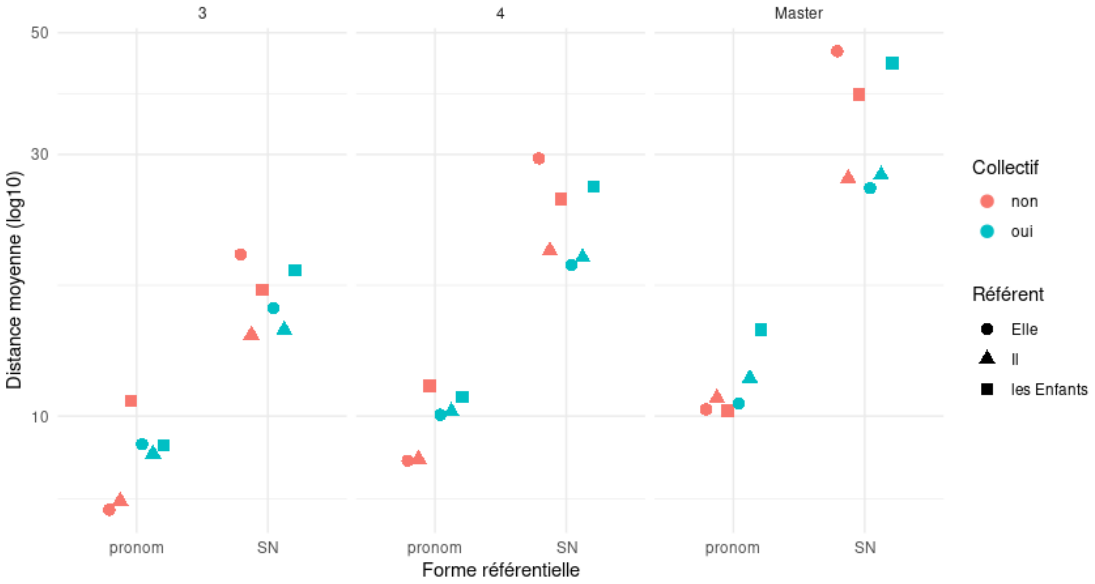



*Figure 12: Distances moyennes par forme, type de maillon, référent et niveau scolaire*

Ce graphique permet d'observer la distance moyenne entre un maillon et le m-1 selon les variables présentées plus haut. Chaque point comporte différentes caractéristiques :

- la forme indique le référent : rond pour *Elle*, triangle pour *Il* et carré pour *les Enfants* ;
- la couleur précise le type de maillon : rouge pour les maillons non collectifs, bleu pour les collectifs ;
- l'axe des abscisses présente les niveaux scolaires et les POS des maillon : pronoms à gauche, SN à droite de l'axe correspondant au niveau ;
- l'axe des ordonnées présente la distance moyenne sur une échelle logarithmique.

Ces caractéristiques autorisent une représentation croisée de l'ensemble des variables et mettent en évidence les liens possibles entre différents facteurs. L'analyse de ce graphique souligne plusieurs tendances, en accord avec les observations formulées dans les parties précédentes.

De manière globale, les syntagmes nominaux sont plus espacés de leur m-1 que les pronoms pour l'ensemble des cycles, des référents et des types de maillons.

L'analyse de la distance entre un maillon et le m-1 montre une évolution au fil de la scolarité. Les distances augmentent avec le niveau pour la grande majorité des combinaisons. Cependant, cette progression n'est pas parfaitement régulière, notamment pour les pronoms non collectifs, qui restent relativement stables. Les pronoms non collectifs de la CR *Les Enfants* voient même leur distance se réduire au Master. En revanche, le niveau Master montre les écarts les plus marqués entre les différentes combinaisons de variables. Nous pouvons remarquer des séparations nettes entre les pronoms et les SN.

L'analyse par référent et caractère collectif apporte un éclairage complémentaire. Lorsque l'on analyse les distances selon les référents, il apparaît que les CR *Elle* et *Il*, présentent des distances moyennes relativement similaires pour les mentions pronoms non collectifs et collectifs, pour tous les niveaux. Les distances moyennes des pronoms collectifs du référent *Les Enfants* se distinguent de celles des deux autres référents aux cycles 3 et 4. Au Master, la distance moyenne des trois référents non collectifs se rapproche et les pronoms collectifs varient selon les référents : les pronoms collectifs *Elle* apparaît dans les mêmes contextes que les maillons non collectifs ; la CR *Il* est plus éloignée et la CR *Les Enfants* davantage encore.

Les tendances semblent donc s'inverser au Master : des distances proches pour les pronoms non collectifs des trois référents et des variations selon les référents pour les pronoms collectifs.

Concernant les SN non collectifs, leur distance progresse tout au long de la scolarité et les tendances restent identiques : la CR *Elle* présente les valeurs les plus élevées, suivie par la CR *Les Enfants* puis la CR *Il*. Les différences entre référents sont davantage marquées au Master. Les SN collectifs présentent également une tendance se confirmant au fil de la scolarité : les distances des SN *Elle* et *Il* ont des valeurs proches et plus basses que celles des SN *Les Enfants*.

Le caractère collectif peut être lié à la répartition des distances pour chaque POS. Au cycle 3 et au cycle 4, le caractère collectif marque un emploi particulier pour les pronoms. Nous pouvons voir que la distance est inférieure pour les pronoms non collectifs, sauf pour *Les Enfants* non collectifs. Cette distinction est moins marquée au niveau Master. Pour les syntagmes nominaux, les différences sont moins présentes.

Pour résumer, l'analyse du graphique 12 montre que la maîtrise de la référence se complexifie avec le niveau scolaire et repose sur la gestion de plusieurs variables : POS, distance, caractère collectif et référent.

### 4.4. Analyse des premiers résultats

Ce travail a mis en lumière le lien entre distance et forme d'un maillon grâce à plusieurs représentations et tests statistiques. Nous avons observé que, de manière générale, la distance moyenne est environ deux fois plus étendue entre un syntagme nominal et le maillon le précédant que celle observée pour un pronom. Nos observations s'inscrivent dans la continuité des théories cognitives qui affirment que plus les maillons sont rapprochés, plus le référent est accessible. Un syntagme nominal plus autonome dans la référence peut donc être employé à une distance plus étendue qu'un pronom qui a besoin d'un référent accessible et donc plus proche, bien que le critère de la distance n'est pas le seul critère permettant d'évaluer l'accessibilité d'un référent. Nos observations vont dans le sens d'Ariel (1990) qui a démontré que le pronom est davantage utilisé lorsque son antécédent est accessible.

Nous avons ensuite affiné l'analyse selon 3 variables distinctes : le niveau scolaire, le référent et son caractère collectif. Nous pouvons d'abord préciser que la distance moyenne et la médiane évoluent au fil de la scolarité, ce qui peut indiquer une complexification des compétences. Cette observation est visible en particulier pour les syntagmes nominaux, ce qui pourrait suggérer que la maîtrise de la reprise nominale s'acquiert plus tardivement. L'augmentation progressive de la distance peut être l'indicateur de l'évolution des compétences référentielles et notamment la capacité à gérer des reprises référentielles distantes.

Lorsque l'on observe la distance selon le référent, nous pouvons remarquer que les pronoms sont généralement employés dans des contextes plus rapprochés, quel que soit le référent. L'écart de distance entre pronom et SN est le plus marqué dans la CR *Elle*. Des variations existent entre référents pour chaque catégorie grammaticale, en particulier pour les SN des CR *Elle* et *Il.* Ces observations suggèrent que le choix de la forme référentielle et à la distance est lié au référent.

Le caractère collectif d'un maillon semble être faiblement lié à la distance et au choix de la forme référentielle. La distance moyenne des pronoms collectifs et non collectifs varie très peu. En revanche, pour les SN, une légère baisse de la distance moyenne peut être observée dans le cas d'un maillon collectif mais les médianes des SN collectifs et non collectifs restent très proches. Les distances médianes des SN sont plus élevées que celles des pronoms, tandis que les expressions collectives montrent une variabilité moindre et moins de valeurs extrêmes. La variabilité moins grande des valeurs des mentions collectives pourrait indiquer que les référents sont employés de manière plus homogène. Ces observations indiquent un impact limité du caractère collectif d'un maillon sur la gestion référentielle. L'analyse croisée des variables suggère cependant que l'effet du caractère collectif varie selon le niveau scolaire et le référent.

Nos analyses nous ont permis de mettre en évidence les liens entre la distance, le référent, le type de maillon et la forme de l'expression référentielle. Les résultats montrent une évolution de la gestion référentielle et suggèrent que les élèves développent des stratégies de reprise de plus en plus complexes. Les résultats confirment donc le lien entre distance et forme référentielle et montrent que les contraintes évoluent selon le référent, son type et le niveau scolaire de manière plus ou moins marquée.

## *Partie 5. Analyse des maillons précédents (m-1)*

Notre étude porte sur le lien entre la distance qui sépare deux maillons d'une chaîne référentielle et la forme de l'expression référentielle qui est employée. Nous avons souhaité analyser si la forme d'un maillon varie selon la distance qui le sépare du maillon précédent et nous avons caractérisé les variations observées selon 3 critères. Dans cette partie, nous proposons un autre regard en nous intéressant également au POS du maillon précédent.

### 5.1. Extraction du POS du m-1 dans un échantillon

Nous élargissons notre travail en incluant à présent la forme du m-1 comme variable. Cette ouverture vise à observer si certains enchaînements de maillons sont plus fréquents que d'autres et si la distance entre ces deux maillons évolue. Nous présenterons les extractions obtenues sur un échantillon avant de proposer une première analyse des résultats. Enfin, nous décrirons nos observations sur l'ensemble du corpus.

#### 5.1.1. Méthode choisie

Pour cette partie, nous avons choisi de travailler dans un premier temps sur un échantillon composé de trois copies. Nous avons sélectionné les deux copies du gold de la première étude (cycle 3 et cycle 4) et nous avons choisi d'ajouter une copie du niveau Master (UN-M2-2018-TUTJ2-D1-R1-V1). Cette sélection permet d'explorer les premières tendances et de repérer d'éventuelles anomalies.

L'affichage du POS du m-1 s'appuie sur l'étiquette attribuée par Stanza au dernier token. Pour explorer les premières tendances, nous avons choisi une approche qui identifie correctement la forme des maillons simples. Nous avons conscience que cette approche simplifie la grande diversité des maillons composés. Ainsi, dans la phrase 43) le pronom *qui* a bien pour antécédent un SN car le maillon m-1 se termine par le nom *garçon*.

> 43) La sonnerie avait à peine retenti que le jeune garçon franchissait déjà la grille qui le séparait des vacances. [CO-3e-2018-FSBJC6-D1-R5-V1]

Cependant, dans la phrase 44) le SN *la tante taciturne* sera identifié comme adjectif.

> 44) Les enfants se jetèrent des regards apeurés, peu rassurés par la seule présence de la tante taciturne. [UN-M2-2021-UCL-D1-R34-V1]

Cette approche, bien qu'imparfaite, nous permet d'obtenir un premier regard général. Lors de nos analyses (5.2.1. Description de l'extraction), nous remarquons en effet des tendances qui confortent nos choix pour ce premier travail.

Contrairement aux précédentes étapes, nous séparons les noms propres des syntagmes nominaux pour le m-1, afin d'obtenir une analyse plus fine du m-1, objet d'étude de cette partie. En revanche, pour les maillons qui suivent, nous préférons conserver les étiquettes attribuées aux étapes précédentes : pronom, SN et autre. En effet, pour ce travail de prolongement, nous souhaitons rester dans le cadre installé dans ce mémoire afin d'assurer une continuité et de pouvoir comparer les résultats. C'est pourquoi nous choisissons de caractériser les m-1 des maillons pronoms et SN.

#### 5.1.2. Description des extractions

Le script python appliqué aux trois copies sélectionnées obtient en sortie un fichier tsv composé de huit colonnes. Le tableau 21 est un extrait du fichier résultat.

| Fichier | Referent | POS | Forme | Collectif | Cycle | Distance | POS_m-1 |
|---|---|---|---|---|---|---|---|
| CO-4e-2018-LSPJJRC-D1-R22-V1.conllu | Elle | SN | une fille nom de Jeanne | non | 4 | | |
| CO-4e-2018-LSPJJRC-D1-R22-V1.conllu | Elle | autre | son | non | 4 | 1 | PROPN |
| CO-4e-2018-LSPJJRC-D1-R22-V1.conllu | Elle | pronom | elle | non | 4 | 3 | DET |

*Tableau 21: Extraction des POS des m-1*

Les colonnes sont identiques à l'extraction présentée dans la partie 3.1.3. Description de la sortie attendue :

- l'identifiant de la copie ;
- le référent de la chaîne ;
- le POS selon notre classement (pronom ou SN) ;
- le texte du maillon extrait ;
- l'indication du caractère collectif ;
- le niveau scolaire ;
- la distance qui sépare le maillon présenté du m-1

La dernière colonne nous intéresse particulièrement car c'est elle qui précise le POS du m-1. Elle permet également de dissocier les POS des maillons m-1 et d'affiner l'analyse en séparant les SN en trois classes : déterminant (DET), nom (NOUN) et nom propre (PROPN). Cette séparation permet notamment d'observer les éventuelles variations entre l'usage d'un nom propre et celui d'un syntagme nominal.

Pour l'exemple :

45) Il était une fois une femme qui s'appela Emse. [CO-6e-2016-PJPR1-D1-R9-V1]

Nous pourrons retrouver l'extraction suivante :

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 6e-2016-PJPR | Elle | SN | une femme | non | 3 | | |
| 6e-2016-PJPR | Elle | pronom | qui | non | 3 | 0 | NOUN |
| 6e-2016-PJPR | Elle | SN | Emse | non | 3 | 2 | PRON |

*Tableau 22: Extraction m-1 copie CO-6e-2016-PJPR1-D1-R9-V1*

Ce tableau indique que le maillon *une femme* est le premier maillon de la CR *Elle.* Aucune étiquette n'est précisée dans la colonne finale car il n'y a pas de maillon précédent. *Une femme* précède le maillon *qui*. La ligne correspondant à ce maillon indique que le m-1 est un *NOUN* car *femme* porte cette étiquette. Pour finir, *Emse* est bien précédé d'un pronom, *qui*.

### 5.1.3. Analyse des extractions de l'échantillon

L'extraction résultant de l'application du script python sur les trois copies contient 160 maillons qui correspondent aux critères établis dans la partie 3.1. Choix méthodologiques. Pour chaque maillon extrait, nous pouvons former une paire avec le m-1. Dans la perspective d'observer les éventuelles variations qui peuvent être liées au POS du m-1, nous avons commencé par quantifier les maillons ainsi que les paires correspondant aux enchaînements des maillons dans les copies. Le tableau 23 résume ces résultats.

| POS m-1 | Pronom | SN |
|---|---|---|
| <NA> | 2 | 7 |
| DET | 8 | 6 |
| INTJ | 1 | 0 |
| NOUN | 12 | 5 |
| PRON | 60 | 31 |
| PROPN | 15 | 8 |
| NUM | 2 | 0 |
| VERB | 2 | 1 |
| Total | **102** | **58** |

*Tableau 23: Quantification des paires maillon/m-1 dans l'échantillon*

Nous pouvons remarquer tout d'abord que 102 maillons courants sont des pronoms et 58 sont des SN. L'extraction permet d'identifier 16 enchaînements, bien que tous ne comportent pas d'exemples pour les deux formes, comme c'est le cas pour un numéro suivi d'un SN qui n'a pas été repéré dans le corpus. La paire la plus fréquente est pronom-pronom avec 60 occurrences, elle représente environ le double des occurrences de la paire pronom-SN (31 occurrences). Le graphique 13 montre la répartition en pourcentage des POS du m-1 selon la forme du maillon.

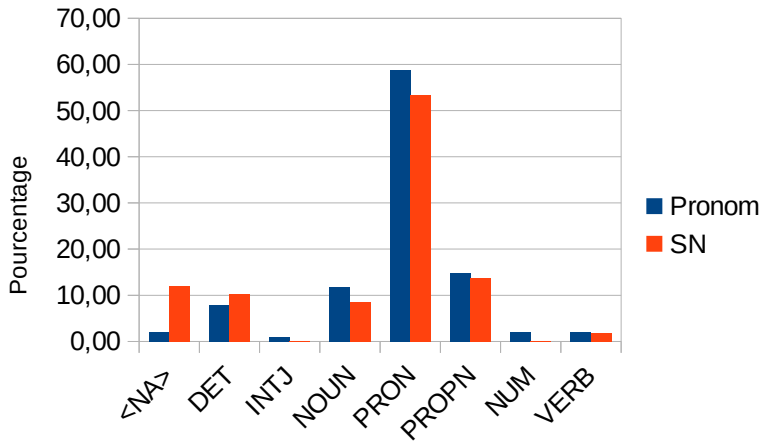



*Figure 13: Répartition en pourcentage des POS du m-1 selon la forme du maillon, dans l'échantillon*

Nous pouvons observer deux tendances principales :

- Les pronoms m-1 sont largement majoritaires. Ils précèdent 58,82 % des maillons pronoms et 53,45 % des maillons SN ;
- La présence importante de noms propres comme maillon précédent. Ils précèdent 14,71 % des maillons pronoms et 13,79% des maillons SN.

Ces observations indiquent tout d'abord une forte présence de la paire pronom-pronom. Ensuite, nous pouvons voir que les noms propres introduisent ou permettent de maintenir des référents sous forme nominale ou pronominale de manière fréquente. Ils sont plus représentés que les noms (11,76 % avant un pronom et 8,62 % avant un SN).

La présence d'une interjection comme maillon m-1 résulte d'une erreur de catégorisation de Stanza. Dans la phrase 46) , *calme-toi* est présenté comme interjection.

46) Mais il fait déja nuit Jeanne, calme-toi et explique-moi ce qu'il t'est arrivé [CO-4e-2018-LSPJJRC-D1-R22-V1]

Cette erreur est présentée de manière plus détaillée dans la partie 3.3.1. Script python appliqué aux copies annotées : analyse des erreurs.

Nous avons examiné les maillons classés comme NUM afin de vérifier l'analyse proposée par Stanza. Tout d'abord, les deux occurrences repérées sont en réalité un seul terme faisant partie des chaînes de référence *ElleColl* et *Les Enfants.* Ce maillon est présenté dans l'exemple 47).

47) Tous trois y furent poussés, et lorsqu'ils se tournèrent en tremblant ils ne virent qu'une seule personne [UN-M2-2018-TUTJ2-D1-R1-V1]

*Trois* est bien un déterminant numéral. Cependant, nous estimons que la locution *tous trois* fonctionne en réalité comme un SN.

### 5.1.4. Analyse du lien entre la forme du m-1 et la distance dans l'échantillon

Cette partie a pour ambition d'approfondir l'analyse de ce mémoire en observant la distance moyenne pour chaque paire identifiée dans la partie précédente. Autrement dit, il s'agit de vérifier si certains enchaînements sont associés à des reprises courtes ou, à l'inverse, si des enchaînements sont associés à des reprises longues. Le diagramme 14 illustre les distances moyennes observées dans notre échantillon.

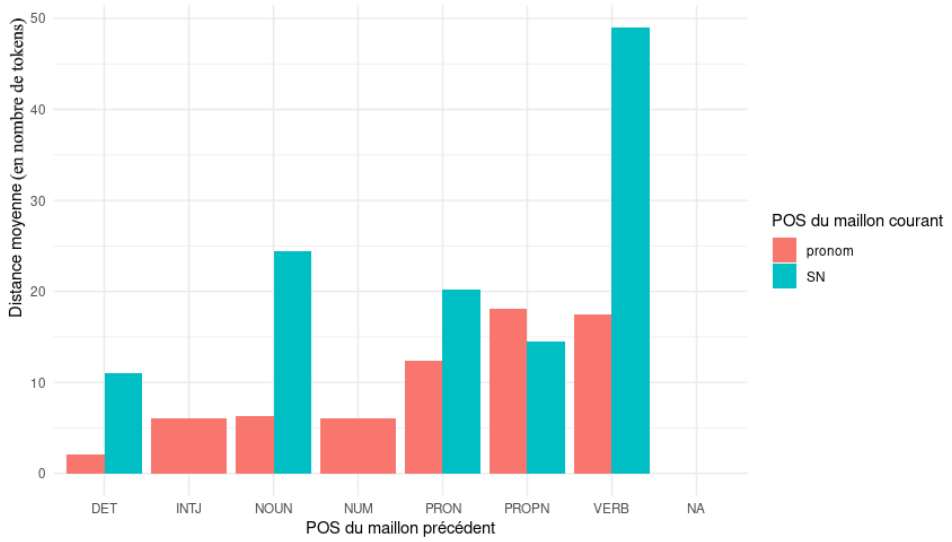



*Figure 14: Distance moyenne par paire (m-1/maillon) dans l'échantillon*

L'analyse de ce diagramme permet de remarquer plusieurs tendances concernant les distances moyennes selon les paires :

- La valeur la plus élevée concerne la paire verbe - SN (49 tokens). Cependant, notre corpus ne contient qu'une seule paire correspondant à cet enchaînement (exemple 48)). Cette observation doit donc être nuancée.

48) Il se retourna en entendant ce grand bruit, et, sans attendre, **s'enfuit** par la porte restée ouverte. Les petits se mirent alors à pleurer plus fort, tout en appelant leurs parents. Ces derniers apparurent aussitôt, et prirent leurs enfants pour les consoler et les coucher. Ils n'étaient pas inquiets : c'étaient eux-mêmes, et **l'un de leurs amis**, qui étaient derrière cet inquiétant épisode. [UN-M2-2018-TUTJ2-D1-R1-V1]

Dans cet exemple, le verbe *s'enfuit* comporte un sujet zéro faisant partie de la CR *Il*. La mention suivante de cette CR est *l'un de leurs amis*. Les deux mentions sont espacés par la reprise du récit principal des *petits*. Les mentions de la CR *Il* peuvent être distanciées car c'est le seul référent singulier de ce passage il n'est donc pas en compétition de plus il est ancré dans le récit.

- Pour les cas où le m-1 est un nom, nous pouvons repérer la distance la plus élevée, si nous choisissons d'ignorer la paire verbe – SN, trop peu représentée. Un nom suivi d'un SN a une distance moyenne de 24,4 tokens. Cette distance baisse largement lorsque le nom est suivi par un pronom (6,25 tokens).
- Pour les cas dans lesquels le m-1 est un pronom, la distance moyenne est de 20,19 tokens lorsqu'il est suivi d'un SN et de 12,35 lorsqu'il est suivi par un autre pronom.
- Les noms propres présentent moins de variabilité. Ils sont séparés d'un pronom par 18,13 tokens en moyenne contre 14,50 pour les SN.

Afin de préciser la distribution des distances selon les paires, nous illustrons la dispersion et les valeurs extrêmes grâce à la boite à moustaches 15 . Nous choisissons de n'illustrer que les 6 paires principales.

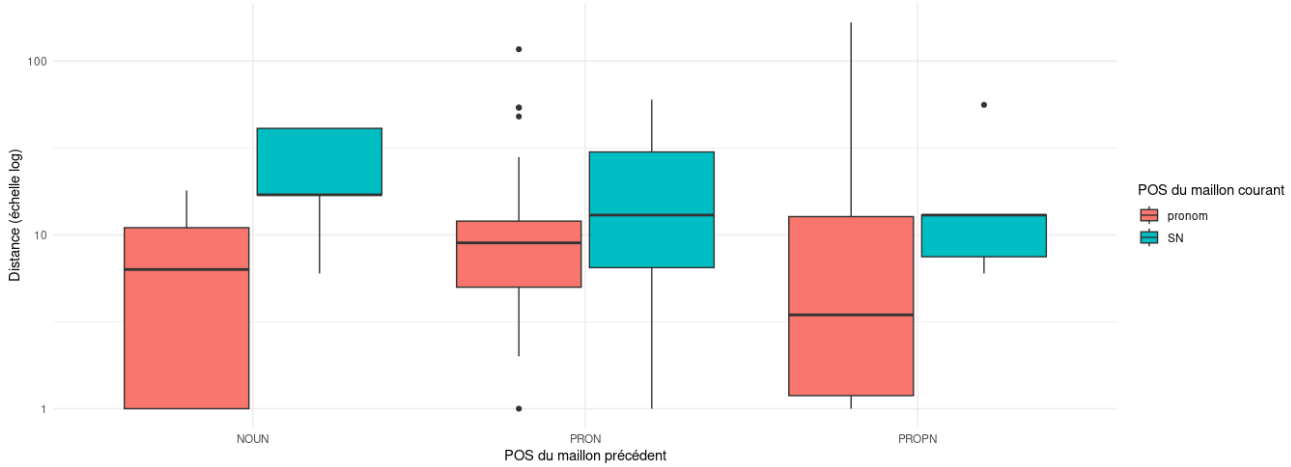



*Figure 15: Distribution logarithmique des distances selon le POS du m-1 dans l'échantillon*

L'observation de cette boîte à moustaches nous permet de mettre en valeur plusieurs éléments.

Lorsque nous observons les valeurs des m-1 pronoms, nous pouvons remarquer la présence de valeurs extrêmes lorsqu'ils sont suivis d'un pronom. Ces valeurs très élevées (ou très basse) rendent la moyenne moins fiable car elle est sensible aux extrêmes, contrairement à la médiane. L'exemple 36), déjà analysé, illustre l'occurrence de la distance la plus élevée. La valeur extrême réduite correspond à l'exemple 49).

49) Une jour, une fille du nom de Jeanne invite son ami Paul chez elle. Elle habitait dans cette maison depuis longtemps.

Dans cet exemple, le pronom *Elle* fait partie de la consigne. L'élève n'avait donc pas la possibilité de choisir, à proprement dit, la forme du référent qui suit le pronom *elle*.

La médiane de la distance est moins élevée pour les pronoms suivis d'un pronoms que celle des pronoms suivis d'un SN. Cependant, la boîte des SN qui suivent un pronom est plus étendue, témoignant de valeurs plus dispersées autour de la médiane. Cette dispersion est le signe d'une plus grande variabilité dans les distances entre un pronom et un SN.

Les valeurs des m-1 noms propres nous permettent d'observer d'autres phénomènes. Les noms propres suivis d'un pronom ont une distance médiane plus courte que lorsqu'ils sont suivis par un SN. La

boîte des pronoms montre une grande variabilité. La boîte des SN montre une faible variabilité des distances autour de la médiane, bien qu'une distance élevée puisse être observée. Cette distance correspond à l'exemple suivant :

50) Au bout de quelques heures de réflexion sur des mathématiques, Jeanne a un creux et va chercher quelque chose à boire pour elle et **Paul**. Mais sur le chemin du retour au salon elle crut entendre une voix. Une voix qui lui était très familière. Qu'elle n'avait plus entendue depuis quelques mois. Il lui semblait entendre la voix de son père derrière la porte d'entrée ! Prise de panique elle hurla « **PAUL** VIENS VITE ! ! (...) » [CO-4e-2018-LSPJJRC-D1-R22-V1]

Dans cet exemple, nous pouvons suivre le récit d'*Elle*. Le référent *Il* semble saillant car il est introduit depuis le début du récit et est associé à *elle* (qui est en position sujet). Cependant, le référent *Il* est ensuite réintroduit dans un discours direct par le nom propre *Paul*. Cette redénomination est nécessaire car un second personnage singulier et masculin est introduit : *son père,* ce qui peut mener à un flou référentiel. Par exemple, si l'élève avait choisi d'écrire : « Viens vite », le lecteur aurait pu se questionner sur le référent : *son père* ou *Paul* ?

Pour résumer, il semblerait que la distance, le pos de la mention précédente et le pos de la mention actuelle soient liés. En effet, les pronoms tendent à être plus éloignés lorsque le m-1 est un pronom que lorsque le m-1 est un nom ou un nom propre mais ces valeurs sont très dispersées. La distance entre un SN et le m-1 semble plus stabilisée : les médianes sont relativement proches et les boîtes montrent une dispersion moins importante que celle observée pour les pronoms.

### 5.2. Extraction du POS du m-1 dans le corpus

Pour terminer, nous avons appliqué le script à l'ensemble du corpus. Dans le cadre de ce mémoire, nous proposerons une description des extractions.

#### 5.2.1. Description de l'extraction

Le script permet d'identifier et de quantifier le POS du m-1 des 11926 maillons de notre étude. Nous avons quantifié la répartition des maillons, par paire, dans le tableau 24. Les premiers maillons ont été exclus de la présentation des résultats car ils ne présentent pas de m-1.

| | Pronom | SN |
|---|---|---|
| ADJ | 77 | 45 |
| ADP | 4 | 6 |
| ADV | 15 | 7 |
| AUX | 7 | 2 |
| CCONJ | 0 | 2 |
| DET | 936 | 771 |
| INTJ | 1 | 0 |
| NOUN | 1209 | 884 |
| NUM | 37 | 23 |
| PRON | 3585 | 1797 |
| PROPN | 804 | 422 |
| PUNCT | 9 | 6 |
| SCONJ | 3 | 1 |
| VERB | 329 | 155 |
| X | 3 | 0 |
| | 7019 | 4121 |

*Tableau 24: Quantification des paires maillon/m-1 dans le corpus*

Dans ce tableau, nous pouvons identifier 30 paires. La paire comportant le plus d'occurrences est la paire pronom-pronom (3585 occurrences), à l'image de notre extraction sur échantillon. Afin de préciser les tendances observées, nous proposons la figure 16 qui présente la répartition en pourcentage des POS du m-1 selon la forme du maillon.

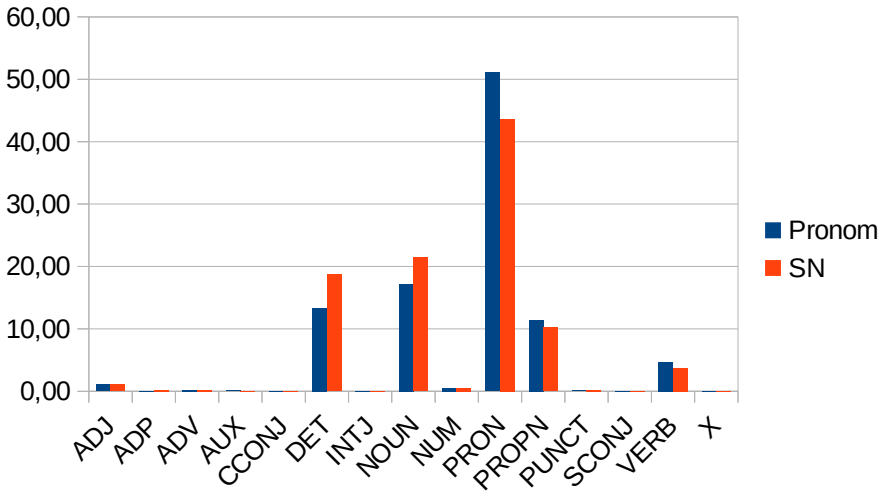



*Figure 16: Répartition en pourcentage des POS du m-1 selon la forme du maillon.*

Sur ce diagramme, plusieurs tendances sont mises en lumière :

- La grande majorité des m-1 pronoms : ils précèdent 51,08 % des pronoms et 43,61 des SN. Cette répartition est semblable à celle observée dans l'échantillon.
- La présence importante des noms, qui précèdent 17,22 % des pronoms et 21,45 % des SN.
- Les déterminants qui précèdent 13,34 % des pronoms et 18,71 %des SN.
- Les noms propres précèdent 11,45 % des pronoms et 10,25 % des SN.
- Les autres catégories grammaticales sont moins présentes.

Ces observations permettent d'affirmer que les m-1 pronoms sont les plus fréquents dans notre corpus. Les déterminant, nom et noms propres sont également bien représentés.

Pour vérifier s'il existe une association entre la forme du m-1 et la forme du maillon, nous avons recours au test Khi-2. La p-value retournée étant inférieure à 0,05 ( $p = 3.842e-07$), nous pouvons en conclure que la forme du m-1 et la forme du maillon sont statistiquement liée.

De plus, nous avons analysé les résidus standardisés pour mettre en lumière les paires surreprésentées ou sous-représentées dans notre corpus. Le graphique 17 illustre les résidus.

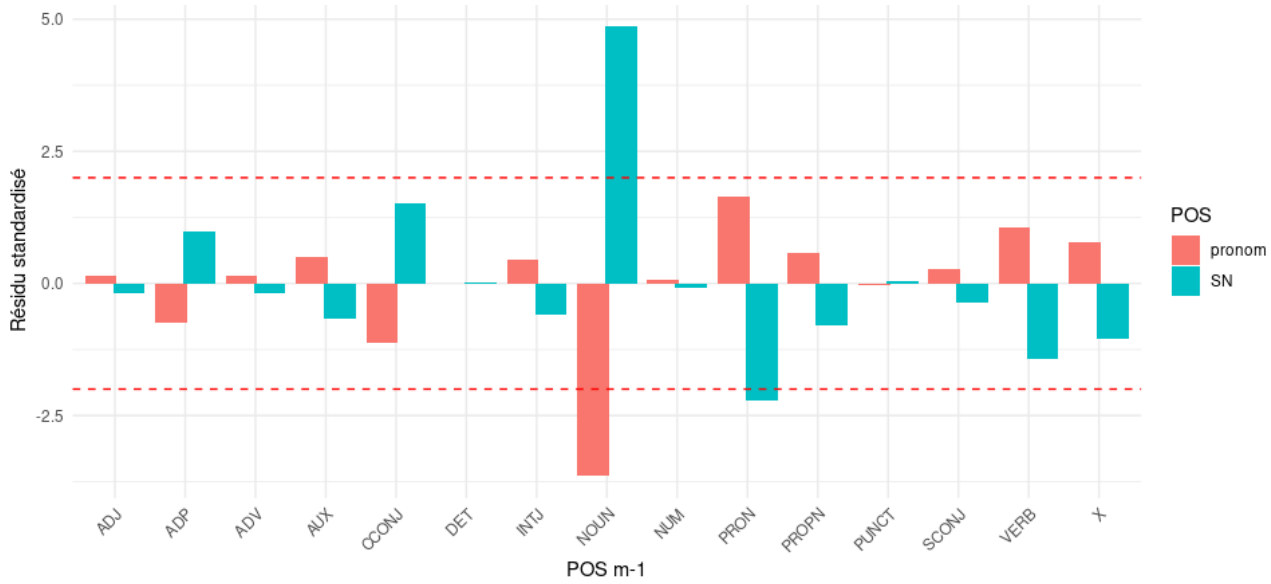



*Figure 17: Résidus standardisés du test Khi-2*

Bien que la paire pronom-pronom soit la plus fréquente, elle n'apparaît pas comme surreprésentée selon le test du Khi-2. La fréquence correspond à ce qui pourrait être observé si les variables étaient indépendantes. Au contraire, les valeurs surreprésentées, qui correspondent aux valeurs dont le résidu dépasse 2, sont plus présentes que ce qui était attendu. Dans notre cas, nous pouvons constater que le nom suivi d'un SN est fortement surreprésenté. Les paires sous-représentées sont les noms suivis d'un pronom et les pronoms suivis d'un SN, car la valeur du résidu est en dessous de -2.

Ces tendances suggèrent qu'il existe des régularités dans l'enchaînement des maillons pronominaux et nominaux et des préférences marquées dans la manière d'introduire, de maintenir ou de réintroduire les référents.

### 5.2.2. Analyse du lien entre la forme du m-1 et la distance dans le corpus

Dans cette partie, nous cherchons à observer le lien entre le POS du m-1, la distance et la forme d'un maillon pour l'ensemble de notre corpus. Le diagramme 18 permet d'observer les tendances.

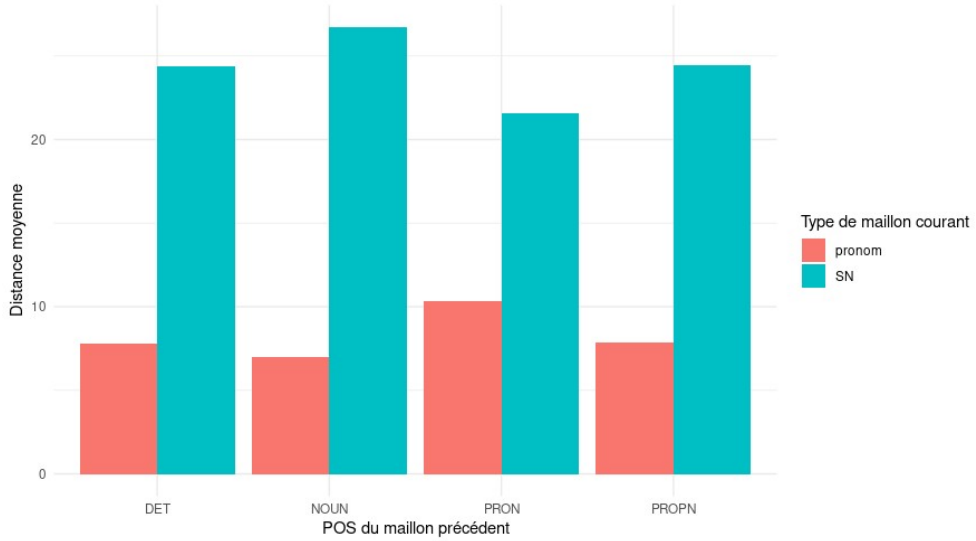



*Figure 18: Distance moyenne par paire (m-1/maillon)*

Ce diagramme nous permet de voir que la distance entre un pronom et le m-1 est toujours plus basse que celle observée entre un SN et le m-1. De plus, il semble que les moyennes varient légèrement selon la catégorie grammaticale du m-1. Ainsi, la distance moyenne entre un nom et un SN est de 26,70 tokens contre 21,54 tokens entre un pronom et un SN. De même, la distance moyenne entre un pronom et un pronom est de 10,36 tokens contre 7,01 entre un nom et un pronom. Cependant, la moyenne est sensible aux valeurs extrêmes, c'est pourquoi nous avons recours à la boîte à moustaches suivante.

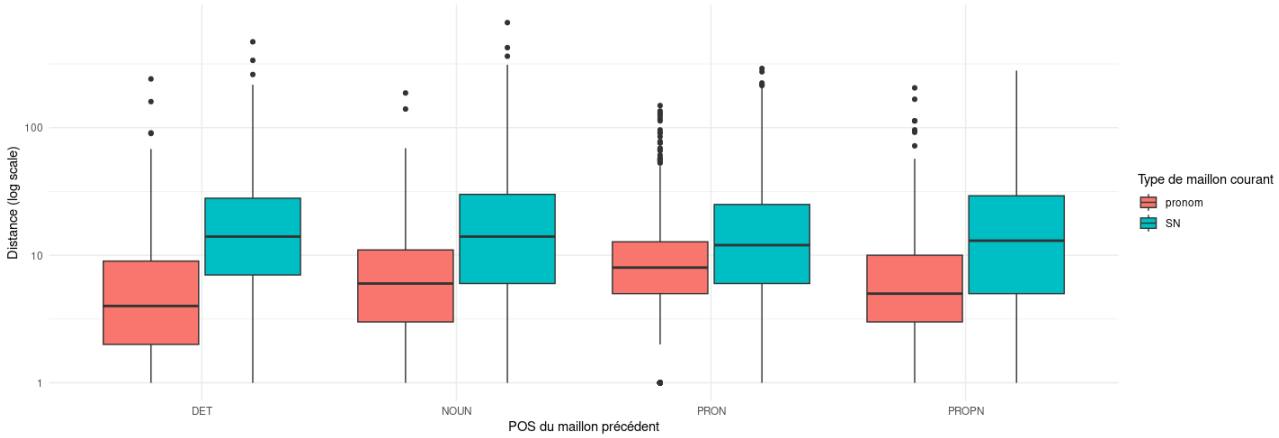



*Figure 19: Distribution logarithmique des distances selon le POS du m-1*

L'échelle logarithmique nous permet d'obtenir une représentation claire notamment pour visualiser les valeurs extrêmes. L'analyse des distributions permet de mettre en lumière des différences entre SN et pronom, selon le POS du m-1. Les médianes des SN sont stables et n'évoluent pas selon la catégorie grammaticale du maillon précédent. Cependant, la variabilité autour de la médiane est variée : la boîte à moustaches la plus étendue est observée pour les SN qui suivent un nom propre. Au contraire, les pronoms sembles plus sensibles au POS du m-1. Les pronoms suivant un pronom présentent la médiane la plus élevée et la variabilité la plus faible malgré de nombreuses valeurs extrêmes.

## *Conclusion et perspectives*

Dans ce travail de recherche, nous avons étudié le lien entre la distance et le choix d'une forme référentielle dans des copies d'élèves et d'étudiant.es. Autrement dit, ce travail visait à caractériser le lien entre la distance qui sépare deux maillons d'une chaîne référentielle et la forme de l'expression référentielle employée. Nous souhaitions décrire l'usage des pronoms et des syntagmes nominaux et comparer ces usages selon le niveau scolaire. Cette analyse s'est basée sur trois chaînes de référence différentes (*Elle*, *Il* et *Les Enfants*) et permet également d'examiner l'influence du type de référence (collective ou non).

### Bilan

Ce mémoire avait pour objectif de caractériser les contextes de maintien de référents humains dans des chaînes de référence à l'intérieur de récits produits par des élèves de cycle 3 et de cycle 4 ainsi que des étudiant.es de Master.

Nous avons commencé par présenter les notions théoriques en mobilisant notamment des travaux de Charolles, Ariel, Schnedecker, Landragin. Nous avons défini des notions clés de la référence et présenté la place des compétences référentielles dans les programmes scolaires (page 3). La seconde partie détaille les données de notre étude, en présentant notamment la constitution et la méthode d'annotation du corpus RésolCo (page 17). Puis nous avons décrit la méthode choisie pour l'extraction des maillons référentiels et des informations nécessaires pour notre analyse (page 22). Nous avons terminé cette partie avec une évaluation de notre extraction puis nous avons proposé une description quantitative des maillons extraits (page 28). Nous avons procédé à une étude statistique de la liaison entre distance et forme référentielle (page 33) avant de proposer une première analyse des m-1 (page 45).

Nos résultats mettent en évidence diverses tendances : plus la distance entre deux maillons augmente, plus il est probable de trouver une forme référentielle informative comme un SN (comme détaillé dans la partie 1.3.1. Le choix de l'expression référentielle selon Ariel). A l'inverse, une distance plus réduite, indice d'un référent plus accessible, favorise l'emploi de formes moins informatives comme les pronoms. Ce phénomène évolue selon le niveau scolaire et semble être plus régulier au Master, bien que les élèves du cycle 3 semblent déjà sensibles à certaines contraintes cognitives. Le référent a une influence sur la distance, en particulier pour les SN. Le caractère collectif du référent a un impact plus modéré mais toutefois notable selon les référents et le niveau scolaire. Il est intéressant d'affiner notre analyse entre référent collectif et non collectif car nous avons pu mettre en évidence des emplois différenciés des pronoms selon le type de référence.

Notre étude permet de confirmer que la gestion de la référence est une compétence sensible à la distance, au référent et qu'elle évolue avec l'âge et l'expérience. Pour finir, la première analyse des m-1 suggère que la forme d'un maillon est liée à la forme du maillon qui le précède et à la distance qui les sépare. Ces observations permettent de penser la chaîne comme un système complexe, sous-tendu par un ensemble de variables.

Ce mémoire propose une contribution à l'étude des chaînes de référence. Il s'appuie sur un cadre théorique pour réaliser une analyse d'un corpus authentique de travaux d'élèves et d'étudiant.es. Le travail de statistiques descriptives permet de croiser des variables tout en quantifiant les variations de la distance et de la forme. La méthode proposée est semi-automatisée, ce qui permet une analyse quantitative reproductible. Nos résultats suggèrent que la compétence référentielle évolue, bien que les élèves du cycle 3 soient déjà sensibles aux contraintes cognitives. Ces résultats peuvent servir de leviers pour proposer des contenus pédagogiques spécifiques. Ils invitent à renforcer l'enseignement de l'identification de référents dans des contextes plus ou moins étendus, à proposer des activités de reformulation ou de correction et ce, dès le cycle 3.

**Limites**

Ce travail présente plusieurs limites qui doivent être prises en compte pour l'analyse des résultats. Tout d'abord, certaines erreurs d'extraction automatique sont difficiles à quantifier. Le travail d'annotation manuelle d'un gold a permis de vérifier la qualité d'une partie des extractions et a obtenu des scores satisfaisants mais nous n'avons pas pu évaluer la qualité de l'ensemble.

Nous insistons sur les erreurs de calculs de distance, pour les maillons composés qui réfèrent à plusieurs entités de manière collective ou non. Les distances négatives ont été délibérément écartées mais doivent faire l'objet d'une mise en garde.

De plus, les analyses statistiques reposent sur des représentations graphiques et des tests bivariés. Nous avons conscience de l'existence de méthodes multifactorielles plus appropriées pour nos données. Cependant, ces approches nécessitent une interprétation plus complexe qui dépasse le cadre de ce mémoire de première année de Master.

**Perspectives**

Plusieurs pistes de prolongement peuvent être envisagées. Tout d'abord il pourrait être intéressant d'affiner l'analyse commencée dans la partie 5. Ce travail qui a pour objectif d'observer la forme référentielle du maillon précédent permettrait d'observer si la distance est également liée à la forme du m-1, comme nos résultats le suggèrent. Dans un second temps, l'analyse pourrait montrer l'éventuelle variabilité de la distribution selon le niveau scolaire, le référent ou le type de maillon. Nous avons proposé un premier travail mais il reste encore trop peu développé et pourra faire l'objet d'une recherche plus approfondie et agrémentée de tests statistiques appropriés pour vérifier la significativité des tendances remarquées.

De plus, dans le cadre de ce travail nous avons souhaité observer les formes pronominales en opposition aux syntagmes nominaux. Pour cela, nous avons choisi de regrouper différents pronoms et divers syntagmes nominaux. Il pourrait être pertinent d'affiner la comparaison en séparant les pronoms relatifs des pronoms personnels ou les différents syntagmes nominaux (défini, indéfini, possessif, démonstratif) et les noms propres.

Il pourrait également être pertinent d'étudier avec précision les valeurs atypiques (*outliers*) remarquées lors de nos analyses statistiques. Cette étude pourrait révéler des stratégies référentielles particulières ou des cas d'usages non cohérents et permettre de mieux accompagner les enseignant.es dans la création de contenus pédagogiques adaptés.

Par ailleurs, au-delà du critère de la distance, il serait pertinent d'étendre l'analyse aux autres critères fondamentaux des théories cognitives du traitement référentiel, tels que la compétition ou la saillance.

Enfin, une réflexion pourrait être menée sur le traitement des référents non humains proposés dans la consigne du corpus RéSolCo (*cette maison*, *cette nuit*, *ce grand bruit*), que nous n'avons pas pris en compte dans ce mémoire.

Ces pistes restent à approfondir dans le cadre de travaux ultérieurs. L'analyse menée dans ce mémoire a tenu compte de contraintes de temps et de compétences.

***Bibliographie***


Ariel, M. (1990). *Accessing Noun-Phrase Antecedents.* Routledge.

Charolles, M. (1978). *Introduction aux problèmes de la cohérence des textes*.

Charolles, M. (1987). Contraintes pesant sur la constitution des chaînes de référence comportant un nom propre. *Cahiers du Centre de recherches sémiologiques* (53), 29-55.

Charolles, M. (1988). Les plans d'organisation textuelle : périodes, chaînes, portée et séquences , *Pratiques*, 57, 3-13.

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

Charolles, M. (2002). *La référence et les expressions référentielles en français*. Ophrys.

Charolles, M. (2021). Les différentes expressions anaphoriques. In Abeillé, A. & Godard, D. (Dir.), *La Grande Grammaire du Français* (Imprimerie nationale Éditions). Actes Sud.

Combettes, B. (2021). La continuité référentielle : Anaphores et cataphores. In A. Abeillé & D. Godard (Eds.), *La Grande Grammaire du Français* (Imprimerie nationale Éditions). Actes Sud.

Corblin, F. (1985). Les chaînes de référence : Analyse linguistique et traitement automatique. *Intellectica. Revue de l'Association pour la Recherche Cognitive*, *Les interactions homme / ordinateur* (1), 123-143.

Corblin, F. (1995). *Les chaînes de référence dans le discours* (Presses Universitaires de Rennes).

Delaborde, M., & Landragin, F. (2019). *De la coréférence exacte à la coréférence complexe : Problèmes typologiques et complexité de mise en œuvre en corpus*. 10èmes Journées Internationales de la Linguistique de Corpus, Grenoble.

Elalouf, M.-L. (2011). Constitution de corpus scolaires et universitaires : vers un changement d'échelle ? *Pratiques. Linguistique, littérature, didactique*, (149), 56-70.

Federzoni, S., Rebeyrolle, J., Ho-Dac, L.-M., Bard, Y., & Garcia-Debanc, C. (2022). *Guide d'annotation RésolCo*.

Garcia-Debanc, C., Ho-Dac, L.-M., Bras, M., & Rebeyrolle, J. (2017). Vers l'annotation discursive de textes d'élèves. *Corpus*, (16), 157-184.

Ho-Dac, L.-M., Federzoni, S., Bras, M., Rebeyrolle, J., & Garcia-Debanc, C. (2020). *RésolCo un corpus de manuscrits d'élèves et d'étudiants pour l'étude de la cohérence.* 10èmes Journées Internationale de la Linguistique de Corpus, Grenoble, France.

Landragin, F. (2011). Une procédure d'analyse et d'annotation des chaînes de coréférence dans des textes écrits. *Corpus*, *10*, 61-80.

Landragin, F. (2021). Approches contrastives des chaînes de référence. Présentation. *Travaux de linguistique*, *82*(1), 7-16.

Landragin, F. (2021bis). L'ancrage des énoncés dans l'énonciation. In Abeillé, A. & Godard, D. (Dir.), *La Grande Grammaire du Français* (Imprimerie nationale Éditions). Actes Sud.

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

Milner, J.-C. (1982). *Ordres et raisons de langue*. Éditions du Seuil.

Oberle, B. (2019). *Détection automatique de chaînes de coréférence pour le français écrit : Règles et ressources adaptées au repérage de phénomènes linguistiques spécifiques*. Conférence sur le Traitement Automatique des Langues Naturelles (TALN-RECITAL), Toulouse.

Ogrodniczuk, M., Glowinska, K., Kopec, M., Savary, A. & Zawislawska, M. (2015). *Coreference: Annotation, Resolution and Evaluation in Polish*. De Gruyter.

Pepin, L. (2009). *La coréférence dans la narration, première partie*.

Philippe, G. (1998). Les démonstratifs et le statut énonciatif des textes de fiction : L'exemple des ouvertures de roman, *Langue française, Les démonstratifs : théorie linguistique et textes littéraires* (120), 51-65.

Ponton, C., Doquet, C., Fleury, S., Ho-Dac, L. M. (2022). E-CALM [Corpus] (v2.1). ORTOLANG (Open Resources and TOols for LANGuage)

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., & Manning, C. D. (2020). *Stanza : A Python Natural Language Processing Toolkit for Many Human Languages* (No. arXiv:2003.07082). arXiv.

[RésolCo] *Corpus de manuscrits d'élèves et d'étudiants pour l'étude de la cohérence*, CLLE (UMR 5263) & Université Toulouse – Jean Jaurès. [http://redac.univ-tlse2.fr/corpus/resolco.html]

Riegel, M., Pellat, J.-C., & Rioul, R. (1994). *Grammaire méthodique du français.* PUF.

Salazar Orvig, A., Fayolle, V., Hassan, R., Leber-Marin, J., Marcos, H., Morgenstein, A., & Pares, J. (2004). ÉMERGENCE DES MARQUEURS ANAPHORIQUES : le cas des pronoms. *Cahiers d'acquisition et de pathologie du langage*, *24*, 57-82.

Salles, M. (2015). Chaînes de référence : La deuxième mention. L'exemple des entités inanimées dans les narrations littéraires. *Travaux de linguistique*, *71*(2), 111-133.

Schnedecker, C., & Landragin, F. (2014). Les chaînes de référence : Présentation. *Langages*, *195*(3), 3-22.

Schnedecker, C. (2019). De l'intérêt de la notion de chaîne de référence par rapport à celles d'anaphore et de coréférence. *Cahiers de praxématique*, *72*, Article 72.

Schnedecker, C. (2021). *Les chaînes de référence en français*. Ophrys.

Tasmowski, L., & Laca, B. (2021). La détermination et la quantification.In Abeillé, A. & Godard, D. (Dir.), *La Grande Grammaire du Français* (Imprimerie nationale Éditions). Actes Sud

Widlöcher, A., & Mathet, Y. (2009). *La plate-forme Glozz : Environnement d'annotation et d'exploration de corpus.* Conférence Traitement Automatique des Langues Naturelles, Senlis, France.

Zribi-Hertz, A. (2021). Les pronoms et les proformes. In Abeillé, A. & Godard, D. (Dir.), *La Grande Grammaire du Français* (Imprimerie nationale Éditions). Actes Sud