Mémoire de Master 1 de Sciences du Langage
Parcours Linguistique, Informatique et Technologies du Langage (LITL)

Les cohabitations des continuités référentielles dans le corpus RésolCo : modélisations et détection automatique

Lucas Aubertin

Sous la direction de Lydia-Mai Ho-Dac & Josette Rebeyrolle

Juin 2021

Remerciements

Je voudrais remercier mes deux directrices de mémoire, Lydia-Mai HO-DAC et Josette REBEYROLLE pour leur aide et leurs conseils avisés tout au long de ce travail de mémoire. Je tiens également à les remercier pour leur disponibilité sans faille.

J'aimerais remercier tout particulièrement mes amis et mes proches qui m'ont soutenu tout au long de l'année et sans qui ce mémoire n'aurait peut-être pas vu le jour.

Table des matières

Liste des tableaux

Table des illustrations

Introduction

Ce mémoire discute des cohabitations des continuités référentielles liées à la notion de chaine de référence. Cependant, les définitions et modèles qui seront présentés seront ceux des chaines de référence car l'annotation réalisée dans le corpus RésolCo s'approche de celle des chaines même si elle diffère sur les types d'éléments qu'elle peut contenir. Ce mémoire ne se concentrant pas sur le contenu linguistique des unités annotées, nous proposons d'analyser les continuités référentielles comme des chaines de référence puisque leurs structurations se révèlent similaires. Afin d'expliciter les différences entre chaine et continuité, nous évoquerons plus en détail les éléments linguistiques annotés dans le corpus au cours de la partie II-3. Nous développerons également, dans la partie II-5, les raisons pour lesquelles, dans le cadre de ce mémoire, l'utilisation de la notion de continuité référentielle est préférable à celle de « chaine de référence ».

L'objectif de ce mémoire est de mettre en place une méthode automatique ou semi-automatique permettant de détecter et d'analyser les cohabitations des chaines de référence dans un texte. Les données utilisées dans le cadre de ce projet sont issues du corpus RésolCo. Les méthodes et les analyses se basent donc sur les écrits d'élèves normalisés et annotés en provenance de ce corpus.

La réalisation de l'objectif se fera en quatre temps. Dans un premier temps, il s'agira de fournir une description qualitative des différents types de cohabitations des chaines de référence. Cette description reposera sur les définitions théoriques présentes dans la littérature et nos observations réalisées en parcourant les données. Dans un deuxième temps, ces descriptions seront appliquées aux particularités et contraintes liées aux textes produits dans le cadre du corpus RésolCo. Dans un troisième temps, ces applications alimenteront la création d'une méthode de recherche automatique permettant de repérer les cohabitations des chaines de référence. Pour finir, les résultats issus de ces recherches seront analysés et nous tenterons d'amener des éléments permettant de comprendre les différences d'utilisation entre les types de cohabitations des chaines de référence.

I - État de l'art

La problématique de ce mémoire traitant de la détection automatique de la cohabitation des chaines de référence, il est nécessaire de décrire au préalable ces objets et domaines d'études. Cette partie servira donc ce but et sera composée de trois parties : la première traitera de la chaine de référence (désormais « CR »), la deuxième de la cohabitation des CR et la troisième de la détection automatique des CR et des cohabitations associées.

1 - Les chaines de référence

A - La chaine et ses composantes

Les chaines de référence sont décrites en 1988 par Charolles (cité dans Schnedecker, 2019) comme : « *constituées par des suites d'expressions coréférentielles [...]. Seules peuvent appartenir (donner lieu à) une chaine 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) ».

Pour éclaircir ces propos, il est possible de définir la chaine de référence comme une succession d'éléments référentiels (qui font référence à quelque chose) qui ont la propriété commune de renvoyer au même référent, réel ou fictif. Ces éléments peuvent être nominaux ou pronominaux mais aussi revêtir d'autres formes particulières que nous serons amenés à discuter. Les expressions référentielles qui composent les chaines de référence entretiennent des relations entre elles, qui peuvent être décrites par les notions d'anaphore et de coréférence.

En quoi la CR est-elle différente de l'anaphore ou de la coréférence ? Le premier élément de réponse possible est le nombre d'éléments associés aux différentes notions. En effet, les recherches linguistiques sur l'anaphore ou la coréférence étudient majoritairement ces relations à travers des paires d'éléments référentiels, quand la CR se distingue en contenant au minimum deux éléments (une paire) jusqu'à une infinité ; ou plutôt un grand nombre « limité », nous y reviendrons. Cette multiplication des expressions référentielles amène des contraintes et des spécificités de traitement différentes de la paire anaphorique ou coréférentielle. La CR est donc un élément linguistique à part entière. Son contenu pouvant être des éléments anaphoriques ou coréférentiels, ou les deux simultanément, nous allons décrire ces relations car si l'anaphore et la coréférence diffèrent, leurs différences ne sont pas toujours aisément perceptibles.
Afin de remédier à cela et pour la compréhension des termes utilisés dans ce mémoire, nous allons définir et caractériser efficacement les deux types de relations référentielles que sont l'anaphore et la coréférence.

B - L'anaphore

Milner (cité par Schnedecker, 2019) la décrit comme suit : « *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 relation implique donc une sorte de relation de soumission interprétative de l'élément B à l'élément A : **l'élément B ne porte pas de sens sans la présence de l'élément A**.

Le type d'élément second qui correspond le mieux à cette idée de dépendance est le pronom personnel, qui remplirait parfaitement sa fonction de renvoi. Mais l'anaphore n'est pas toujours une relation nom - pronom comme Erkü & Gundel (1987) en ont fait l'observation. Ils proposent alors le terme d'anaphore *indirecte* pour qualifier ces autres anaphores. Pour illustrer, nous reprenons les exemples de Schnedecker (2019) :

- L'anaphore conceptuelle (« lexicale », chez Milner, 1982) : *Paul a tué trois lions. Pierre en a tué cinq.*
- L'anaphore associative (d'après Guillaume, 1919, cité dans Riegel et al., 2009 : 1039) : *l'église… les vitraux… l'autel*
- L'anaphore générique (dans Webber, 1983, cité par Kleiber, 1991) : *Paul a acheté une Toyota, car elles/ces voitures sont robustes.*
- L'anaphore possessive : *Paul… son chien…*
- L'anaphore collective (Kleiber, 1991) : *À Strasbourg, ils roulent comme des fous.*

Dans ces exemples, le référent premier est bien un nom mais l'on observe que l'élément second n'est pas un pronom (personnel), pourtant, il fait bien référence au même objet ou concept que

le premier. Ce type de relation indirecte peut compliquer la détermination de l'anaphore et donc être un frein au développement d'une méthode de détection automatique ainsi qu'à la production d'annotations humaines assurées.
La difficulté réside essentiellement dans le fait que l'interprétation de l'anaphore indirecte demande une considération de son contexte ainsi que la « visualisation » d'un lien plus ou moins discret entre les deux éléments. En effet, le lien dans l'exemple de l'anaphore possessive nous semble plus facilement saisissable puisqu'il est explicité par l'adjectif possessif « son », qui relie immédiatement « Paul » et « chien ». Mais dans cet exemple, il n'y a qu'un autre référent présent. L'interprétation du lien se complique quand ils se multiplient. Par exemple, dans la phrase « Marie va voir Paul. Son chien… », il est difficile, sans avoir plus d'éléments, de savoir à qui fait référence « son ».
Dans l'exemple de l'anaphore collective, c'est encore plus délicat car il n'existe pas de lien grammatical immédiat entre « Strasbourg » et « ils » (probablement les habitants de Strasbourg). Si pour un programme informatique l'interprétation du lien entre les deux mots sera, à notre connaissance, très difficile voire quasi impossible, pour un humain, elle peut varier d'un interprétant à l'autre car la relation n'est pas explicite. Cette variabilité ne facilite donc pas la production d'annotations certaines.

C - La coréférence

En quoi la coréférence se différencie-t-elle alors de l'anaphore ? La différence majeure réside dans le fait que dans la coréférence **l'identité des éléments référentiels ne repose ni exclusivement, ni nécessairement sur la relation d'anaphore** (de dépendance). Les maillons, mentions, éléments référentiels peuvent être référentiellement autonomes et pourtant désigner la même chose. Pour illustrer ce concept, nous reprenons une succession d'éléments tirés de Schnedecker (2019) : « *Barak Obama… Le 44ième président des États-Unis… L'époux de Michelle Obama…* ».
Les trois éléments référentiels sont indépendants et coréférentiels, c'est-à-dire, qu'ils sont interprétables indépendamment de la présence des autres et qu'ils désignent tous la même chose : M. Obama.

Pour reprendre Milner (1982 : 32-33, cité par Schnedecker, 2019), il est possible de définir avec des termes plus logiques la notion de coréférence comme suit :

- Coréférence symétrique : « **Si A est coréférent à B alors B est coréférent à A** »
- Coréférence transitive : « **Si A est coréférent à B et B est coréférent à C alors A est coréférent à C** »

Par contraste, la relation d'anaphore est dite « asymétrique ». Boudreau (2004 : 21, cité par Schnedecker, 2019) précise que *« les anaphores sont aussi des relations transitives, mais elles sont asymétriques car la relation de dépendance oriente la relation. De plus, elles sont irréflexives car un élément référentiel ne peut pas être anaphorique à lui-même* ».

Après avoir défini les relations anaphoriques et coréférentielles, nous pouvons dès à présent considérer la chaine de référence comme une série de points, c'est-à-dire des unités linguistiques délimitées qui se font référence, que nous pouvons également considérer comme des « maillons » de la chaine. Ces unités sont reliées par des traits invisibles qui expriment leur référentialité anaphorique ou coréférentielle. Ces traits invisibles reliant deux unités sont également des « indices de cohésion » qui, d'après Halliday & Hasan (1976), tissent des liens

à travers le texte, créant ainsi sa texture. La chaine de référence est donc la somme de toutes les itérations se rapportant à un même référent et de tous les liens entre elles.

Cependant, cette définition de la CR pose la question de sa taille et de son dénombrement interne. En effet, si deux CR peuvent avoir la même longueur en termes de caractères parcourus, elles peuvent être composées d'un nombre de maillons différent. Pour répondre à cette problématique différentes mesures ont été imaginées, nous les verrons dans la section suivante.

De plus, les CR peuvent suivre le découpage du texte (Schnedecker & Landragin, 2014), les paragraphes peuvent donc voir le maillon initial de la CR réintroduit sous une forme identique au début d'un nouveau paragraphe. Dans ce cas, il est important de déterminer si la CR initiale doit être maintenue ou s'il faut l'arrêter à ce moment-là pour en démarrer une nouvelle. Nous privilégions les CR maintenues comme c'est le cas dans le corpus RésolCo. Nous pensons que le fait de reprendre un référent préalablement introduit et de le maintenir permet de créer la texture et la cohérence d'un texte. Il est donc important de privilégier la continuité de la CR tout au long d'un texte afin de la mettre en évidence et ce même si la CR est maintenue par-dessus des discontinuités telles que des sauts de paragraphes, voire de sections, des portions de texte sans mention du référent de la CR, etc.

D - Mesures des chaines de référence

Pour mesurer et comparer des CR, il existe plusieurs méthodes qui peuvent se cumuler. Il est en effet possible de mesurer (Schnedecker, 2019) :

- La longueur en comptant le nombre de maillons uniquement. **C'est la mesure que nous utiliserons dans ce mémoire.**
- Le nombre de maillons rapporté au nombre de référents (du texte) et à ceux présents dans une CR
- La complexité syntaxique (variations des classes grammaticales dans ou entre les CR)
- Le coefficient de stabilité de Perret. Pour un référent donné, on divise le nombre total d'anaphores nominales par le nombre de désignations différentes : *« Par exemple, dans la Mélusine de Jean d'Arras, pour la désignation de l'héroïne on rencontre 164 anaphores nominales, et 17 désignations différentes ; le coefficient de stabilité est donc 164/17 = 9,64.* (Perret, 2000 : 17) »

Et plus amplement :

- Le nombre de CR dans le texte (peut différer du nombre de référents présents)
- La distance entre les maillons. Cependant, à ce jour, à notre connaissance, il n'existe pas d'unité de mesure commune. Ces différentes unités peuvent être, par exemple, des syllabes, des mots, des syntagmes nominaux ou des phrases.
- La portée des CR ou le pourcentage de texte couvert par chaque CR
- La persistance des CR, étudiée par Givón (1983) qui est définie par le nombre de mentions antérieures et subséquentes à un point référentiel donné
- La composition des CR avec catégories et fonctions grammaticales des éléments
- L'ordre d'apparition dans le texte
- Le mode de cohabitation des CR : **succession, entrecroisement, fusion, dissociation, parallélisme, etc.**

Ainsi que d'autres points d'observations tels que le nombre de paragraphes qu'elles traversent, les domaines sémantiques abordés, etc.

2 - Les cohabitations des chaines de référence

Comme nous venons de le voir dans la section précédente, il existe plusieurs modes de cohabitations que nous allons décrire et illustrer. Dans ce but, nous reprenons la liste de cohabitations proposée par Schnedecker (2006) et l'approfondissement de Schnedecker & Landragin (2014). Les différents types de cohabitation y sont définis comme suit :

- **Succession**
  - Arrêt d'une chaine dès qu'une autre apparait (changement de référent).
  - Exemple : **Michel** vit paisiblement dans sa maison de campagne. **Michel** est aimé de **ses voisins**. Mais **Michel** possède une nouvelle voiture. Elle risque de causer beaucoup de bruit. Elle va vite, elle est belle mais elle risque de faire trembler *les murs*. *Ils* supportent de vieilles toitures, *ils* sont affaiblis. (Exemple construit)

- **Entrecroisement**
  - 2 référents ou plus se mêlent consécutivement et/ou simultanément.
  - Exemple : **Michel** conduit *sa voiture*. **Il** y prend du plaisir, *elle* est très confortable. *Cette voiture* était chère mais l'annonce était irrésistible. Alors **il** a craqué et ne *la* regrette pas. (Exemple construit)

- **Dérivation**
  - Tiré de l'anaphore associative ou possessive[1]. L'apparition d'un nouveau référent dans le texte n'est pas matérialisée par un syntagme nominal indéfini ou un nom propre mais par dérivation d'un autre référent.
  - Exemple pour l'anaphore associative : **Michel** conduit un peu trop vite et perd le contrôle. **Il** percute *une voiture*. La conductrice est tuée sur le coup. Elle rentrait de son travail. **Michel** est sonné mais vivant. (Exemple construit).
    Le transfert de référent se produit à l'apparition d' « une voiture » et il est possible de faire le lien logique entre la voiture et sa conductrice, nouvellement citée.
  - Exemple pour l'anaphore possessive : La conductrice s'appelait **Mathilde**. **Son** amant comptait **la** demander en mariage dans les jours à venir. **Sa** *sœur* **l'**attend chez **elle** pour **lui** annoncer une grande nouvelle. Tous les deux ne s'attendent pas à recevoir la terrible annonce de **son** décès. (Exemple construit).
    Les nouveaux référents sont associés à « Mathilde » par un pronom possessif les devançant.

- **Dissociation**
  - Séparation de 2 référents ou plus qui étaient inclus sous un seul dénominateur.

[1] Voir B - L'anaphore

- Exemple[2] : La police arrive rapidement sur le lieu de l'accident. **Deux hommes$_{(R)}$** sortent de la voiture. L'un$_{(D1)}$ très grand, l'air heureux. *Le second$_{(D2)}$*, plus petit, essuie désespérément la tâche que son café renversé vient de faire sur *sa* chemise. (Exemple construit)

- **Fusion**
  - Rassemblement de deux référents ou plus sous un seul dénominateur.
  - Exemple : Le grand$_{(D1)}$ lance un regard à *l'autre$_{(D2)}$*. *Il* l'interprète immédiatement et sait qu'il ne présage rien de bon. **Ils$_{(R)}$** se connaissent depuis longtemps. **Ils** sont allés à la même école de police. **Ils** communiquent presque par télépathie, ce qui fait la jalousie de **leurs** collègues. (Exemple construit)

- **Parallélisme**
  - Reprise alternée de deux référents avec le même niveau de saillance, c'est-à-dire d'importance équivalente.
  - Exemple : **Le premier policier**, **Jean**, s'approche de la voiture de Michel. Le second, Paul, de la voiture de Mathilde. **Jean** vérifie l'état de santé de Michel et Paul constate que Mathilde est décédée. **Jean** tente de faire sortir Michel, coincé dans sa voiture mais Paul **l'**interpelle. **Jean** laisse son accidenté et se dirige alors vers Paul… (Exemple construit)

Ces différentes cohabitations ne sont pas toujours présentes dans chacun des textes et certains genres textuels en favorisent certaines par rapport à d'autres. Les genres textuels peuvent également avoir une influence sur la composition des différentes expressions référentielles de la chaine de référence (Schnedecker, 2006).

3 - La détection des chaines de référence en français

Nous venons de voir la composition des chaines de référence et les interactions possibles entre elles. Afin de les étudier dans des productions écrites, il faut dans un premier temps les détecter. La détection débute en repérant les éléments référentiels dans un texte. Quand tous les éléments faisant référence au même référent ont été relevés, il suffit de les relier entre eux pour former la chaine de référence. Ce mémoire s'inscrivant dans le domaine du Traitement Automatique des Langues (TAL), nous nous sommes demandé s'il était possible de détecter automatiquement les chaines de référence.

D'après nos recherches, il s'avère que la détection automatique des chaines de référence est un domaine relativement nouveau et très peu fourni pour le français. En effet, d'après Landragin (2020), avant l'arrivée du projet ANR Democrat en 2016, les seuls systèmes de détection automatique de CR pour le français étaient des systèmes assez anciens, à base de règles comme le système ODACR (Outil de Détection Automatique des Chaines de Référence) développé par Oberlé (2017). La conception à base de règles de ces systèmes les rendait assez peu propices

[2] R = Référent, D1 = Dissociation 1 et D2 = Dissociation 2

aux évolutions et par conséquent, leur robustesse n'était plus acceptable. Ce type de système était de plus conceptuellement dépassé car les recherches actuelles portent sur l'apprentissage autonome et/ou artificiel.

Concernant la langue anglaise, la recherche est plus avancée. Plusieurs systèmes capables de traiter du texte brut à base d'apprentissage, voire de deep-learning, existent. En effet, ces systèmes peuvent s'avérer très utiles pour les moteurs de recherches et par conséquent de grosses entreprises ont investi dans leurs développements (Landragin (2020) et Longo (2014)).

Mais pour le français, en 2016, il n'existait pas de système capable de transcrire les avancées actuelles sur la détection automatique de CR. Il a donc fallu en construire et c'est une des missions que s'est donnée le projet ANR Democrat. Le projet *Description et Modélisation des Chaines de Référence : outils pour l'Annotation de corpus (en diachronie et en langues comparées) et le Traitement automatique* est un projet lancé en mars 2016 et qui s'est achevé en février 2020. Ce projet a été motivé par le besoin d'un modèle théorique de la référence et des chaines de référence, un besoin de données linguistiques attestées, en particulier diachroniques pour l'observation de chaines de référence mais aussi la création d'un corpus de référence en langue française, le besoin d'une plateforme pour la gestion de corpus, de visualisation et de calculs statistiques mais surtout par le besoin d'un système de détection automatique des chaines de référence pour le français. C'est ce dernier point qui va particulièrement nous intéresser. En effet, en 2020, le projet a rendu publiques deux livrables diffusées gratuitement sous licence ouverte concernant le traitement automatique et la détection des chaines de référence : les systèmes COFR et DeCOFR. Ces deux systèmes sont les premiers à réaliser la détection automatique de (co)référence(s) dans des textes tout venant du français.

Le premier système COFR « COreference resolution tool for FRench » s'apparente à une adaptation du système de Kantor & Globerson (2019), conçu pour l'anglais, à la langue française avec un entrainement réalisé sur le corpus Democrat construit lors du projet ANR Democrat. Le second, DeCOFR correspond à l'application de recherches nouvelles sur les architectures de réseaux neuronaux. Pour des raisons d'emploi du temps, ce second système a lui été entrainé sur un corpus antérieur au projet, le corpus ANCOR (Muzerell, 2013).

Cependant, ces systèmes n'offrent pas encore les performances que des linguistes peuvent réaliser avec une annotation manuelle des CR. Pour illustrer, l'outil COFR a été évalué avant sa publication et a obtenu un score de 75% dans la mesure CoNLL (métrique standard d'évaluation des systèmes de résolution de la coréférence). L'outil DeCOFR n'a lui pas fait l'objet de mesure (Landragin, 2020).

Compte tenu des limitations que peuvent rencontrer ces systèmes de détection et du cadre spécifique des écrits utilisés pour ce mémoire, l'exploration et la manipulation des chaines de référence sera réalisée manuellement, à l'aide d'outils informatiques spécialisés.

II - Présentation et recueil des données utilisées

1 - Le corpus RésolCo

Les données qui vont être utilisées dans le cadre de ce mémoire proviennent du corpus RésolCo, constitué dans le cadre du projet E-Calm. Le corpus RésolCo est un ensemble de textes narratifs transcrits et anonymisés d'élèves allant du CE2 au M2 (Université). Il est composé de copies

récoltées en 2015, 2016, 2017 et/ou 2018 à différents niveaux de scolarité chaque année[3]. Ces copies ont été enrichies par des annotations sur les traces du processus lié à l'écriture et sur les variantes orthographiques existantes. L'annotation de ces copies portera également sur les chaines de référence.

RésolCo totalise 394 textes transcrits et normalisés au format numérique. Parmi ces textes, 383 ont été annotés. Le reste des copies est actuellement en cours de traitement. Ces textes sont consultables dans plusieurs formats tels que la version originale ou normalisée de la copie au format texte, sa transcription fidèle à l'original et son annotation au format XML ainsi que les copies originales ou normalisées annotées au format Glozz[4] (.aa et .ac). Pour les copies annotées, plusieurs couches d'annotation sont visualisables comme les traces d'écriture, la normalisation orthographique ou les chaines de référence[5].

La particularité de ce corpus est que tous les textes partagent la même origine : celle d'une seule consigne. La consigne fournie avait pour but de provoquer chez le scripteur la mise en œuvre de stratégies de résolution des problèmes de cohérences soulevés par l'intégration de trois phrases :

- 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.

La consigne indiquait : « *Racontez une histoire dans laquelle vous insérerez séparément et dans l'ordre donné, les trois phrases (ci-dessus). Vous pouvez découper les bandelettes contenant les phrases (ci-dessus) ou bien recopier chaque phrase avec soin à l'identique de celles qui vous sont données* ». Comme le précisent Garcia-Debanc et al. (2017), « *les trois phrases en jeu dans cette consigne impliquent des stratégies discursives variées, amenant le scripteur à gérer plusieurs continuités référentielles et planifier son discours afin d'assurer la cohérence de son texte.* ».

Les écrits récoltés suite à cette consigne ont été traités comme suit[6] :

- Préparation des copies : classement des copies scannées et préparation des fichiers XML dédiés à la transcription.

- Transcriptions et vérifications des transcriptions : deux étapes réalisées indépendamment par deux personnes différentes.

---

[3] La liste des écoles qui ont participé au projet est disponible sur le site : http://redac.univ-tlse2.fr/corpus/resolco/resolco.html.
[4] Se reporter au « Manuel de l'utilisateur » présent sur le site : http://www.glozz.org/
[5] À venir
[6] http://redac.univ-tlse2.fr/corpus/resolco/constitution.html

- Génération des fichiers pour l'annotation des erreurs d'orthographe : application d'un script Python permettant de modifier les fichiers XML traités en format exploitable par Glozz.

- Normalisation orthographique dans Glozz : annotation des erreurs d'orthographe et indication de la version corrigée. Précision importante, si à cette étape les annotateurs rencontrent des erreurs liées à la transcription, il faut recommencer le processus depuis l'étape de transcription afin de corriger les fichiers fautifs.

- Génération des fichiers normalisés : un script Python est appliqué pour récupérer les textes normalisés, et donc dépourvus de formes incorrectes. Les textes restent cependant au format Glozz.

- Vérification des fichiers normalisés : opération réalisée par une personne différente de celle qui a effectué l'annotation. Précision importante, si à cette étape, le vérificateur rencontre des erreurs de normalisation, le processus est recommencé depuis l'étape de normalisation orthographique.

- Parsing des fichiers normalisés : les fichiers .ac sont traités par Talismane (Urieli, 2013) afin de réaliser un étiquetage morphosyntaxique et un parsing.

Ce corpus d'écrits d'élèves est intéressant dans le cadre de ce mémoire car il nous permet deux choses. Premièrement, étant déjà construit, il nous permet de gagner un temps précieux en évitant de récolter nous-même un corpus et ainsi passer rapidement à l'annotation de chaines de référence qui signera véritablement le début de ce mémoire. Secondement, les écrits d'élèves étendus sur une aussi large fourchette de niveaux scolaires peuvent rendre compte de beaucoup de constitutions et de cohabitations de chaines de référence différentes, ce qui en fait des données pertinentes pour notre étude. Ces variations devraient donc permettre de rendre la méthode de détection robuste car la diversité permet de renforcer les processus. Cependant, une trop grande diversité peut également nuire à la construction d'une méthode, nous y serons sensibles dans les manipulations à venir.
Pour terminer, si un ensemble de seulement 383 textes annotés semble relativement faible, l'analyse des chaines de référence et de leurs cohabitations se fera de manière qualitative plutôt que de manière quantitative. De plus, chaque texte présente plusieurs chaines induites par la consigne, ce qui multiplie en réalité leur nombre ainsi que leurs possibilités de cohabitation.

2 – Le logiciel Glozz pour l'annotation de RésolCo

L'annotation de ce corpus est réalisée manuellement à l'aide du logiciel Glozz (version 2.1)[7]. Il s'agit d'une plateforme d'annotation produite pour le projet ANR Annodis (Péry-Woodley et al., 2009) et qui est le fruit de la collaboration de plusieurs laboratoires (CLLE, GREYC et IRIT)[8]. Le projet ANR Annodis avait pour mission la mise en place d'un corpus de référence pour l'analyse de discours. Pour réaliser des analyses sur le corpus, les chercheurs avaient besoin d'une plateforme d'annotation et d'exploration de corpus. Ce n'est pas le premier outil de ce type à être développé car d'autres projets similaires ont vu le jour par le passé et se sont accompagnés de leurs propres outils d'annotation, malheureusement souvent limités par leurs liens aux modèles théoriques ou aux objets linguistiques étudiés. Pour remédier à ces limites, la plateforme Glozz a donc été conçue de façon à proposer un environnement « fortement configurable et non limité *a priori* au contexte discursif dans lequel elle a initialement vu le jour ». (Widlöcher & Mathet, 2014, p. 3). En plus d'offrir cette liberté d'utilisation, la plateforme Glozz offre également une visualisation ergonomique du texte à annoter qui se trouve être proche de la réalité (cf. Illustration 1). Visuellement, c'est comme si l'annotateur annotait un PDF ce qui rend l'activité d'annotation plus confortable.
Ce principe de configuration nous permet d'exploiter un modèle d'annotation de chaines de référence spécifique au corpus RésolCo qui nous permet de mettre en évidence manuellement les éléments correspondant aux maillons de la CR, leur référent commun ainsi que les liaisons entre les éléments, le tout coloré et visualisable en fonction des préférences de l'annotateur.

L'annotation en chaines de référence du corpus RésolCo ne porte à ce jour que sur les trois référents les plus saillants donnés par la consigne, à savoir : « elle », « il » et « les enfants ». Voici l'illustration de ce que permet de faire l'interface Glozz avec, encadrées en rose, les phrases de la consigne annotées préalablement (visualisation désactivable).

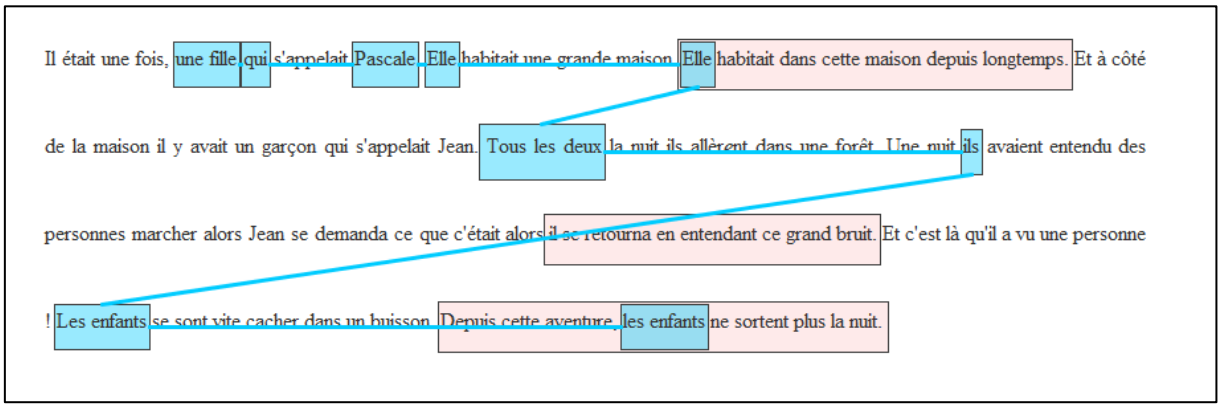


Illustration 1 - CR_elle dans Glozz

[7] Le site http://www.glozz.org/ indique que la dernière version du logiciel est la version 2.0.1, datant de juillet 2014 mais la version distribuée actuellement, accessible via le formulaire de téléchargement, et celle que nous utilisons est la version 2.1.
[8] CLLE : *Cognition, Langues, Langage, Ergonomie*, Université de Toulouse UTM
GREYC : *Groupe de Recherche en Informatique, Image et Instrumentation de Caen*, Université de Caen
IRIT : *Institut de Recherche en Informatique de Toulouse*, Université de Toulouse UPS

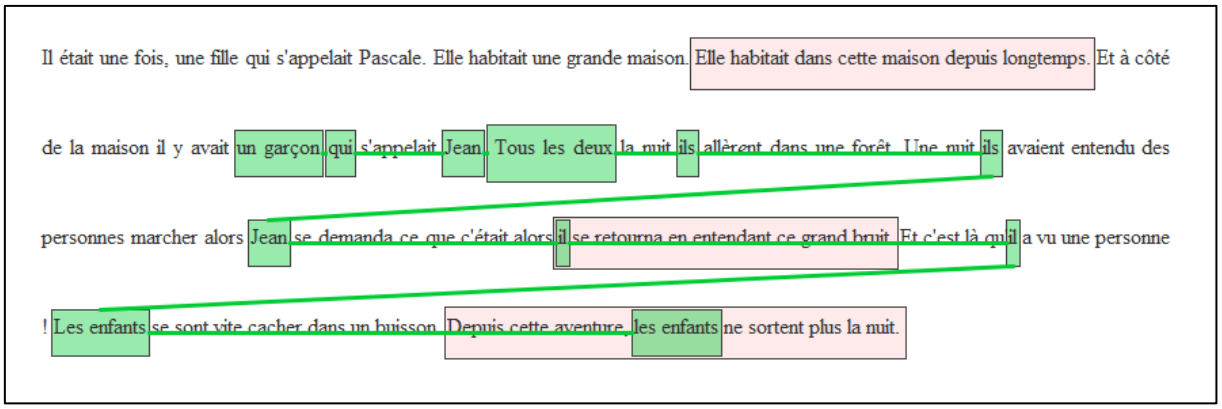


Illustration 2 - CR_il dans Glozz

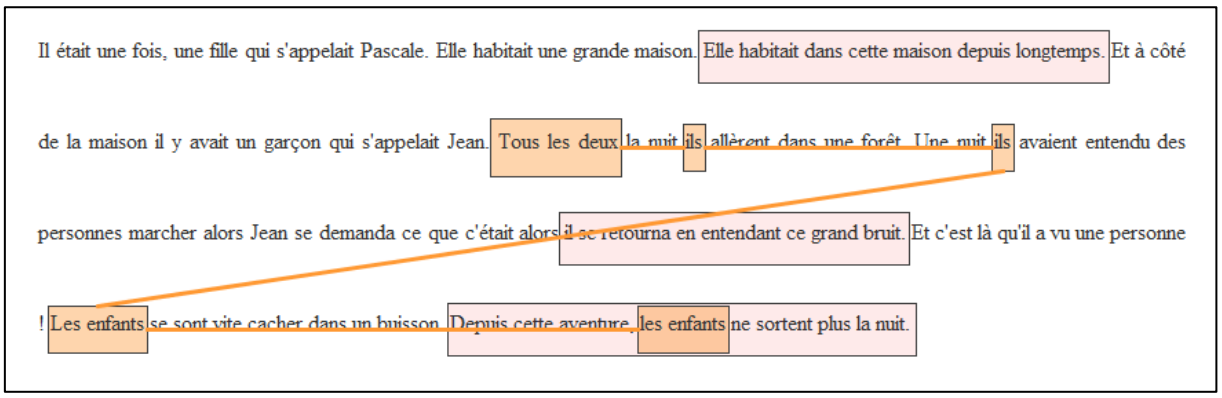


Illustration 3 - CR_les enfants dans Glozz

Il est également possible de visualiser la superposition des chaines comme sur l'Illustration 4. Les couleurs se mélangent et deviennent difficiles à distinguer mais cela donne un aperçu des relations existantes.

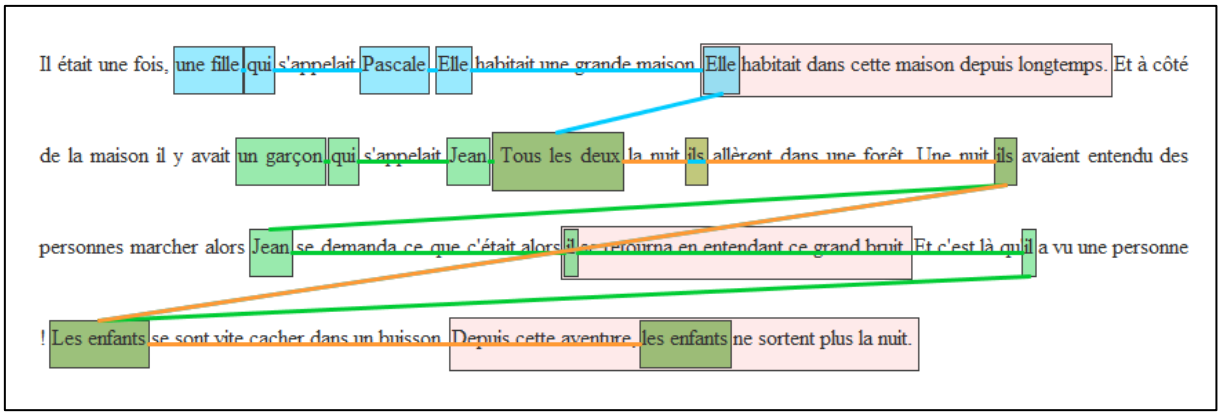


Illustration 4 - Les 3 CR dans Glozz

Chaque maillon ainsi annoté comporte plusieurs informations (Illustration 5). On y retrouve le référent associé aux maillons ainsi que des précisions :

- « Groupe » permet de spécifier si le référent, ici « elle », est intégré dans un autre référent, par exemple « Tous les deux » dans l'Illustration 4.

- « Incertitude sur la délimitation » permet de spécifier si l'annotateur a exprimé un doute sur la délimitation du maillon.

- « Incertitude sur le rattachement » permet de spécifier si l'annotateur a exprimé un doute sur le rattachement du maillon sélectionné à une chaine de référence.

- « Commentaire » permet de laisser un commentaire (facultatif) sur l'annotation réalisée. Ce commentaire pourra être lu par un autre annotateur lors de l'étude des annotations et pourra permettre d'aider à la correction de l'annotation ou à sa compréhension. De plus, ces commentaires peuvent être utilisés pour indiquer des phénomènes mentionnés dans le guide d'annotation (voir la section suivante) mais pour lesquels il n'y a pas de modèle d'annotation prédéfini comme par exemple, un membre du référent « les enfants » qui n'est ni associé au référent « elle », ni au référent « il » mais qui est intéressant à mettre en valeur via l'annotation.

- L'onglet « Informations » garde les données du type d'annotation : coref_Elle pour un maillon de la CR « elle » par exemple, l'identifiant de l'auteur de l'annotation, la date et l'heure à laquelle elle a été réalisée, sa dernière modification et la date de la dernière modification.

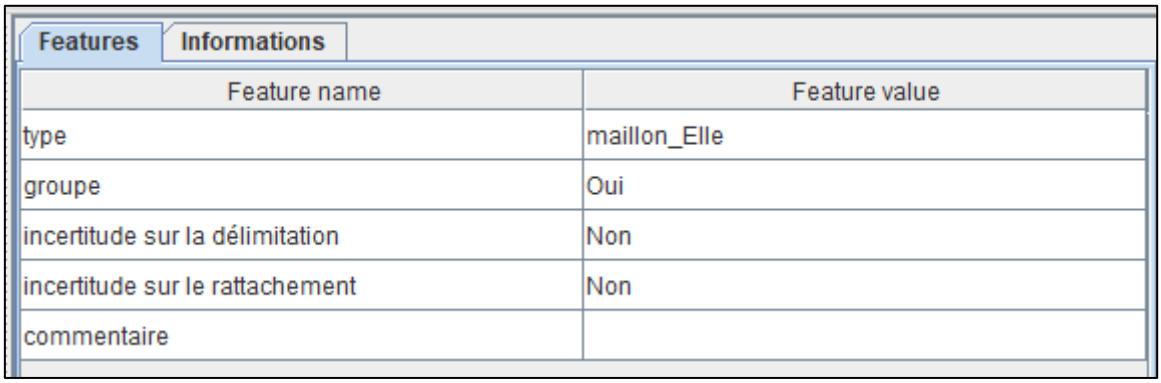


Illustration 5 - Caractéristiques des maillons dans Glozz

3 - L'annotation dans RésolCo

Nous venons de voir que l'outil Glozz permet d'annoter mais qu'est-ce qui est véritablement annoté dans le corpus ? La tâche d'annotation dans le corpus RésolCo est celle de l'annotation des **marques référentielles** qui se trouvent sous la forme d'éléments linguistiques. D'après le guide d'annotation de RésolCo[9], il est possible de transformer les éléments suivants en **maillons** (se reporter au guide d'annotation pour plus de précisions concernant l'annotation) :

- **Les syntagmes nominaux** : s'ils sont accompagnés de modifieurs, ils doivent être inclus dans la délimitation du maillon.
  Par exemple : *Une vieille femme* constitue un maillon complet, avec la prise en compte du modifieur souligné « vieille ».

[9] Guide d'annotation de RésolCo : https://hodaclm.github.io/resolco/

- **Les possessifs** : les expressions possessives coréférent au possesseur, elles sont donc annotées, cependant, seul le déterminant possessif est considéré comme maillon.
  Par exemple : *Elle gare sa voiture*. Le déterminant possessif « sa » référant au pronom « elle » sera considéré comme un maillon.

- **Les pronoms** : il faut annoter les pronoms personnels, démonstratifs, indéfinis, relatifs et corrélés mais pas les pronoms réfléchis.
  Par exemple : *Les enfants entendent un bruit. Ils sortent dans le jardin.* Le pronom personnel « ils » est annoté comme maillon de la chaine « les enfants ».

- **Les pronoms relatifs** : lorsque le nom ou le groupe nominal composant un maillon est modifié par une proposition relative, seul le pronom relatif est à considérer comme un nouveau maillon, et doit être annoté. Le nouveau maillon est donc dans ce cas composé du seul pronom relatif.
  Par exemple : *La jeune fille qui habitait dans le village.* Toute la proposition suivant le pronom relatif « qui » n'est pas annotée, seul le pronom l'est.

- **Les noms propres :** lorsque l'expression référentielle prend la forme d'un nom propre (NPP), elle est systématiquement annotée, sauf dans les cas de didascalies nominatives.
  Par exemple : *Jade est pilote de rallye*. Le NPP « Jade » sera annoté.

- **Sujet zéro** : il est fréquent dans les cas de coordination verbale que le sujet ne soit pas répété pour chaque verbe mais uniquement instancié en sujet du premier verbe ("Elle est venue, a vu et a vaincu" *vs.* "Elle est venue, elle a vu et elle a vaincu"). Dans ces contextes, la seule façon d'annoter le sujet de tous les verbes est d'annoter le groupe verbal.
  Par exemple : *Johanna se prépara, alla à l'école et donna cours*. Dans cet exemple, le sujet des verbes soulignés n'est pas répété. La seule façon d'annoter la référence est d'annoter le groupe verbal (en incluant les négations, les auxiliaires et certains pronoms s'ils sont présents).

- **Référents multiples :** dans le cas de référents multiples, désignés par des pronoms personnels pluriels, des SN pluriels (*les deux frères*), collectifs (*le groupe de copines*), deux situations sont à distinguer :
  - Le référent multiple ne réfère pas à « les enfants » et inclut un des référents de la consigne.
    Par exemple : *Maxime est en salle de classe avec les autres étudiants. Ils passent un partiel important.* Disons que « Maxime » fait référence au référent singulier masculin « il » de la consigne, le maillon « Ils » sera alors annoté comme référant à « il » mais prendra la valeur « Oui » dans la caractéristique « groupe ».

  - Le référent multiple réfère à « les enfants » et inclut un des référents, ou les deux, de la consigne.
    Par exemple : *Marion est la sœur de Sam. Ce sont les enfants de leurs parents.*
    Si « Marion » fait référence au référent « elle » et « Sam » au référent « il » et que « les enfants » fait référence au maillon « les enfants » de la consigne, ce dernier sera annoté trois fois. Une première fois avec le type « les enfants » et la

valeur de « groupe » valant « Non », puis une deuxième fois avec le type « il » et la valeur de « groupe » valant « Oui » et une dernière fois avec le type « elle » et la valeur de « groupe » valant « Oui ». Trois annotations sont donc superposées sur la même unité linguistique.

- **Le discours direct** : si le scripteur a inséré du discours direct dans son texte, il n'y a pas de rupture de la continuité référentielle, il faut annoter normalement les CR.
  Par exemple : *Jean – « Pourquoi es-tu si triste ? »*. Dans ce cas, « tu » sera annoté s'il réfère à un des trois référents de la consigne. « Jean » étant une didascalie nominative, il ne sera pas annoté.

- **Le titre** : s'il y a des marques de continuité référentielle dans le titre, il faut les annoter.
  Par exemple : *L'histoire de la vie de Laura*. Dans ce cas, si « Laura » réfère à un des référents de la consigne, le nom sera annoté.

Remarque complémentaire concernant l'annotation

Lorsque l'élève n'a pas respecté la phrase consigne uniquement en termes de genre ou de nombre sur les référents (Exemple : Ils habitaient dans cette maison depuis longtemps), la copie est écartée mais sauvegardée dans un dossier à part. Cela permet d'éviter de biaiser les analyses quantitatives avec des copies trop éloignées des autres. Cependant, si les phrases consignes sont en désordre ou modifiées mais que le référent est correct en nombre et en genre, la copie est traitée comme les autres.

4 – Mon expérience d'annotation

Je n'avais jamais annoté de copies auparavant. J'ai donc découvert l'annotation ainsi que le logiciel d'annotation Glozz en même temps. Cela n'a pas été tâche aisée mais ce fut bénéfique car en annotant, j'ai pu me rendre compte de phénomènes particuliers. En annotant des copies du CE2 au M2, je me suis rendu compte que les liens de cohérence faiblissaient à mesure que le niveau scolaire s'affaiblissait. Pourtant, les copies produites au début de la scolarité me semblant plus courtes, leur cohérence aurait dû être plus aisée. Mais il ne faut pas oublier que ces copies de primaire sont produites par de jeunes apprenants qui ne maitrisent pas encore la langue française, ce qui peut perturber le maintien de la cohérence dans les écrits.
Lors de cette étape, j'ai également commencé à repérer quelques cohabitations parmi celles que je présentais dans l'état de l'art (section I-2), telles que la succession, la dissociation et la fusion.
Lors de l'annotation, j'ai également mesuré le temps que me prenait chaque copie. La moyenne du temps passé sur chacune des copies se situe entre 10 et 15 minutes mais cette moyenne représente mal la distribution des différents temps passés. Certaines copies, comme les copies de CE2, ne possédant que peu de texte et donc peu de maillons sont traitées plus rapidement que les copies plus généreuses et complexes de Master 2. De fait les copies de Master 2 peuvent aisément requérir une demi-heure de traitement.

Le logiciel Glozz est simple d'utilisation et sa visualisation simple et colorée ne décourage pas lors de l'annotation. Grâce à ça, l'expérience n'aura pas été désagréable et s'est même révélée primordiale pour la suite de ce mémoire. Je peux affirmer sans le moindre doute que sans l'annotation, je n'aurais pas pris la même direction que celle prise dans ce mémoire car cette exploration et cette manipulation des données m'a permis d'ancrer le sujet dans la réalité de la production d'écrits d'élèves.

5 – Changement de terme : la continuité référentielle

Lors de mon annotation, je me suis rendu compte que l'annotation réalisée dans le cadre du corpus RésolCo n'est pas strictement celle des chaines de référence évoquées dans l'état de l'art. En effet, il m'a semblé que le terme « chaine de référence » était un peu trop étendu dans le cadre de cette annotation. Le terme le plus adapté pour discuter du type d'annotation réalisée serait alors celui de « continuité référentielle » décrit par Garcia-Debanc et al. (2021) comme visant « l'identification des formes linguistiques utilisées par les élèves pour construire la **cohésion référentielle** d'un texte ». L'annotation de ces formes linguistiques (section II-3) diffère donc de celle qui ne s'appliquerait qu'aux éléments référentiels « directs », c'est-à-dire, ceux de la chaine de référence. En effet, bien que la notion de chaine de référence ait pour objectif, comme l'écrit Schnedecker (citée dans Garcia-Debanc et al., 2021), d' « appréhender le déroulement ou encore le suivi de l'expression référentielle dans sa continuité textuelle ou discursive », la définition que nous en avons trouvé dans la littérature semble trop contrainte pour être le terme désignant le type d'annotation réalisée sur le corpus.

Ainsi Garcia-Debanc et al. (2021) privilégient l'utilisation du terme de « continuité référentielle qui s'inspire des travaux de Givón (1983) ». Ce terme semble plus adapté dans la mesure où « les phénomènes en jeu ne renvoient pas toujours strictement aux phénomènes de coréférence et/ou d'anaphore qui caractérisent les chaines de référence. » Ce choix d'annoter, dans RésolCo, les continuités référentielles et non pas seulement les référents a également été motivé par la volonté d'inclure les « continuités référentielles qui s'établissent […] entre les référents. […] Ces continuités ont pour fonction de « tisser des liens » entre une grande partie des propositions qui composent le texte, contribuant ainsi à construire ce que Halliday & Hasan (cités par Garcia-Debanc et al., 2021) appellent la « texture » [du texte] ».

Afin de respecter au mieux l'annotation réalisée sur le corpus, nous utiliserons dorénavant le terme de « continuité référentielle » en lieu et place de « chaine de référence ». L'abréviation du terme reste toujours « CR ».

III – Modélisations théoriques des cohabitations de continuités référentielles

Après nos observations sur la constitution du corpus RésolCo et son annotation, nous allons maintenant passer à l'étude des cohabitations des continuités référentielles. Dans un premier temps, nous allons présenter les deux modèles théoriques de cohabitations que nous avons sélectionnés dans le cadre de ce mémoire. Dans un deuxième temps, nous présenterons la modélisation des relations possibles entre les maillons qui illustrent les continuités référentielles dans les textes annotés du corpus RésolCo. Dans un troisième et dernier temps, nous appliquerons les deux profils de cohabitation sélectionnés aux textes constituant le corpus.

Dans le cadre de ce mémoire, nous avons décidé de ne garder que deux types de cohabitations : **la succession et l'association/dissociation**. D'autres modèles existent cependant, notamment le parallélisme, mais ne pouvant tous les étudier dans le temps imparti, nous avons décidé de ne sélectionner que deux types, de conceptions profondément différentes, nous permettant ainsi de les comparer. Cette comparaison devrait permettre d'appréhender le contraste entre deux complexités « extrêmes » de la gestion des référents : le « pas à pas » face à la « simultanéité

». Les modélisations de ces deux types de cohabitations, et de leurs formes particulières, sont présentées dans l'ordre que nous jugeons du plus simple au plus complexe.

1 – La succession

La succession est pour nous le niveau le plus « simple » de la gestion des référents avec une stratégie de traitement « pas à pas ». C'est-à-dire que les référents sont traités les uns après les autres, sans se croiser. Cette stratégie ne nécessite donc que la manipulation d'un seul référent à la fois lors de l'écriture.

A - Succession stricte

C'est la cohabitation la plus simple qui puisse exister. Elle illustre le traitement « pas à pas » des référents. C'est une cohabitation que nous avons pu observer pendant notre annotation.

Définition :

Soit une CR A, une CR B, une CR C et une CR D, composées de marques w, x, y et z :

➔ $A_0, A_1, \ldots, A_w, B_0, B_1, \ldots, B_x, C_0, C_1, \ldots, C_y, D_0, D_1, \ldots, D_z$

Une CR ne peut pas contenir de marques de continuité référentielle associées à d'autres référents. La première et la dernière marque d'une CR ne peuvent pas être en contact avec une marque d'une autre CR.

Modélisation :

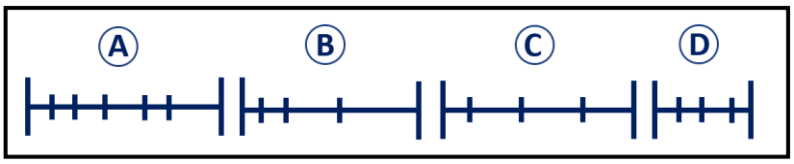


Illustration 6 - Modélisation de la succession stricte

Remarque : sur ce modèle, quatre continuités référentielles sont représentées sous la forme linéaire. Il faut alors imaginer le texte comme une longue ligne déroulée et non pas comme un empilage de lignes comme nous avons l'habitude de le visualiser.
Ainsi, la première barre verticale à gauche représente la première marque de CR d'un référent, les petites barres qui suivent sont les marques suivantes dans la CR et la dernière barre à droite, représente la dernière marque existante de CR d'un référent. Ces barres verticales sont donc des représentations abstraites des éléments linguistiques présents dans un texte, qui portent en eux une marque de continuité référentielle par rapport à un référent donné.

B - Succession chevauchée minimale groupée

Cette cohabitation est théoriquement possible mais nous ne l'avons pas observée de manière empirique**.**

Définition :

Soit une CR A, une CR B, une CR C et une CR D, composées de marques w, x, y et z :

➔ $A_0, A_1, \ldots, A_{w-1}, [A_w + B_0], B_1, \ldots, B_{x-1}, [B_x + C_0], C_1, \ldots, C_{y-1}, [C_y + D_0], D_1, \ldots, D_z$

Remarque : la notation […] représente **le regroupement**, c'est à dire que deux marques référentielles se confondent dans la même unité linguistique.

Une CR ne peut pas contenir de marques de CR associées à d'autres référents, sauf pour sa première et sa dernière marque qui doivent être **regroupées.** La première marque d'une CR doit être regroupée avec la dernière marque de la CR antérieure, s'il y en a une et la dernière marque de la CR doit être regroupée avec la première marque de la CR suivante, s'il y en a une. De fait, la première marque de la première CR ne peut pas être regroupée car il n'existe pas de marque antérieure et la dernière marque de la dernière CR ne peut pas être regroupée car il n'existe pas de marque postérieure.

Modélisation :

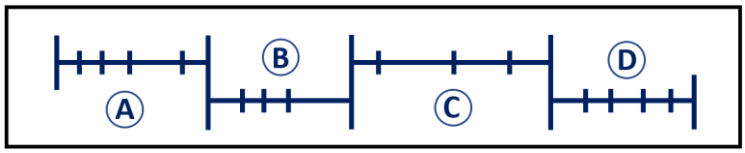


Illustration 7 - Modélisation de la succession chevauchée minimale groupée

Remarque : afin de faciliter la visualisation, la modélisation des cohabitations fera désormais alterner une CR en « haut » et une CR en « bas » mais le format linéaire reste identique à celui décrit dans la modélisation de la succession stricte.

C - Succession chevauchée minimale distincte

Cette cohabitation est théoriquement possible mais nous ne l'avons pas observée de manière empirique**.**

Définition :

Soit une CR A, une CR B, une CR C et une CR D, composées de marques w, x, y et z :

➔ A_0, A_1, …, A_{w-1}, B_0, A_w, B_1, …, B_{x-1}, C_0, B_x, C_1, …, C_{y-1}, D_0, C_y, D_1, …, D_z

Une CR ne peut pas contenir de marques de CR associées à d'autres référents, sauf entre son avant-dernière et sa dernière marque, ce qui implique que la dernière marque de la CR se trouve entre la première et la deuxième marque de la CR postérieure. **Aucune** des marques d'aucune des CR ne peut être **regroupée/en contact** avec une autre.

Modélisation :

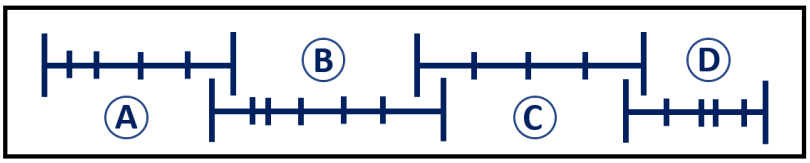


Illustration 8 - Modélisation de la succession chevauchée minimale distincte

D - Succession chevauchée

Nous jugeons cette forme comme la plus complexe de la cohabitation de succession. Cette cohabitation est théoriquement possible mais nous ne l'avons pas observée de manière empirique, même si sa modélisation nous a été inspirée par nos observations.

Définition :

Remarque : à partir d'ici, il devient difficile de proposer une formule générale comme pour les formes précédentes car la position des marques ainsi que leurs relations deviennent moins contraintes. Pour pallier cet inconvénient, nous proposerons une formule figée avec sa modélisation que nous détaillerons.

Soit quatre CR A, B, C et D :

➔ A_0, A_1, A_2, [A_3 + B_0], [A_4 + B_1], A_5, B_2, B_3, B_4, B_5, C_0, B_6, C_1, C_2, [C_3 + D_0], D_1, D_2, D_3, D_4

Une CR doit contenir la présence d'une ou plusieurs marques de la CR postérieure mais ne peut pas la contenir en intégralité, sauf pour la dernière CR. Cette dernière n'ayant pas de CR postérieure, elle doit contenir la présence d'une ou plusieurs marques de la CR antérieure mais ne peut pas la contenir en intégralité.
Toutes les marques des CR *peuvent* être regroupées à une autre tant que la première condition est remplie.

Attention : si le premier et le dernier maillon de deux CR différentes se trouvent regroupés comme dans Illustration 9 - Cas particulier de deux CR contenues mutuellement, il n'y a pas de succession chevauchée car les CR se contiennent mutuellement.

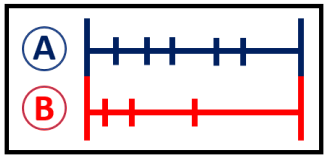


Illustration 9 - Cas particulier de deux CR contenues mutuellement

Modélisation :

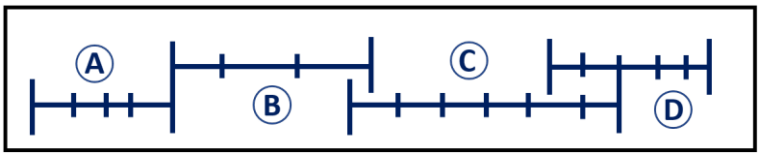


Illustration 10 - Modélisation de la succession chevauchée

L'Illustration 10 tente de figurer les différentes relations qui peuvent se produire entre les CR composant la succession chevauchée. De gauche à droite :

- Entre la CR A et B, nous pouvons observer le même rapport entre la dernière marque de la CR A et la première marque de la CR B que celui décrit dans la cohabitation de la succession chevauchée minimale groupée. Ce rapport unique ne permet pas de définir la cohabitation entre les 4 CR présentes ici mais illustre que la succession chevauchée accepte ce type de « transition ».

- Entre la CR B et la CR C, nous pouvons observer le même rapport entre la dernière marque de la CR B et la première marque de la CR C que celui décrit dans la cohabitation de la succession chevauchée minimale distincte : en effet, la première marque de la CR C se situe entre l'avant-dernière et la dernière de la CR B. Ce rapport

unique ne permet pas de définir la cohabitation entre les 4 CR présentes ici mais illustre que la succession chevauchée accepte ce type de « transition ».

- Entre la CR C et la CR D, nous pouvons observer un chevauchement comme celui décrit dans la définition de la succession chevauchée. La CR C et la CR D contiennent mutuellement des marques de l'une et de l'autre, sans jamais contenir l'intégralité de l'une, ni de l'autre. De plus, les marques 5 et 6 de la CR C (C_5 et C_6) sont regroupées avec, respectivement, les marques 1 et 2 et de la CR D (D_1 et D_2). Pour reprendre la notation précédente, de façon ciblée, cela donne : D_0, $[C_5 + D_1]$, $[C_6 + D_2]$, D_3

Afin de compléter la définition de la succession chevauchée, nous allons illustrer avec deux autres exemples qui permettront de mieux saisir cette cohabitation : l'absence de succession chevauchée et son étendue maximale.

Illustrations complémentaires pour la succession chevauchée

Absence de succession chevauchée

Définition :

Soit quatre CR A, B, C et D :

➔ A_0, B_0, A_1, A_2, A_3, A_4, B_1, B_2, B_3, C_0, D_1, $[C_1 + D_1]$, $[C_2 + D_2]$, D_3, C_3, D_4, D_5, D_6

Modélisation :

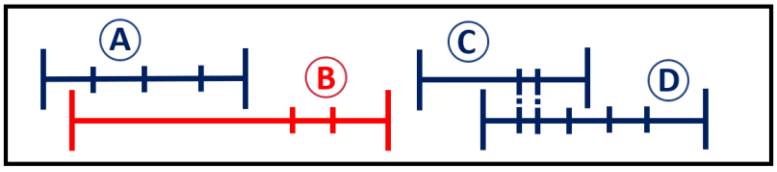


Illustration 11 - Modélisation de l'absence de succession chevauchée

Dans ce cas, nous ne pouvons pas dire que la cohabitation présentée est celle d'une succession chevauchée, ni d'aucune de toutes celles présentées dans cette partie. En effet, la CR B, représentée en rouge, ne contient pas de marque de la CR postérieure (CR C). La condition pour qu'il y ait une cohabitation dite de « succession chevauchée » n'est donc pas remplie.

Succession chevauchée d'étendue maximale

Définition :

Soit quatre CR A, B, C et D :

➔ A_0, B_0, A_1, $[A_2 + B_1]$, $[A_3 + B_2]$, A_4, C_0, B_3, B_4, C_1, B_5, D_0, $[C_2 + D_1]$, D_2, D_3, C_3, D_4

Modélisation :

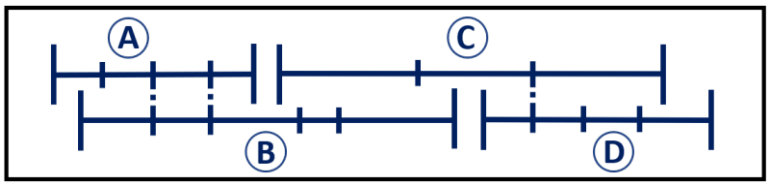


Illustration 12 - Modélisation de la succession chevauchée d'étendue maximale

Dans ce cas, la succession chevauchée est d'étendue maximale. En effet, d'après la définition, une CR ne peut pas en contenir intégralement une autre. Ici, la première marque de la CR A (A_0) est en dehors de la CR B. Cette seule marque permet donc de respecter la définition car la CR B ne contient pas intégralement la CR A. Le schéma inverse est répété entre la CR C et la CR D car la dernière marque de la CR D se trouve en dehors de la CR C, qui par conséquent ne la contient pas intégralement.

Ces illustrations complémentaires concluent les modélisations de succession. La liste n'est pas exhaustive et d'autres formes spécifiques peuvent exister sous cette appellation mais dans le cadre de ce mémoire, nous nous limiterons à celles que nous venons de décrire. Nous allons maintenant passer au second type de cohabitation que nous voulons étudier : l'association/dissociation.

2 - L'association/dissociation

L'étude de ce type de cohabitation, ainsi que celle de ses formes spécifiques, a été motivée par l'observation de certains comportements des marques de continuités référentielles que nous avons pu faire lors de notre annotation sur le corpus. De plus l'association/dissociation (A/D) est pour nous un des niveaux les plus « complexes » de la gestion des référents avec une stratégie de traitement « prévisionnelle » ou « simultanée ». C'est-à-dire, qu'à l'inverse de la succession, les référents sont traités en parallèle les uns des autres, peuvent se croiser, et nécessitent donc la manipulation d'au moins deux référents en « simultané » lors de l'écriture.

Pour commencer la modélisation de ce type de cohabitation, nous proposons de commencer par la forme d'association/dissociation que nous trouvons la plus simple à appréhender, puisqu'elle ne possède pas de dissociation : **la fusion**.

Remarque : dans cette partie concernant l'association/dissociation, la modélisation en alternance n'est plus d'actualité. En effet, elle permet difficilement de rendre compte des phénomènes décrits. La linéarité est cependant conservée et s'accompagne d'un « zoom » permettant d'illustrer au mieux les relations entre les marques des CR.

A - Fusion

Nous avons pu observer cette cohabitation pendant notre annotation. Nous voudrions attirer ici l'attention du lecteur sur le fait que nous utilisons le même terme que celui donné dans l'état de l'art : « fusion ». Cependant, nous n'en avons pas tout à fait la même définition.

Définition :

Soit quatre CR A, B, C et D :

➔ A_0, A_1, A_2, A_3, A_4, A_5, A_6, B_0, B_1, B_2, B_3, B_4, C_0, B_5, C_1, C_2, C_3, D_0, C_4, C_5, $[C_6 + D_1]$, $[C_7 + D_2]$

Une marque de continuité référentielle se confond avec une autre marque associée à un autre référent afin de former un groupe. Ce groupe est alors composé d'au moins deux marques associées à deux référents différents mais il n'y a pas de limitation quant au nombre maximal de marques confondues dans le même groupe. Ces marques maintenant regroupées ne se quittent plus jusqu'à la fin du texte.

Modélisation :

Vue d'ensemble :

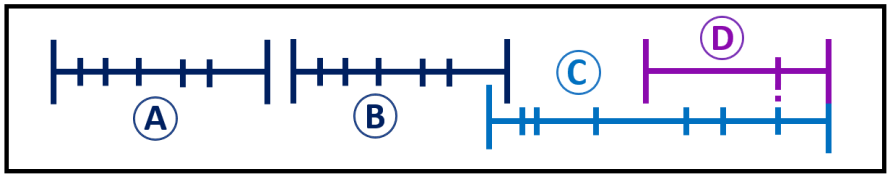


Illustration 13 - Modélisation de la fusion

Zoom :

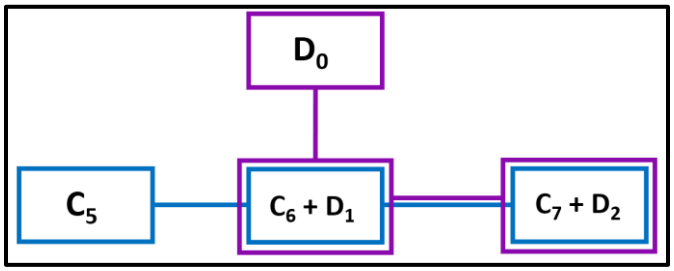


Illustration 14 - Zoom sur la fusion

Note complémentaire

Ce phénomène peut se produire plusieurs fois. À chaque groupe formé peut s'ajouter une nouvelle marque associée à autre référent, et ce de façon quasi infinie, uniquement limitée par le nombre total de référents. Plusieurs marques associées à différents référents peuvent se joindre en même temps dans le même groupe.

Soit quatre CR A, B, C et D :

➔ A_0, A_1, A_2, A_3, C_0, A_4, A_5, C_1, A_6, B_0, A_7, C_2, A_8, B_1, A_9, C_3, A_{10}, C_4, C_5, A_{11}, B_2, C_6, D_0, C_7, B_3, C_8, B_4, $[B_5 + C_9 + D_1]$, $[B_6 + C_{10} + D_2]$

Modélisation :

Vue d'ensemble :

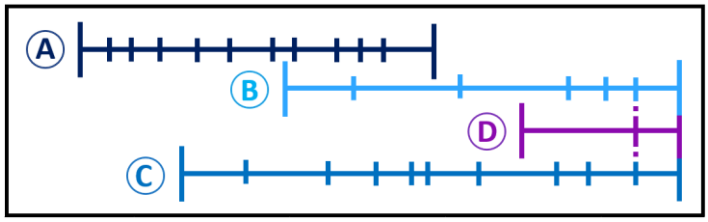


Illustration 15 - Modélisation de la fusion multiple

Zoom :

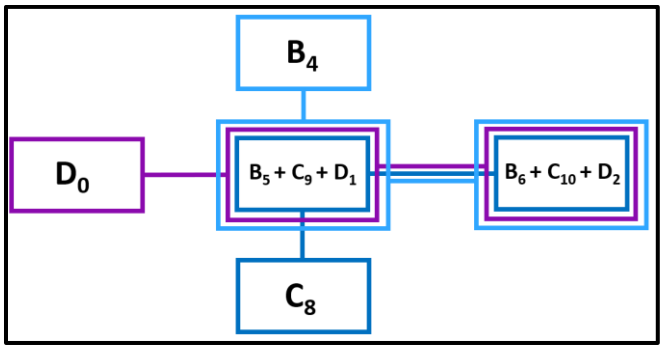


Illustration 16 - Zoom sur la fusion multiple

B - Association/dissociation

Cette cohabitation est la seconde plus complexe de toutes celles que nous présentons. En effet, elle nécessite la manipulation simultanée d'au moins deux référents. Nous avons pu observer cette cohabitation pendant notre annotation.

Définition :

Soit quatre CR A, B, C et D :

➔ A_0, A_1, A_2, A_3, B_0, A_4, $[A_5 + B_1]$, A_6, C_0, $[B_2 + C_1]$, $[B_3 + C_2]$, D_0, B_4, C_3, C_4, C_5, C_6, D_1, D_2

Une marque de continuité référentielle se confond avec une autre marque associée à un autre référent afin de former un groupe. Ce groupe est alors composé d'au moins deux marques associées à deux référents différents mais il n'y a pas de limitation quant au nombre maximal de marques confondues dans le même groupe. Après une période minimale de regroupement d'au moins un maillon, l'une des marques - ou plusieurs - se dissocie du groupe précédemment formé. Ce phénomène peut se produire un nombre de fois illimité dans le texte.

Modélisation :

Vue d'ensemble :

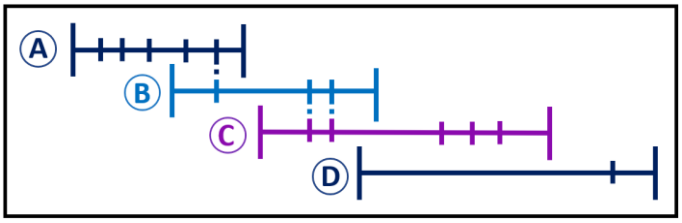


Illustration 17 - Modélisation de l'association/dissociation

Zoom :

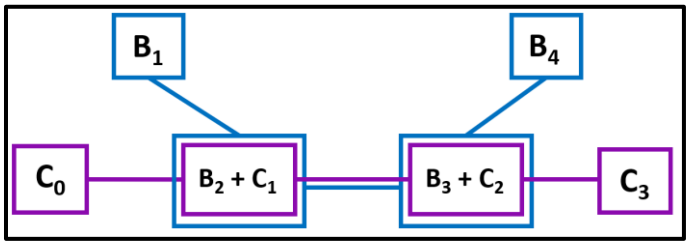


Illustration 18 - Zoom sur l'association/dissociation

C - Association/dissociation puis fusion

Il est possible de cumuler association/dissociation et fusion si au cours d'un texte se produit au moins une **association/dissociation** puis une **fusion**. Cette cohabitation est la plus complexe de toutes celles que nous présentons. En plus de nécessiter une gestion simultanée d'au moins deux référents, elle les contraint à fusionner. Nous avons pu observer cette cohabitation pendant notre annotation.

Définition :

Soit quatre CR A, B, C et D :

➔ A_0, A_1, A_2, A_3, B_0, B_1, A_4, A_5, B_2, C_0, B_3, C_1, B_4, B_5, C_2, B_6, D_0, C_3, **[A_6 + D_1]**, A_7, D_2, D_3, **[A_8 + D_4], [A_9 + D_5]**

Modélisation :

Vue d'ensemble :

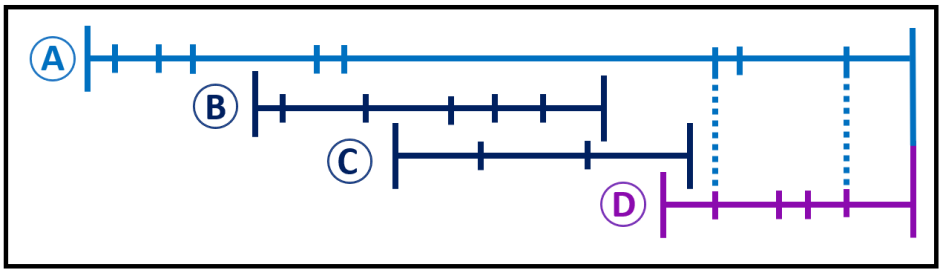


Illustration 19 - Modélisation de l'association/dissociation puis fusion

Zoom :

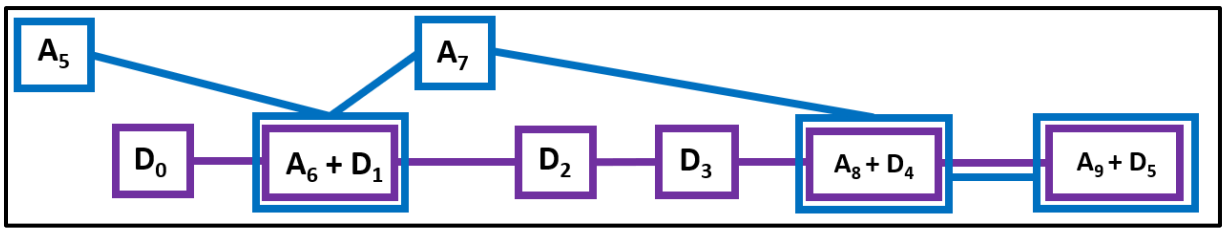


Illustration 20 - Zoom sur l'association/dissociation puis fusion

IV – Application des modélisations théoriques au corpus RésolCo

1 – Les relations entre les maillons des CR

Les modèles présentés dans la partie précédente traitent des marques de continuité référentielle. Dans le corpus RésolCo, l'annotation des textes permet de visualiser ces marques à l'aide de **maillons**. Dans cette partie, nous allons donc décrire les différentes relations possibles entre les maillons associés aux trois référents donnés par la consigne. Pour la suite, cette association au référent sera représentée par un cadre de couleur autour d'une unité linguistique. Il existe trois couleurs pour les trois référents donnés par la consigne : bleu pour le référent « elle », vert pour le référent « il » et enfin, orange pour le référent « les enfants ».

Pour finir, dans la partie précédente, nous discutions de la possibilité que deux marques, ou deux maillons donc, soient regroupées. Le regroupement peut ici prendre deux formes : **l'inclusion** ou **la superposition**.
Si nous considérons les maillons comme des boites, l'inclusion se produit quand une « boite » est contenue dans une autre « boite » plus grande (plus longue). La superposition, quant à elle, se produit quand les deux « boites » ont exactement les mêmes limites (longueur). Par opposition au regroupement, une « boite » peut être seule. C'est-à-dire qu'elle ne contient pas de « boite », n'est pas contenue par une « boite » et n'est pas superposée à une « boite ».

Afin d'illustrer nos propos, nous avons réalisé des arbres qui reprennent les relations possibles entre les différents maillons étudiés. Pour compléter ces arbres, nous proposons des mises en situation à partir de phrases inventées. Pour mieux les comprendre, voici la légende :

Inclus : contenu dans → Il parlait de sa sœur.
Superposé : mêmes limites → Mathilde et Joseph étaient jeunes. Ils s'aimaient.

Elle : maillon « elle »
Il : maillon « il »
Les E : maillon « les enfants »

Nous allons commencer par les relations possibles entre les maillons associés au référent « elle » :

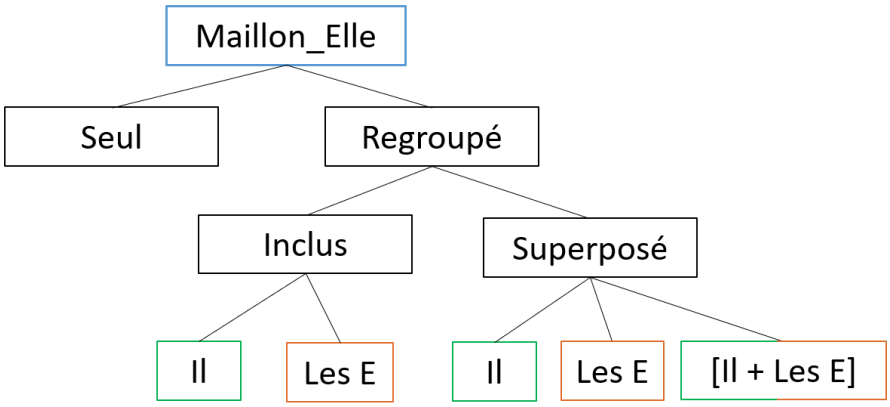


Illustration 21 - Relations possibles des maillons associés au référent "elle"

Mises en situation

Seul : Elle allait au marché.

G, I, Il : Elle parlait de son frère.

G, I, Les E : Les enfants, dont Clara, savaient qui était le loup.

G, S, Il : Mathilde et Joseph étaient jeunes. Ils s'aimaient.

G, S, Les E : Les amis de Clara la rejoignirent. Ensemble, ils allèrent à la rivière.

G, S, [Il + Les E] : Mathilde avait rejoint ses enfants dans la voiture. Le père démarra.

La famille roulait sur la R95.

S'ensuivent les relations possibles entre les maillons associés au référent « il », qui sont les mêmes que pour le référent « elle » avec une inversion des référents nécessaire :

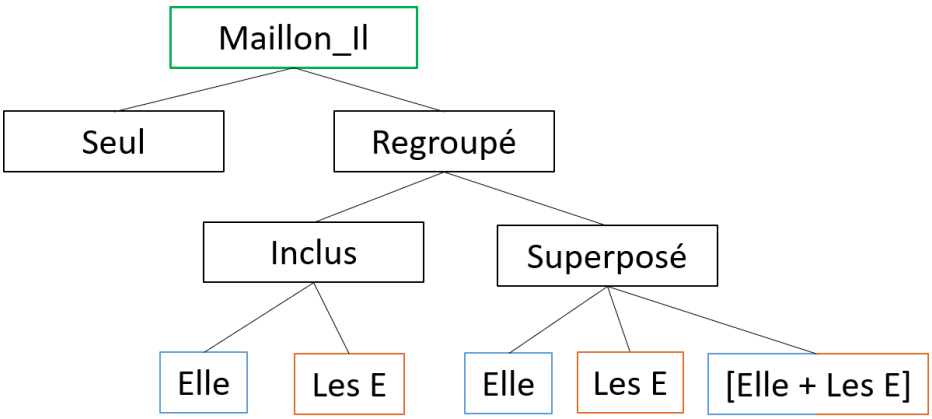


Illustration 22 - Relations possibles des maillons associés au référent "il"

Mises en situation

Seul : Il allait au marché.

G, I, Elle : Il parlait de sa sœur .

G, I, Les E : Les enfants, dont Jacques, savaient qui était le loup.

G, S, Elle : Mathilde et Joseph étaient jeunes. Ils s'aimaient.

G, S, Les E : Les amis de Michel le rejoignirent. Ensemble, ils allèrent à la rivière.

G, S, [Elle + Les E] : Mathilde avait rejoint ses enfants dans la voiture. Le père démarra.

La famille roulait sur la R95.

Ce dernier arbre représente les relations possibles entre les maillons associés au référent « les enfants » :

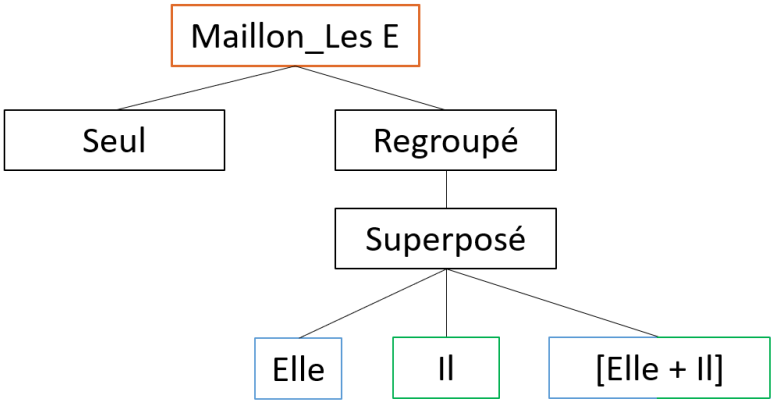


Illustration 23 - Relations possibles des maillons associés au référent "les enfants"

Mises en situation

Seul : Les enfants jouaient.

G, S, Elle : Les gamins arrivent. Elle part. Ils se manquent de peu.

G, S, Il : Nos aventuriers se présentent. Il écoute. Ils s'allient.

G, S, [Elle + Il] : Anel et Laura sont parents de deux enfants . Les parents ont envie de vacances. Toute la famille part en vacances.

Les « Membre de lesEnfants »

Il existe cependant une annotation un peu spéciale dans le corpus. En effet, il est possible, à la volonté de l'annotateur, d'annoter les « Membre de lesEnfants ». Cette annotation permet de signaler un membre de « les enfants » qui n'est ni associé au référent « elle », ni au référent « elle ». Les maillons ainsi créés se voient rattachés à la CR_les enfants mais l'incertitude sur

le rattachement est signalée par un « Oui » dans les caractéristiques du maillon, accompagnée d'un commentaire explicite : « Membre de lesEnfants ». Cette annotation étant laissée au choix de l'annotateur, et n'étant donc pas « stable », nous avons pris la décision, dans le cadre de ce mémoire, de laisser ces maillons de côté et de ne pas les considérer comme des maillons de la CR_les enfants. Dans la suite du mémoire, les CR_les enfants présentées ne possèdent pas de maillon « Membre de lesEnfants ».

Les maillons dans la version « gold » du corpus RésolCo

Niveau	Copies « gold »	nb maillon_elle	nb maillon_il	nb maillon_les	nb maillon_MdE	Total	Part MdE
CE2	53	461	386	230	67	**1144**	5,86%
CM1	50	356	386	225	45	**1012**	4,45%
CM2	57	678	720	320	59	**1777**	3,32%
6e	106	1393	1413	682	129	3617	3,57%
5e + 4e	49	754	881	472	122	2229	5,47%
3e	55	921	947	374	149	2391	6,23%
M2	13	220	281	175	36	**712**	5,06%
Total	383	**4783**	**5014**	**2478**	607	**12882**	4,71%

Tableau 1 - Répartition des maillons par type sur le corpus « gold »

Comme le corpus RésolCo est toujours en cours de création, pour ce mémoire, nous nous sommes appuyés uniquement sur les copies « gold » disponibles. Ces copies sont les copies dont l'annotation est **certaine**. Il n'y a donc qu'une seule annotation par copie.
Le Tableau 1 présente la répartition des types de maillons en fonction du niveau scolaire sur le corpus. Nous observons la présence des maillons « Membre de lesEnfants » séparés des maillons « les enfants ». Nous pouvons ainsi visualiser la part de maillons « éliminés » par cette différence qui est d'environ 5% du nombre total des maillons présents dans le corpus. Dans les cellules en vert, nous pouvons observer que les 6^{e}, les 5^{e} + 4^{e} et les 3^{e} sont les niveaux qui possèdent le plus de maillons mais ce résultat suit le nombre de copies par niveau qui est également le plus élevé pour ces trois niveaux. Nous pouvons également observer que les 5^{e} sont rassemblés avec les 4^{e}. En effet, le corpus ne comptant que 4 copies en 5^{e} et comme c'est un niveau équivalent à celui de la 4^{e}, qui lui comporte 45 copies, nous avons pris la décision de les rassembler pour que les résultats soient plus simples à étudier. Il en sera ainsi dans tous les tableaux ou graphiques suivants.

La caractéristique « groupe »

Pour faire suite aux relations possibles entre les maillons, il existe une caractéristique majeure associée aux maillons : la caractéristique « groupe ». Comme nous l'avons vu dans la partie II-2, l'annotateur peut modifier la valeur par défaut « Non » de la caractéristique « groupe » par « Oui » quand la marque de continuité référentielle se trouve faire partie d'un groupe. Pour un maillon associé au référent « elle » par exemple, dans la phrase « Les parents travaillent » si un des parents est associé au référent « elle », un maillon « elle » sera créé par-dessus « Les parents » et ce maillon prendra « Oui » comme valeur de « groupe ». Ce phénomène peut se cumuler à la superposition. Par exemple, si l'autre parent est associé au référent « il », une superposition sera réalisée sur « Les parents » et le maillon « elle » ainsi que le maillon « il » prendront la valeur « Oui » dans « groupe ».

Niveau	Maillons elle et g:oui	Maillons elle et g:non	Maillons il et g:oui	Maillons il et g:non	Maillons les enfants et g:oui	Maillons les enfants et g:non
CE2	116	345	126	260	5	225
CM1	105	251	108	278	5	220
CM2	128	550	232	488	23	297
6e	315	1078	551	862	35	647
5e + 4e	132	622	365	516	72	400
3e	177	744	385	562	38	336
M2	50	170	87	194	9	166
Total	1023	**3760**	1854	**3160**	187	**2291**
Répartition	21,39%	**78,61%**	36,98%	**63,02%**	7,55%	**92,45%**

Tableau 2 - Répartition des types de maillons par la caractéristique groupe sur le corpus « gold »

Le Tableau 2 présente la répartition, pour les 3 types de maillons, des valeurs de « groupe » en fonction du niveau scolaire. Nous pouvons observer que le nombre de maillons avec la valeur « Oui » dans « groupe » est nettement inférieur à celui des maillons avec la valeur « Non ». Les maillons avec la valeur « Oui » ne représentent, sur le total des maillons du type, que 21,39% pour les maillons « elle », 36,98% pour les maillons « il » et seulement 7,55% pour les maillons « les enfants ».

2 – Filtrage du corpus

Nous avons défini les modèles théoriques et les relations possibles entre les maillons qui composent les CR, nous pouvons donc passer à l'application de ces modèles sur les données à l'étude : le corpus RésolCo. Dans cette partie, nous ne reviendrons pas sur les définitions données dans la partie III mais nous tenterons cependant de les illustrer, quand cela est possible, avec des copies issues du corpus ainsi qu'avec les modélisations linéaires des cohabitations appliquées aux marques de continuités référentielles associées aux trois référents donnés par la consigne de RésolCo.

Premièrement, pour que ces modélisations puissent être appliquées sur le corpus, le fichier d'annotation doit impérativement contenir au moins deux maillons associés à chaque type de référent (« elle », « il » et « les enfants »), puisque sinon, les CR ne peuvent pas exister. Cette condition constitue le premier filtre à travers lequel nous devons faire passer les copies.

Niveau scolaire	Nombre de copies « gold »	Nombre de fichiers possédant au moins 2 maillons par type de référent	Perte filtre 1
CE2	53	35	-33,96%
CM1	50	28	-44,00%
CM2	57	36	-36,84%
6e	106	69	-34,91%
5e + 4e	49	38	-22,45%
3e	55	31	-43,64%
M2	13	12	-7,69%
Total	**383**	249	**-34,99%**

Tableau 3 - Application du filtre 1

Dans le Tableau 3, nous pouvons observer que l'application du premier filtre est assez agressive car nous perdons déjà 34,99% du total des copies, soit 141 copies. Les trois cellules sur fond rouge représentent les trois niveaux qui sont le plus pénalisés par ce premier filtre. Dans l'ordre décroissant, ce sont les CM1, les 3e et les CM2 qui sont le plus pénalisés avec près de la moitié des copies écartées pour les CM1. Même le niveau le plus élevé, les Master 2, perdent une copie. Ce filtre touche donc tous les niveaux et nous nous sommes alors demandé quels types de maillons sont le plus souvent inférieurs à 2, ce qui cause l'élimination de la copie.

Niveau	Copies « gold »	2 maillons elle minimum	Perte	2 maillons il minimum	Perte	2 maillons les enfants minimum	Perte
CE2	53	48	-9,43%	45	-15,09%	42	-20,75%
CM1	50	43	-14,00%	40	-20,00%	35	-30,00%
CM2	57	54	-5,26%	51	-10,53%	39	-31,58%
6e	106	103	-2,83%	98	-7,55%	73	-31,13%
5e + 4e	49	49	0,00%	43	-12,24%	42	-14,29%
3e	55	53	-3,64%	49	-10,91%	33	-40,00%
M2	13	13	0,00%	13	0,00%	12	-7,69%
Total	383	363	-5,22%	339	-11,49%	**276**	**-27,94%**

Tableau 4 - Détail des pertes par type de maillon dues au premier filtre

La réponse à cette question se trouve dans le Tableau 4 où nous pouvons observer le type de maillon possédant le moins fréquemment un minimum de 2 maillons : « les enfants ». En effet, 27,94% des copies « gold » contiennent moins de 2 maillons associés au référent « les enfants ». C'est plus du double de la deuxième perte la plus importante et plus de 5 fois la perte la moins importante. Ces résultats peuvent surprendre mais ils semblent cohérents avec ce que nous avons pu observer lors de notre annotation. En effet, « les enfants » sont fréquemment représentés par un dernier maillon unique, celui lié à la dernière phrase consigne, souvent placée comme dernière phrase de la copie.
Nous remarquons également qu'il semble exister une corrélation entre l'ordre d'apparition de la phrase consigne et le nombre minimum de maillons associés au référent de ladite phrase. En première position dans la consigne se trouve la phrase du référent « elle » et c'est le type de maillon qui souffre le moins de la demande de la présence minimale de 2 maillons. Il en va de même pour la deuxième phrase consigne donnée, la phrase associée au référent « il ». Ce type de maillon se trouve en deuxième position en termes de pertes. La dernière phrase de la consigne, celle du référent « les enfants » se trouve en troisième et dernière position et ce type de maillon souffre le plus de la demande de la présence minimum de 2 maillons.

À propos de la consigne, nous voudrions rappeler que ces textes ne sont pas des textes tout venant. Ils sont tous produits dans le cadre d'une seule et même consigne, qui concerne l'utilisation des trois phrases consignes évoquées dans la partie II-1. Ainsi nous avons pris la décision de ne garder que les copies qui présentent strictement les trois phrases consignes. Cette condition constitue le deuxième filtre à travers lequel nous devons faire passer les copies.

Niveau scolaire	Nombre de fichiers possédant au moins 2 maillons par type de référent	Application du filtre 2	Perte entre filtre 1 et 2
CE2	35	35	0,00%
CM1	28	24	-14,29%
CM2	36	34	-5,56%
6e	69	67	-2,90%
5e + 4e	38	38	0,00%
3e	31	31	0,00%
M2	12	12	0,00%
Total	249	241	-3,21%

Tableau 5 - Application du filtre 2

Le Tableau 5 présente les résultats suite à l'application du deuxième filtre : la présence stricte des trois phrases consignes. En conséquence, nous perdons 3,21% des copies du total précédent (ou 37,08% depuis le début). Cette faible perte indique que la consigne originale de construire un récit autour des trois phrases imposées semble avoir été correctement respectée sur les copies restantes. Seuls trois niveaux sont touchés par ce filtre, dans l'ordre décroissant, les CM1, les CM2 et les 6e. Nous pouvons remarquer que ce sont plutôt des niveaux scolaires bas avec une décroissance entre le CM1 et la 6e.

Cependant, dans le corpus RésolCo, comme indiqué dans la partie II-3, lorsque l'élève n'a pas respecté la consigne uniquement en termes de genre ou de nombre sur les référents (exemple : Ils habitaient dans cette maison depuis longtemps), la copie est écartée du corpus. Par contre, si dans la copie les phrases consignes sont en désordre ou modifiées mais que le référent est correct en nombre et en genre (elle/il/les enfants), la copie est conservée et peut donc se retrouver dans cette version « gold » du corpus.
Afin de mettre les copies sur un pied d'égalité, nous avons pris la décision, pour l'application des modèles, de ne le faire que sur les copies qui présentent les trois phrases consignes **dans l'ordre** et donc d'appliquer un dernier filtre. Malheureusement, notre technique de vérification de l'ordre des phrases consignes (voir V-2) ne me permet pas de vérifier que le référent annoté dans la phrase consigne ne soit pas modifié, ni en genre, ni en nombre.

Niveau scolaire	Filtre 1 + 2	Application du filtre 3	Perte entre filtre 2 et 3
CE2	35	32	-8,57%
CM1	24	23	-4,17%
CM2	34	32	-5,88%
6e	67	62	-7,46%
5e + 4e	38	36	-5,26%
3e	31	31	0,00%
M2	12	11	-8,33%
Total	241	**227**	**-5,81%**

Tableau 6 - Application du filtre 3

Le Tableau 6 présente les résultats liés à l'application de ce dernier filtre. De façon surprenante, tous les niveaux sauf les 3e sont impactés. Nous nous séparons de 5,81% des copies en

appliquant ce filtre (ou 40,73% des copies depuis le début) amenant le total de copies restantes à 227 sur les 383 du départ. Sur les niveaux scolaires touchés, cela représente en moyenne 6,61% des copies qui ne présentent pas les trois phrases consignes dans l'ordre. Les niveaux les plus impactés sont, dans l'ordre décroissant, les CE2, les M2 et les 6e. Si la forte perte des copies de CE2 peut se justifier par le fait que les très jeunes apprenants ne respectent pas toujours correctement les consignes données, la présence de Master 2 est plus étonnante. Nous avons donc observé la copie écartée.

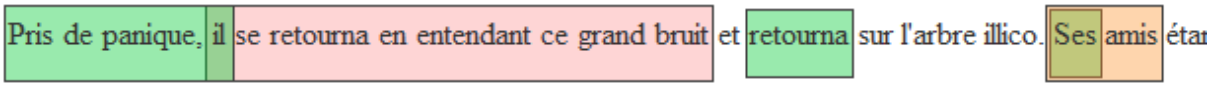


Illustration 24 - Extrait de copie annotée : UN-M2-2018-TUTJ2-D1-R2-V1_N_coref

Dans cette copie, les trois phrases consignes sont pourtant dans le bon ordre mais c'est cet extrait de la copie qui permet de comprendre ce qu'il se passe. Pour simplifier, la technique de vérification de l'ordre des phrases consignes est de vérifier si le maillon associé à « il » dans la deuxième phrase consigne s'y trouve. Il doit se trouver dans les limites de la phrase consigne, ici mise en valeur par la couleur rose. Le problème réside dans le fait que le maillon « il » est bien présent comme attendu mais qu'il dépasse des limites de la phrase consigne et donc la vérification n'est pas validée. Notre méthode ne prend pas en compte ce genre de cas spécifiques plus complexes à vérifier ce qui fait que, malheureusement, cette copie ne sera pas prise en compte ni dans l'application des modèles, ni dans les analyses.

Après l'application successive des 3 filtres présentés, il ne reste donc plus que 227 copies.

Niveau scolaire	Nombre de copies dans le corpus « gold »	Nombre de copies retenues après application des 3 filtres	Perte totale
CE2	53	32	-39,62%
CM1	50	23	-54,00%
CM2	57	32	-43,86%
6e	106	62	-41,51%
5e + 4e	49	36	-26,53%
3e	55	31	-43,64%
M2	13	11	-8,33%
Total	383	**227**	**-40,58%**

Tableau 7 - Comparaison entre corpus « gold » et corpus filtré

Le Tableau 7 présente les pertes totales par niveau, ainsi que la perte globale qui s'élève à près de 41%, soit 157 copies en moins. Nous ne nous attendions pas à perdre autant de copies annotées en appliquant nos trois critères de sélection mais le jeu de données est ainsi plus homogène et proche du résultat voulu lors de la présentation de la consigne aux scripteurs.

Niveau scolaire	Nombre de copies retenues	Pourcentage du total
CE2	32	14,10%
CM1	23	10,13%
CM2	32	14,10%
6e	62	27,31%
5e + 4e	36	15,86%
3e	31	13,66%
M2	11	4,85%
Total	**227**	**100,00%**

Tableau 8 - Répartition des copies par niveau scolaire sur le corpus résultant

Le Tableau 8 présente la répartition des copies restantes en fonction des niveaux scolaires. La répartition s'avère assez équilibrée sauf pour deux niveaux : 6e et M2. En effet, le nombre de copies en 6e est double par rapport aux autres niveaux hors M2. En M2, seules 11 copies annotées sont conservées mais cette faible présence est due au faible nombre de copies annotées de ce niveau dans la version « gold » du corpus RésolCo.

Les maillons dans le corpus filtré

Maintenant que nous avons filtré le corpus, nous pouvons étudier la répartition des maillons en fonction de leurs types sur les 227 copies.

Type de maillon	Nombre total	Pourcentage du total	Moyenne	Perte
Elle	3154	34,39%	13,89	-34,06%
Il	3513	38,31%	15,48	-29,94%
Les enfants	2008	21,90%	8,85	-18,97%
MdE	495	5,40%	2,18	-18,45%
Total	**9170**	100,00%	**40,40**	**-28,82%**

Tableau 9 - Répartition des types de maillons sur le corpus filtré

Nous avons inclus les maillons « Membre de lesEnfants » (MdE) dans le Tableau 9 dans le but de faire une comparaison avec le corpus « gold ». Nous perdons 18,45% de ce type de maillon dans le corpus filtré. Une tendance suivie par les maillons « les enfants » dont la perte s'élève à 18,97%. Les deux maillons associés aux référents singuliers, « elle » et « il », ont quant à eux le plus souffert de ce filtrage avec 39,94% de perte pour les maillons « il », suivi de 34,06% de perte pour les maillons « elle ». La perte totale sur les maillons est d'un peu plus d'un quart sur le total.

Concernant leur répartition, les maillons associés au référent « il » sont les plus présents avec 38,31% du total, suivis de près par les maillons associés au référent « elle » avec 34,39%. Leurs moyennes se suivent également avec respectivement, 15,48 et 13,89 maillons en moyenne par copie. Ces moyennes nous donnent la longueur moyenne des CR dans les copies. Les CR les plus longues sont donc, en moyenne, les CR_il, un résultat qui semble cohérent avec nos observations réalisées sur le corpus.

Pour les maillons associés au référent « les enfants », les valeurs sont plus faibles. Ils représentent seulement 21,90% du total des maillons et leur moyenne de 8,85 maillons par copie est nettement inférieure aux maillons associés aux référents singuliers. Les CR_les enfants sont donc, en moyenne, les CR les plus courtes dans le corpus.

La moyenne du total des maillons est d'environ 40, ce qui signifie qu'il y a en moyenne 40 maillons par copie annotée.

Nous nous sommes ensuite intéressés à la répartition de ces maillons en fonction du niveau scolaire.

Niveau	Nombre de copies	Nombre de maillons elle	Nombre de maillons il	Nombre de maillons les enfants	Nombre de maillons MdE	**Total**
CE2	32	320	307	161	55	843
CM1	23	228	258	142	23	651
CM2	32	426	468	260	44	1198
6e	**62**	**911**	**899**	**572**	91	**2473**
5e + 4e	36	585	716	386	99	1786
3e	31	497	620	322	149	1588
M2	11	187	245	165	34	631
Total	227	3154	3513	2008	495	**9170**

Tableau 10 - Répartition des types de maillons sur le corpus filtré en fonction du niveau scolaire

Dans le Tableau 10 nous pouvons observer, qu'à l'évidence, le niveau scolaire contenant le plus de copies, soit la 6e, possède les plus grands nombres de maillons pour chacun des types (sauf pour les maillons MdE). En deuxième position, nous retrouvons le groupe 5e + 4e avec les deuxièmes plus grands nombres de maillons pour chaque type (sauf pour les maillons MdE). Un résultat cohérent puisque ce groupement est le second niveau le plus prolifique en nombre de copies sur le corpus filtré. Ensuite, les résultats deviennent contre-intuitifs car à la 3e position du total du nombre de maillons se trouve le niveau de 3e. Ce niveau n'est pourtant que le 4e en termes dans la répartition des copies dans le corpus filtré.
Nos échantillons de copies n'étant pas de tailles équivalentes d'un niveau à l'autre, nous proposons d'étudier les moyennes du nombre de maillons par type et par niveau scolaire.

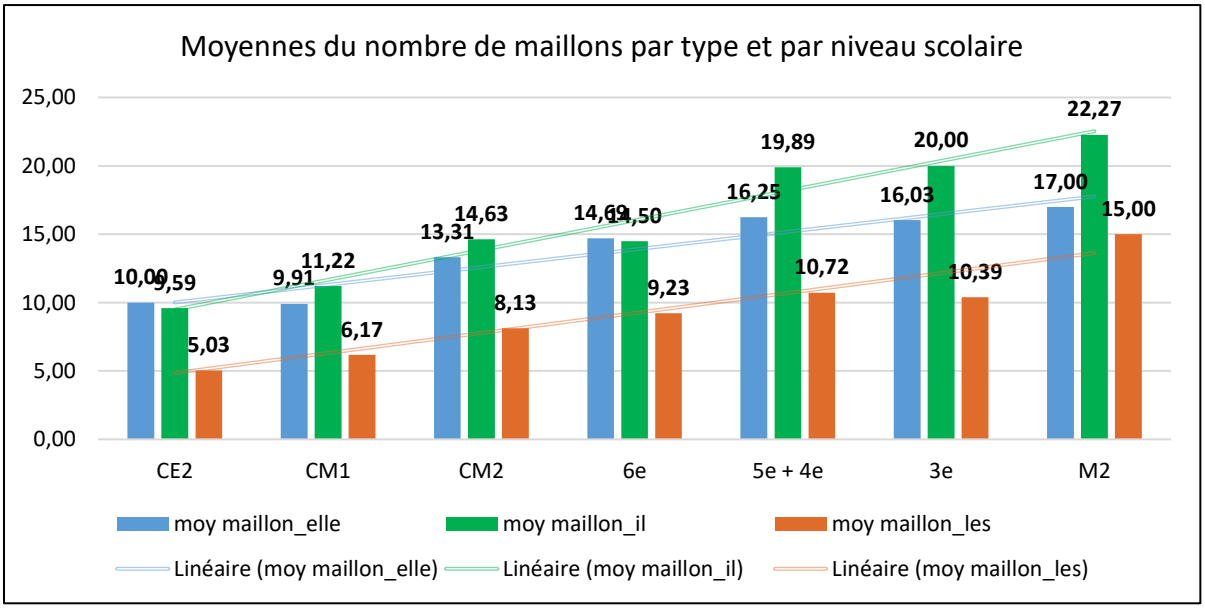


Illustration 25 - Histogramme des moyennes du nombre de maillons par type et par niveau scolaire

Sur l'Illustration 25, nous pouvons observer l'augmentation des moyennes du nombre de maillons par type à mesure que le niveau scolaire augmente. Cette augmentation se fait de façon quasi linéaire entre le CE2 et le M2. Elle est de 70% pour les maillons « il », de 132,16% pour les maillons « elle » et de 198,14% pour les maillons « les enfants ».

Niveau	Maillons elle et g:oui	Maillons elle et g:non	Maillons il et g:oui	Maillons il et g:non	Maillons les enfants et g:oui	Maillons les enfants et g:non
CE2	84	236	110	197	1	160
CM1	79	149	96	162	5	137
CM2	85	341	186	282	22	238
6e	260	651	442	457	25	547
5e + 4e	117	468	306	410	60	326
3e	101	396	270	350	35	287
M2	49	138	76	169	9	156
Total	775	**2379**	1486	**2027**	157	**1851**
Répartition	24,57%	**75,43%**	42,30%	**57,70%**	7,82%	**92,18%**

Tableau 11 - Répartition des types de maillons par la caractéristique groupe sur le corpus filtré

Le dernier tableau de cette section, le Tableau 11 nous permet de visualiser la répartition des valeurs « Oui » et « Non » de la caractéristique « groupe » en fonction du type de maillon et du niveau scolaire. La répartition suit à peu près la même tendance que celle que nous avions mesurée sur la version « gold » de RésolCo (Tableau 2). Les maillons associés au référent « elle » avec « groupe » valant « Oui » sont plus présents de 3,18%. Pour les maillons associés au référent « il », l'augmentation est de 5,32% et pour les maillons associés au référent « les enfants », l'augmentation est de 0,27%. Le filtrage a donc conservé des copies où les maillons sont un peu plus en « groupe » même si la majorité des maillons de chaque type restent encore « seuls ».

3 - Application des deux cohabitations au corpus

Dans la consigne donnée aux élèves, il leur est demandé de conserver l'ordre des phrases consignes. Nous avons répercuté cette consigne en filtrant le corpus « gold » et l'une des conséquences est de limiter les « stratégies discursives » possibles. En effet, si l'ordre des phrases consignes n'était pas contraint, il y aurait potentiellement plus de variété dans les stratégies employées. Ainsi, l'ordre d'apparition attendu des CR se fait d'abord avec la CR_elle, puis la CR_il, puis la CR_les enfants puisque c'est celui demandé par la consigne. C'est donc dans cet ordre que nous allons appliquer les modélisations théoriques de la partie III.

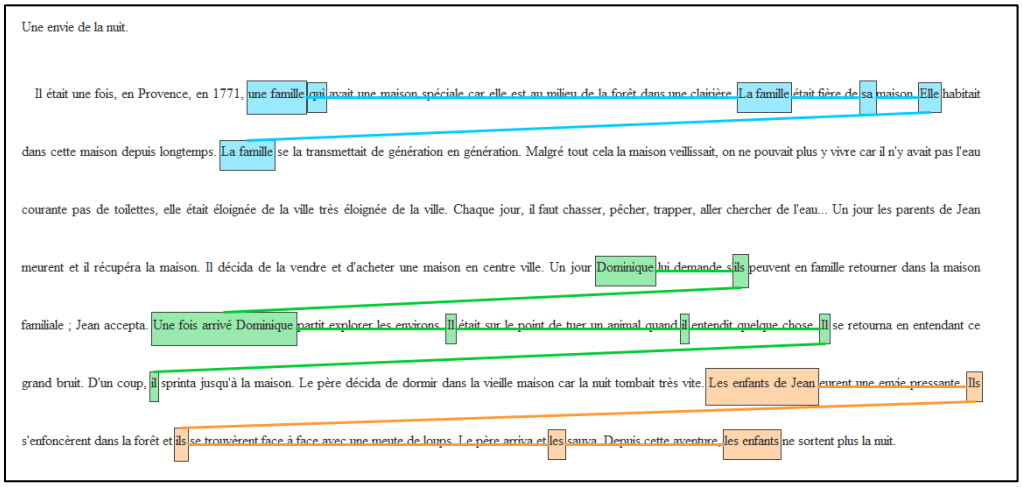
Toujours à propos de la consigne, nous voudrions mettre en avant un phénomène que nous avons remarqué lors de l'annotation. La consigne de RésolCo contient une phrase avec un référent pluriel : « les enfants ». Cette phrase peut alors générer des comportements particuliers, comme l'implication des référents singuliers « il » ou « elle », donnés par les deux autres phrases, dans le référent « les enfants » afin de former un groupe. Cette formation en groupe est peut-être avantagée par la proximité des phrases consignes dans leur présentation ainsi que par la dualité des référents singuliers/pluriels mais nous y reviendrons par la suite.

A – La succession

Comme nous l'évoquions dans la section III-1, la succession est pour nous le niveau le plus simple de la gestion des référents avec une stratégie de traitement « pas à pas ». Cette stratégie pourrait être favorisée par la consigne qui demande de placer dans un ordre précis des phrases contenant chacune un référent distinct. Ainsi chaque phrase peut être traitée en construisant un morceau de l'histoire autour, puis le scripteur sélectionne la phrase suivante et recommence. Comme les référents sont traités un par un, les marques de continuités référentielles disparaissent rapidement, voire immédiatement, quand une nouvelle phrase est sélectionnée.

Succession stricte

Étant la cohabitation la plus simple, elle devrait revenir fréquemment dans le corpus. Elle s'illustre par exemple dans la copie suivante, issue de RésolCo :


Une envie de la nuit.

Il était une fois, en Provence, en 1771, une famille qui avait une maison spéciale car elle est au milieu de la forêt dans une clairière. La famille était fière de sa maison. Elle habitait dans cette maison depuis longtemps. La famille se la transmettait de génération en génération. Malgré tout cela la maison veillissait, on ne pouvait plus y vivre car il n'y avait pas l'eau courante pas de toilettes, elle était éloignée de la ville très éloignée de la ville. Chaque jour, il faut chasser, pêcher, trapper, aller chercher de l'eau... Un jour les parents de Jean meurent et il récupéra la maison. Il décida de la vendre et d'acheter une maison en centre ville. Un jour Dominique lui demande s'ils peuvent en famille retourner dans la maison familiale ; Jean accepta. Une fois arrivé Dominique partit explorer les environs. Il était sur le point de tuer un animal quand il entendit quelque chose. Il se retourna en entendant ce grand bruit. D'un coup, il sprinta jusqu'à la maison. Le père décida de dormir dans la vieille maison car la nuit tombait très vite. Les enfants de Jean eurent une envie pressante. Ils s'enfoncèrent dans la forêt et ils se trouvèrent face à face avec une meute de loups. Le père arriva et les sauva. Depuis cette aventure, les enfants ne sortent plus la nuit.

Illustration 26 - Copie annotée : CO-4e-2018-LSPJJRD-D1-R2-V1_N_coref

Modélisation appliquée :

Illustration 27 - Modélisation appliquée de la succession stricte

Ordre attendu : Elle – Il – Les enfants

Succession chevauchée minimale groupée

Ce type de cohabitation n'est pas illustré car il n'a pas été trouvé dans le corpus RésolCo.

Remarque : Le regroupement elle-il et il-les peut correspondre à trois cas de figure dans le corpus RésolCo : inclusion A dans B, inclusion B dans A et superposition. Dans l'étude des formes de la succession (succession chevauchée minimale groupée et succession chevauchée), les trois cas sont possibles quand il s'agit de regroupement. Par exemple, le regroupement entre « elle » et « il » correspond soit à un maillon « elle » contenu dans un maillon « il », soit à un maillon « il » contenu dans un maillon « elle », soit à la superposition du maillon « il » et du maillon « elle ».

Modélisation appliquée :

Illustration 28 - Modélisation appliquée de la succession chevauchée minimale groupée

Ordre attendu : Elle – Il – Les enfants

Succession chevauchée minimale distincte

Ce type de cohabitation n'est pas illustré car il n'a pas été trouvé dans le corpus RésolCo.

Modélisation appliquée :

Illustration 29 - Modélisation appliquée de la succession chevauchée minimale distincte

Ordre attendu : Elle – Il – Les enfants

Succession chevauchée

Elle s'illustre par exemple dans les copies suivantes, issues de RésolCo :

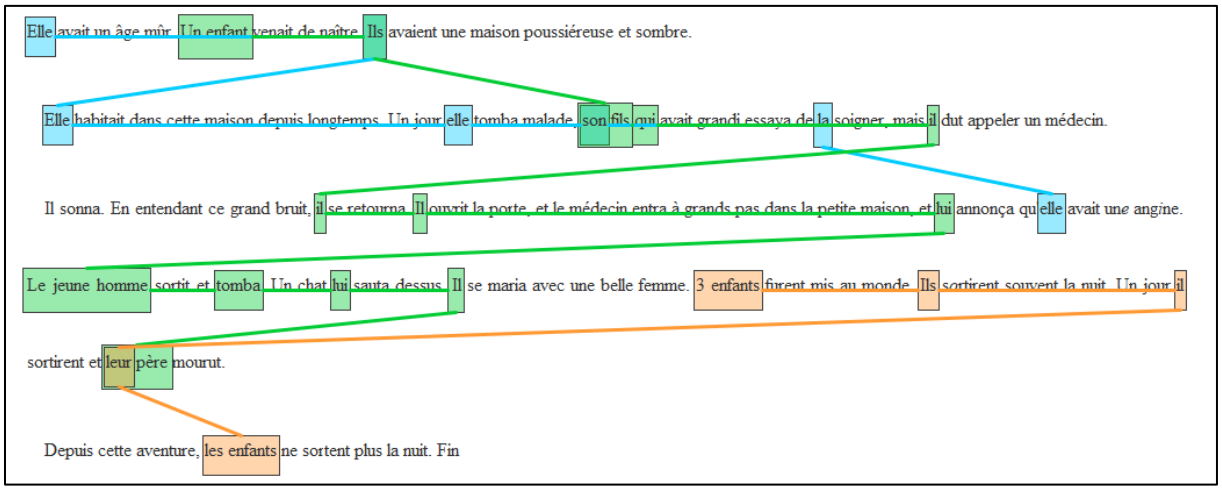


Illustration 30 - Copie annotée : EC-CE2-2017-TBZX-D1-R1-V1_N_coref

Illustration 31 - Copie annotée : CO-4e-2018-LSPJJRC-D1-R29-V1_N_coref

Modélisation appliquée :

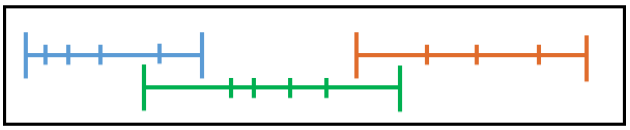


Illustration 32 - Modélisation appliquée de la succession chevauchée

Ordre attendu : Elle – Il – Les enfants

Remarque : théoriquement, d'autres formes de succession chevauchée sont possibles tout en respectant l'ordre des phrases consignes (« pc ») comme dans l'Illustration 33.

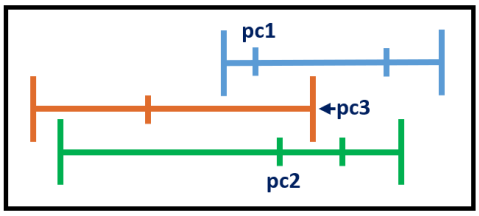


Illustration 33 - Modélisation appliquée d'un autre ordre de succession théorique

La CR_elle pourrait être la dernière CR à démarrer que cela ne poserait pas d'inconvénient. Cependant, afin de garder une continuité dans l'étude de la succession, nous ne cherchons que les successions chevauchées formées dans l'ordre attendu, celui illustré dans la modélisation appliquée (Illustration 32) de cette cohabitation.

Illustrations complémentaires pour la succession chevauchée

Ces illustrations ne sont pas recherchées dans le corpus. Elles viennent cependant compléter l'application de la succession chevauchée au corpus filtré.

Absence de succession chevauchée

Modélisation appliquée :

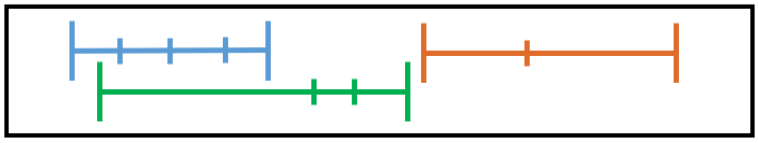


Illustration 34 - Modélisation appliquée de l'absence de succession chevauchée

Ordre attendu : Elle – Il – Les enfants

Explication :

Dans ce cas, la cohabitation présentée n'est pas celle d'une succession chevauchée car, la CR_il ne contient pas de maillon de la CR_les enfants. La condition pour qu'il y ait une cohabitation dite de « succession chevauchée » n'est donc pas remplie.

Succession chevauchée d'étendue maximale

Modélisation appliquée :

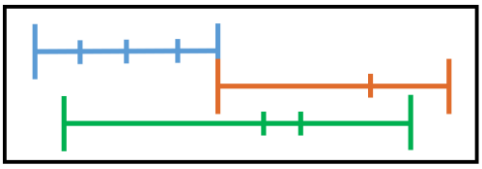


Illustration 35 - Modélisation appliquée de la succession chevauchée d'étendue maximale

Ordre attendu : Elle – Il – Les enfants

Explication :

Conformément à la définition de la succession chevauchée, la CR_il ne peut pas contenir une CR complète, elle ne peut donc pas commencer avant le premier maillon_elle et ne peut pas finir après le dernier maillon_les enfants. Ce schéma présente donc l'étendue maximale possible dans les copies du corpus RésolCo.

B – L'association/dissociation

L'étude de ce type de cohabitation déjà motivée par sa position opposée à la succession sur l'échelle de difficulté estimée à manipuler les référents, l'est également par la volonté de

connaitre un peu mieux le comportement que nous évoquions dans la partie IV-3 : **l'inclusion du référent singulier dans le référent pluriel « les enfants ».**

Si l'association/dissociation peut se faire avec des référents singuliers/singuliers ou pluriels/pluriels, nous ne traiterons ici que du phénomène de regroupement d'un référent singulier « il » ou « elle » dans le référent pluriel « les enfants ». Ce phénomène est indépendant de la consigne, car elle ne demande pas explicitement d'associer les référents ensemble. Il serait peut-être lié à une « envie » d'inclure un référent singulier dans le référent « final » à cause de l'ordre de présentation des phrases consignes. Ce phénomène est peut-être également avantagé par la proximité des phrases consignes dans leur présentation ainsi que par la dualité des référents singuliers/pluriels avec la dualité toute particulière « il » / « les enfants ». En effet, il semble, en français, plus « facile » ou « naturel », en comparaison avec un référent singulier féminin, d'inclure le référent singulier masculin dans un référent pluriel masculin.

En appliquant la définition à l'annotation réalisée sur le corpus, la détection de l'inclusion évoquée précédemment ne se fera, dans les prochains cas appliqués, qu'à travers la superposition. C'est en effet le type de regroupement qui rend le mieux compte de la réalité d'un groupe de personnes ; ce qui n'est pas toujours le cas de l'inclusion. Cependant, il existe un cas particulier où « il » et « elle » sont tous les deux « les enfants », de fait, il peut exister un maillon « les enfants » qui contienne un maillon « il » et un maillon « elle » ainsi que l'unité linguistique les reliant, comme une conjonction de coordination par exemple (Illustration 36).

Illustration 36 - Extrait de copie annotée : CO-6e-2016-PJPR5-D1-R14-V1_N_coref

Ce cas, plus rare, et plus compliqué à détecter n'est pas pris en compte dans le cadre de ce mémoire. Seule la superposition nous permettra de détecter l'association de deux référents.

Fusion

Elle s'illustre par exemple dans les copies suivantes, issues de RésolCo :

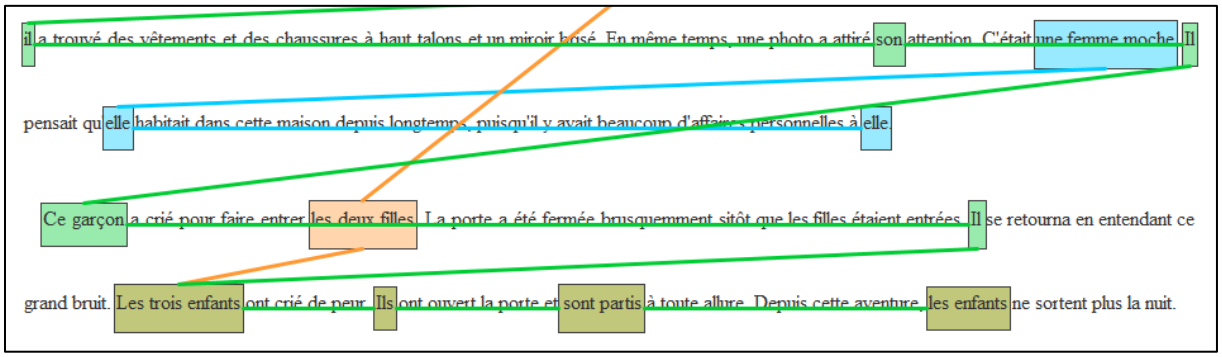


Illustration 37 - Extrait de copie annotée : UN-M2-2018-TUTJ2-D1-R12-V1_N_coref

Dans l'Illustration 37, « il » **fusionne** avec « les enfants ».

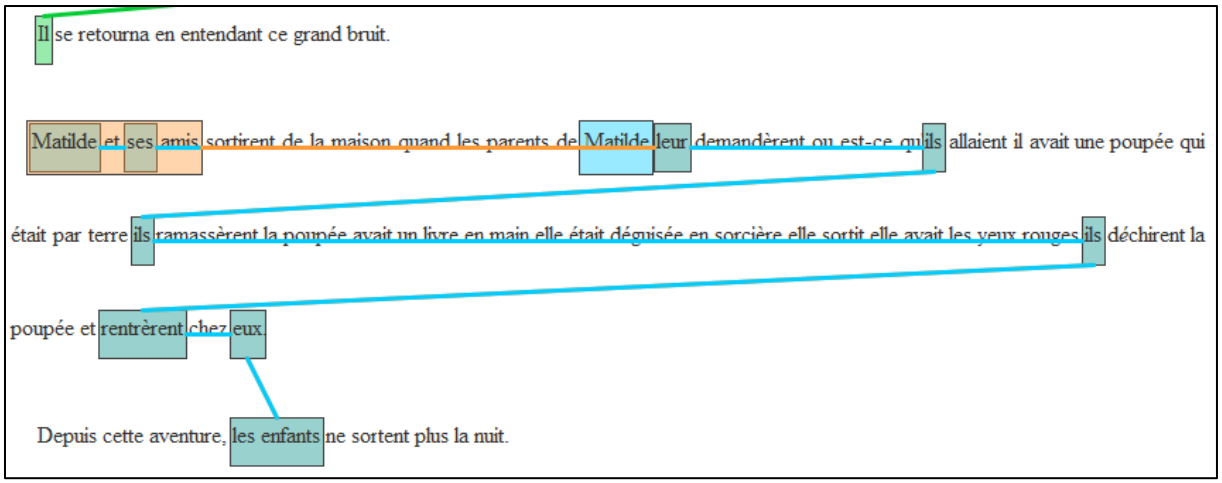


Illustration 38 - Extrait de copie annotée : EC-CM1-2015-TFGLX-D1-R6-V1_N_coref

Dans l'Illustration 38, « elle » **fusionne** avec « les enfants ».

Définition appliquée :

Un maillon associé à un référent singulier (« il » ou « elle ») **s'associe** avec un maillon lié au référent pluriel « les enfants » afin de former **un groupe**. Les maillons associés au référent singulier sont **superposés** avec ceux de « les enfants » jusqu'à la fin du texte.

Modélisation appliquée :

Vue d'ensemble

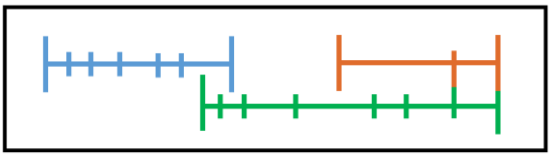


Illustration 39 - Modélisation appliquée de la fusion

Ordre attendu : Elle – Il – Les enfants

Zoom

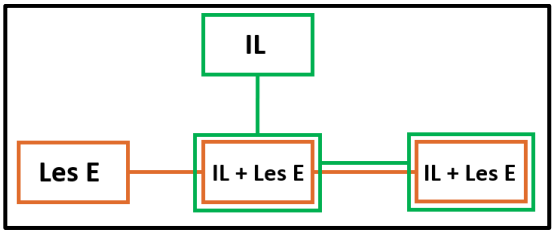


Illustration 40 - Zoom sur la fusion appliquée

Note complémentaire

Ce phénomène peut se produire jusqu'à deux fois. Une première fois avec un maillon associé à un référent singulier et une seconde fois avec un maillon associé au référent singulier restant, ce qui donne **une double fusion** et des maillons superposés dans un groupe [« les enfants » +

« il » + « elle »].
Cela s'illustre par exemple dans la copie suivante, issue de RésolCo :

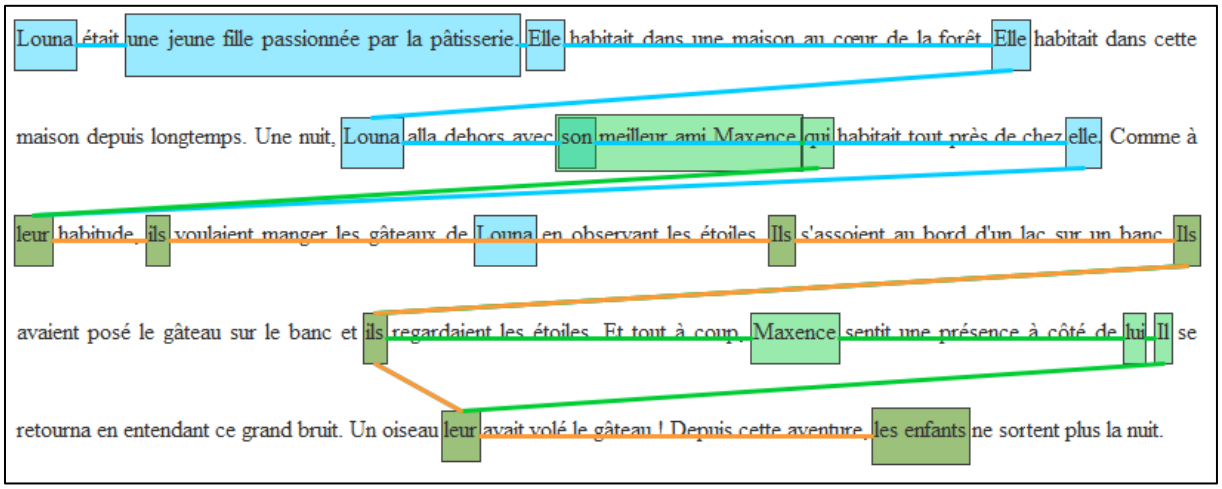


Illustration 41 - Copie annotée : CO-6e-2016-VTAC603-D1-R10-V1_N_coref

Dans l'Illustration 41, « elle » fusionne avec les enfants à partir du maillon « ils » dans la phrase « Ils s'assoient […] ». Les maillons de CR_il ne fusionnent qu'à partir du maillon « leur » dans la phrase « Un oiseau leur […] » créant ainsi une double fusion en deux temps.

Les 3 CR peuvent également fusionner en même temps comme c'est le cas dans l'Illustration 42.

Illustration 42 - Extrait de copie annotée : EC-CM2-2016-SGLEA-D1-R11-V1_N_coref

Dans cette illustration, « elle » et « il » fusionnent en même temps à partir du maillon « Ils » dans la phrase « Ils sortirent dehors […] » créant ainsi une **double fusion**.

Zoom

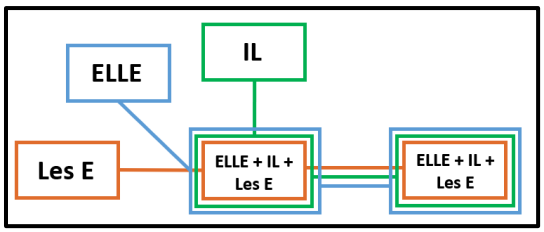


Illustration 43 - Zoom sur la fusion appliquée instantanée de tous les référents

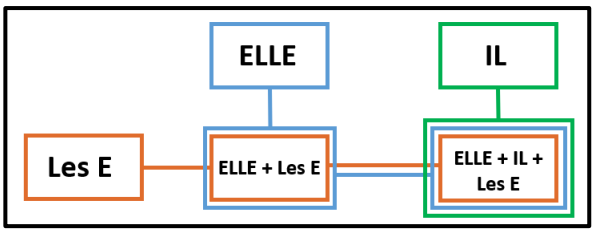


Illustration 44 - Zoom sur la fusion appliquée en deux temps de tous les référents

L'association/dissociation

Elle s'illustre par exemple dans les copies suivantes, issues de RésolCo :

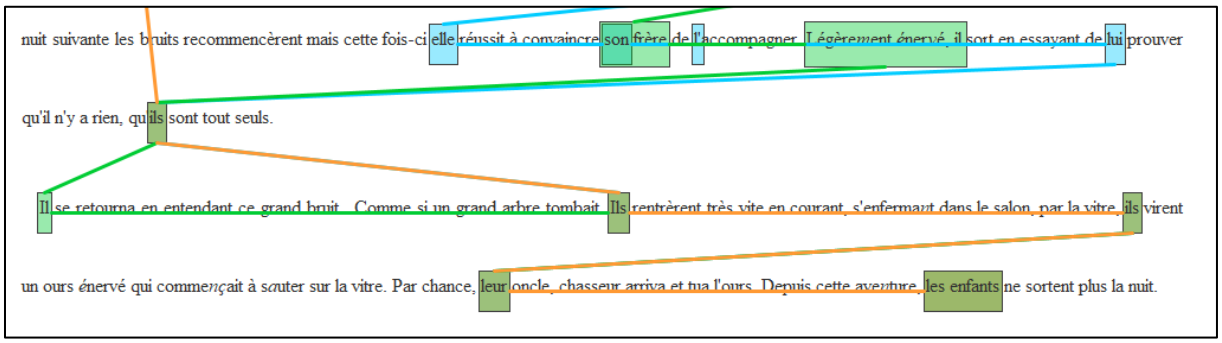


Illustration 45 - Extrait de copie annotée : CO-3e-2016-VTAC305-D1-R20-V1_N_norm_coref

Dans l'Illustration 45, « il » **s'associe** avec le regroupement [elle + les enfants] pendant un maillon, s'en **dissocie** pendant un maillon puis **fusionne** avec le regroupement [elle + les enfants].

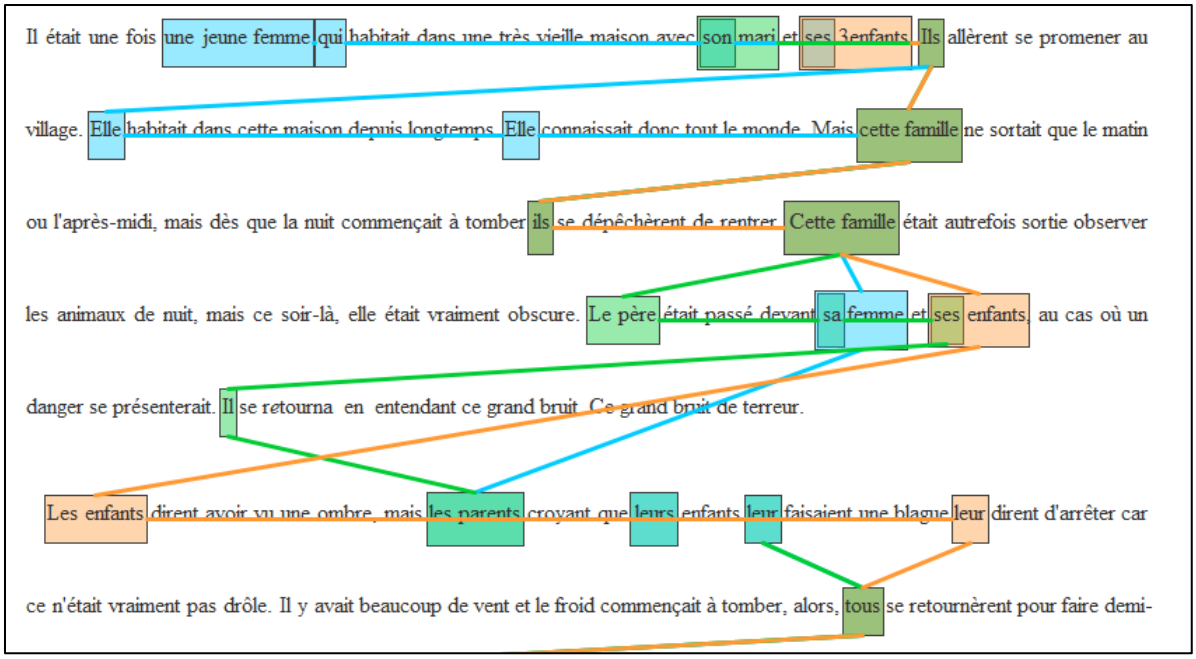


Illustration 46 - Extrait de copie annotée : CO-4e-2018-LSPJJRC-D1-R3-V1_N_coref

Dans l'Illustration 46, nous pouvons observer plusieurs A/D. Premièrement, « elle », « il » et « les enfants » s'associent à partir du maillon « Ils » dans la phrase « Ils allèrent se promener […] ». La CR_elle se dissocie pendant 2 maillons avant de s'associer à nouveau avec « il » et « les enfants » dans une **triple superposition**. Ensuite, après « Cette famille », les trois référents se dissocient.
Fait remarquable, « il » et « elle » s'associent au préalable dans le maillon « les parents » dans la phrase « […] les parents croyant que […] » avant de s'associer une dernière fois (dans cet extrait) avec « les enfants » dans le maillon « tous » dans la phrase « tous se retournèrent […] ».

Définition appliquée :

Un maillon associé à un référent singulier (« il » ou « elle ») **s'associe** avec un maillon lié au référent pluriel « les enfants » afin de former un **groupe**. Les maillons ainsi **superposés** sur une période minimale d'un maillon finissent par se **dissocier** et ne se trouvent plus regroupés. Ce phénomène peut se produire un nombre de fois illimité dans un texte.

Modélisation appliquée :

Vue d'ensemble

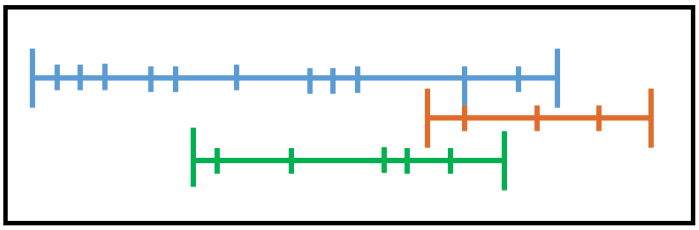


Illustration 47 - Modélisation appliquée de l'association/dissociation

Ordre attendu : Elle – Il – Les enfants

Zoom

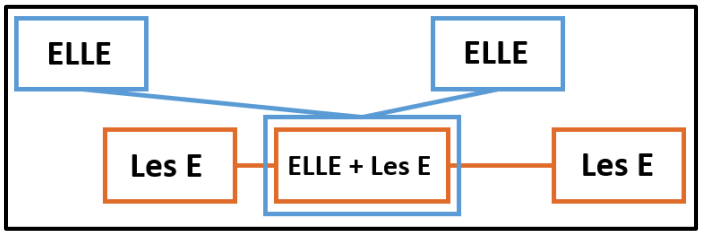


Illustration 48 - Zoom sur l'association/dissociation appliquée

L'association/dissociation puis fusion

C'est la forme de l'association/dissociation la plus complexe car la dernière A/D se transforme en fusion comme dans l'Illustration 45.
Dans l'application de notre modèle au corpus, nous distinguerons deux phénomènes :

- Le cas « simple », quand le regroupement étudié entre deux référents est le même pour l'A/D et la fusion. Par exemple : A/D il + les puis fusion il+les.
- Le cas « croisé », quand les regroupements étudiés entre deux référents ne sont pas les mêmes pour l'A/D et la fusion. Par exemple : A/D il+les puis fusion **elle+les**.

Dans les illustrations suivantes (Illustration 49 et Illustration 50), la cohabitation d'association/dissociation puis fusion est donc « simple » puisque seules deux CR s'associent (et se dissocient) : la CR_elle et la CR_les enfants.

Modélisation appliquée :

Vue d'ensemble

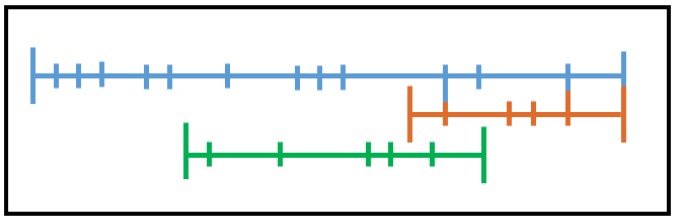


Illustration 49 - Modélisation appliquée de l'association/dissociation puis fusion « simple »

Zoom

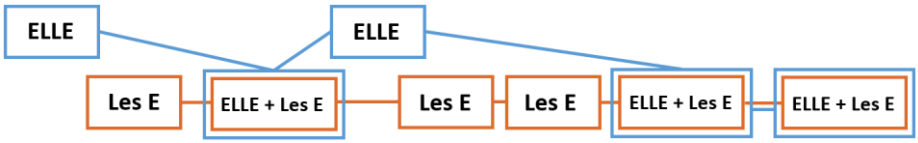


Illustration 50 - Zoom sur l'association/dissociation puis fusion « simple »

V – Manipulations et programmes

1 – Accéder aux données de l'annotation

Avant de commencer cette partie, nous voudrions apporter une précision technique. Dans le corpus RésolCo, les annotations se font sur la base d'une copie d'élève numérisée et normalisée. Chaque caractère, même le caractère d'espace, possède une position dans la copie. La copie commence avec le premier caractère qui possède donc une position initiale, le plus souvent « 0 ». À chaque caractère qui suit, la position augmente d'un jusqu'au dernier caractère. Chaque unité linguistique peut ainsi être identifiée avec une position et une longueur. Sa position démarre à un premier caractère x et se termine à un dernier caractère y. Les longueurs sont calculées en soustrayant x à y. L'annotation par maillon possède les mêmes caractéristiques de position et de longueur. Un maillon commence à une position et se finit à une autre, avec une longueur minimale d'un caractère (y-x >= 1).
Ainsi, en étudiant les positions de début et de fin des maillons, il est possible de les placer les uns par rapport aux autres. Ce sont ces placements relatifs qui vont être la base de nos programmes de recherche des cohabitations des continuités référentielles.

Comme nous l'avons vu précédemment dans la partie II-2, l'annotation réalisée via Glozz est en réalité une « surcouche » enregistrée dans un fichier .aa, qui est un fichier semblable à un fichier XML. Dans ces fichiers se trouvent toutes les informations en rapport avec les maillons créés lors de l'annotation, telles que sa position de début, sa position de fin, des caractéristiques comme le « groupe », les incertitudes, etc.

L'idée pour accéder à ces données est donc de construire un programme qui va lire les fichiers d'annotations ligne par ligne afin de repérer les unités qui sont les « blocs » qui contiennent toutes les informations relatives à un maillon unique.

```
<unit id="laubertin_1622032561686">
    <metadata>
        <author>laubertin</author>
        <creation-date>1622032561686</creation-date>
        <lastModifier>n/a</lastModifier>
        <lastModificationDate>0</lastModificationDate>
    </metadata>
    <characterisation>
        <type>maillon_Elle</type>
        <featureSet>
            <feature name="type">maillon_Elle</feature>
            <feature name="groupe">Non</feature>
            <feature name="incertitude sur la délimitation">Non</feature>
            <feature name="incertitude sur le rattachement">Non</feature>
            <feature name="commentaire"/>
        </featureSet>
    </characterisation>
    <positioning>
        <start>
            <singlePosition index="62"/>
        </start>
        <end>
            <singlePosition index="66"/>
        </end>
    </positioning>
</unit>
```

Illustration 51 - Visualisation d'un maillon_Elle dans un fichier .aa

Nous pouvons ici observer une unité (« unit ») qui correspond à un maillon associé au référent « elle ». Dans un premier temps, nous retrouvons les « metadata » qui concernent l'auteur et la date de création/modification. Nous n'exploitons pas ces « metadata » dans le cadre de ce mémoire.

Ensuite vient la deuxième partie « characterisation » qui contient des informations sur les caractéristiques du maillon. La balise « type » nous indique à quel référent il est associé : dans l'Illustration 51, il s'agit du référent « elle ». Dans cette partie, c'est l'ensemble « featureSet » qui est le plus intéressant. On y trouve à nouveau le type du maillon, s'il appartient à un groupe (Oui/Non), s'il y a des incertitudes sur la délimitation du maillon (Oui/Non), s'il y a des incertitudes sur le rattachement de ce maillon au référent du « type » (Oui/Non) ainsi que « commentaire » qui reste vide par défaut mais qui peut contenir n'importe quel commentaire laissé par l'annotateur. Comme nous l'avons vu dans la partie IV-1, cette caractéristique est d'autant plus importante que c'est à cet endroit que l'annotateur va laisser le commentaire « Membre de « lesEnfants » » au maillon « les enfants », permettant ainsi de le distinguer.

La dernière partie « positioning » de cette unité contient la position du maillon. La position est déterminée par rapport au texte. Dans cette partie, nous trouvons donc la position de départ du maillon (à quel caractère il commence) ainsi que sa position de fin (à quel caractère il s'arrête). Cette partie, tout comme le bloc unité, ne contient aucune information concernant le contenu linguistique présent entre le départ et la fin du maillon. Le fichier d'annotation est donc une suite d'unités ne décrivant que les caractéristiques données à un maillon par l'annotateur ainsi que sa position par rapport au texte.

Afin d'extraire ces informations, nous avons réalisé un programme en Python qui se charge de chercher ces unités. Il accumule les caractéristiques des unités, dont la position du maillon, dans des dictionnaires[10] avec en clé l'identifiant unique du maillon[11] et la valeur que nous souhaitons conserver comme la position de départ par exemple.
À cette recherche d'informations dans les unités s'ajoute un comptage qui permet de répartir les résultats des recherches des cohabitations en fonction des niveaux scolaires présents dans le corpus RésolCo, de la même manière que dans les sections IV-1 et IV-2.

Nous avons donc créé 6 programmes[12] qui, en combinant l'extraction d'information et les recherches conditionnelles que nous allons présenter, permettent de chercher les différentes cohabitations à l'étude dans ce mémoire : succession stricte, succession chevauchée minimale groupée, succession chevauchée minimale distincte, succession chevauchée, fusion, association/dissociation et association/dissociation puis fusion. Nous allons expliquer leurs fonctionnements et présenter les résultats obtenus dans la partie qui suit.

En complément de ces programmes, il en existe un dernier qui est uniquement dédié au comptage du nombre de maillons en fonction de leurs types et de leur répartition par niveau scolaire qui a servi à la réalisation des sections IV-1 et IV-2.

2 – Recherches conditionnelles utilisées dans les programmes

Les programmes fonctionnent tous avec le même principe. Avant tout, pour que la recherche soit appliquée, il faut que le fichier valide les 3 étapes de filtrage : contenir au minimum deux maillons de chacun des trois types (« elle », « il » et « les enfants ») et contenir les 3 phrases consignes, dans l'ordre. Si ce n'est pas le cas, le fichier est écarté.

Premièrement, pour vérifier que les copies contiennent au moins deux maillons de chaque type, nous vérifions le nombre d'éléments dans nos dictionnaires de départ de chaque type de maillon. S'il y a au moins deux éléments dans chacun des 3 dictionnaires de départ des maillons, alors la copie est conservée[13].
Deuxièmement, pour vérifier que les phrases consignes sont présentes et dans l'ordre, nous faisons deux opérations successives. Comme pour le nombre minimum de maillons, nous vérifions que le dictionnaire de départ des phrases consignes contient bien 3 et uniquement 3 éléments. Si c'est le cas, les 3 phrases consignes sont bien présentes. Nous pouvons maintenant vérifier si elles sont bien dans l'ordre demandé par la consigne.
Troisièmement, si les informations de départ et de fin des phrases consignes sont récupérées dans le fichier d'annotation depuis des « blocs » similaires à ceux des maillons associés aux

[10] Le dictionnaire en Python est une collection non ordonnée, modifiable et indexée de paires clé-valeur.
[11] Donné par l'attribut « id » dans la balise « unit ». Cet attribut a pour valeur « laubertin_1622032561686 » dans l'Illustration 51. C'est cette valeur, unique pour chaque maillon, qui sert de clé dans les différents dictionnaires.
[12] Il est théoriquement possible de combiner ces différents programmes afin de n'en former mais qu'un mais il était plus facile pour nous de travailler sur de plus petits programmes séparés.

[13] Nous ne vérifions pas dans les dictionnaires de fin car si un maillon a un départ, il a forcément une fin, grâce aux conditions d'annotations de Glozz. Mais le test de vérification pourrait être renforcé en regardant également la longueur des dictionnaires de fin.

référents étudiés, ces blocs ne contiennent pas d'informations sur le type de référent associé à la phrase, ni à son contenu linguistique. Alors, pour vérifier l'ordre attendu, nous utilisons une ruse. Nous formons une liste de tuples[14] avec les positions de départ et de fin des 3 phrases consignes. Nous ordonnons cette liste dans l'ordre croissant. Ensuite, nous vérifions si dans le premier tuple (et donc la première phrase consigne) se trouve bien un maillon « elle », dans le deuxième, un maillon « il » et dans le dernier, un maillon « les enfants ». Mais pour effectuer ce test, il faut au préalable avoir reconstruit les maillons de chacune des CR. Alors avant de vérifier si les phrases consignes contiennent bien des maillons, nous constituons en amont des listes de tuples contenant la position de départ, de fin ainsi que la caractéristique « g:oui » si la caractéristique « groupe » est annotée « oui » et « g:non » si elle est annotée « non ». Nous formons 3 listes pour les 3 types de maillons. Nous ordonnons ces listes dans l'ordre croissant et nous obtenons ainsi les 3 CR reconstituées[15]. Il nous est maintenant possible de vérifier si les phrases consignes contiennent bien le maillon qu'elles sont supposées contenir à leurs positions. Si les phrases consignes contiennent bien leurs maillons respectifs, la copie a passé les 3 filtres et nous pouvons réaliser nos recherches de cohabitations.

Précisions techniques concernant la formation des tuples : pour chaque clé dans le dictionnaire « départ », nous récupérons sa valeur, ainsi que sa valeur associée dans le dictionnaire « fin », et finalement sa valeur associée dans le dictionnaire « groupe ». Les trois valeurs misent bout à bout forment un tuple qui est ensuite ajouté à une liste représentant la CR d'un des trois référents étudiés. Ces listes sont ordonnées de manière croissante, avec comme critère la position de départ du maillon. Cette organisation est possible car deux maillons associés à un même référent ne peuvent pas empiéter l'un sur l'autre. C'est-à-dire que le maillon suivant dans une CR commence forcément, et au minimum, à la position de fin du maillon d'avant + 1.

Sur le même principe que la récolte des informations contenues dans les « unit » (section V-1), les informations concernant les longueurs des copies sont également récupérées depuis des blocs qui se différencient par la balise « type » contenant alors le mot « paragraphe ». Chaque copie pouvant être composée d'un ou plusieurs paragraphes, les valeurs de départ et de fin de chaque paragraphe sont extraites dans deux dictionnaires « début » et « fin ». Afin d'obtenir la longueur totale de la copie, nous prenons la valeur minimale contenue dans le dictionnaire « début » ainsi que la valeur maximale contenue dans le dictionnaire « fin » et nous calculons la différence entre les deux. Cette différence nous donne le nombre de caractères que contient la copie, et donc sa longueur. Cette longueur est ensuite ajoutée aux autres informations et finalise le tuple unique de chaque fichier, composé de son nom, son nombre de maillons « elle », « il », « les enfants » et la longueur de la copie.

Nous venons de voir les principes qui lient les différents programmes utilisés, nous allons maintenant détailler leurs spécificités. L'explication de chaque programme contient un rappel

[14] Le tuple en Python est une liste immuable. À la différence d'une simple liste, ses éléments ne peuvent pas être déplacés ou modifiés.
[15] Par exemple, pour une CR_elle de deux maillons : CR_elle = [(120, 128, g:non), (150, 154, g:non)]

de la modélisation appliquée, sa transcription littéraire conditionnelle ainsi que les résultats rapportés par le programme.

Succession stricte

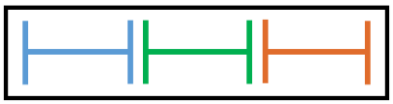


Illustration 52 - Modélisation appliquée et simplifiée de la succession stricte

Recherche appliquée

Si le premier maillon est un maillon_elle et que les maillon_il et les maillon_les ne se trouvent strictement pas avant la fin du dernier maillon_elle
et si le premier maillon après le dernier maillon_elle est un maillon_il et que les maillon_les ne se trouvent strictement pas avant la fin du dernier maillon_il
alors il y a **succession stricte**

Résultats

Niveau	Nombre de fichiers	Occurrences	Proportion occurrences
CE2	32	3	9,38%
CM1	23	1	4,35%
CM2	32	1	3,13%
6e	62	1	1,61%
5e + 4e	36	1	2,78%
3e	31	0	0,00%
M2	11	0	0,00%
Total	**227**	**7**	**2,65%**

Tableau 12 - Nombre d'occurrences de la succession stricte

Le Tableau 12 nous indique que la succession stricte n'est trouvée que 7 fois seulement. Les élèves de primaire semblent utiliser un peu plus cette cohabitation que les collégiens. Un peu moins de la moitié des occurrences se trouvent en CE2. Nous pouvons observer une décroissance de l'utilisation entre le CE2 et le début de collège.

Succession chevauchée minimale groupée

Illustration 53 - Modélisation appliquée et simplifiée de la succession chevauchée minimale groupée

Recherche appliquée

Si le premier maillon est un maillon_elle et les maillons_les ne se trouvent strictement pas avant la fin du dernier maillon_elle **et si** le premier maillon_il se trouve regroupé au dernier maillon_elle
et si le premier maillon_les se trouve regroupé au dernier maillon_il **et s'il** existe un maillon_les

après la fin du dernier maillon_il
alors il y a **succession chevauchée minimale groupée**

Remarque : dans l'application de mes recherches au corpus RésolCo, le regroupement désigne 3 possibilités de relations entre deux maillons. Par exemple, le regroupement du dernier maillon_elle et du premier maillon_il comporte 3 possibilités exclusives :

- Les deux maillons sont superposés
- Le maillon_il est inclus dans le maillon_elle
- Le maillon_elle est inclus dans le maillon_il

Résultats

Niveau	Nombre de fichiers	Occurrences	Proportion occurrences
CE2	32	0	0,00%
CM1	23	0	0,00%
CM2	32	0	0,00%
6e	62	0	0,00%
5e + 4e	36	0	0,00%
3e	31	0	0,00%
M2	11	0	0,00%
Total	**227**	**0**	**0,00%**

Tableau 13 - Nombre d'occurrences de la succession chevauchée minimale groupée

Aucune succession chevauchée minimale groupée n'a été trouvée dans le corpus filtré.

Succession chevauchée minimale distincte

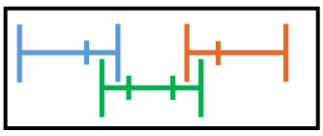


Illustration 54 - Modélisation appliquée et simplifiée de la succession chevauchée minimale distincte

Recherche appliquée

Si le premier maillon est un maillon_elle **et si** le premier maillon_il se trouve strictement après l'avant-dernier maillon_elle et le début du dernier maillon_elle
et si le premier maillon_les se trouve strictement après l'avant-dernier maillon_il et le début du dernier maillon_il
et s'il existe au moins un maillon_les après le dernier maillon_il
alors il y a **succession chevauchée minimale distincte**

Résultats

Niveau	Nombre de fichiers	Occurrences	Proportion occurrences
CE2	32	0	0,00%
CM1	23	0	0,00%
CM2	32	0	0,00%
6e	62	0	0,00%
5e + 4e	36	0	0,00%
3e	31	0	0,00%
M2	11	0	0,00%
Total	**227**	**0**	**0,00%**

Tableau 14 - Nombre d'occurrences de la succession chevauchée minimale distincte

Comme pour la succession chevauchée minimale groupée, la succession chevauchée minimale distincte ne présente aucune occurrence dans le corpus étudié. Ces deux formes de la succession étant très spécifiques, nous attendions de faibles résultats mais pas l'absence complète d'occurrence sur les deux formes. En prenant en considération que le corpus étudié est relativement restreint, il n'est pas impossible qu'en l'élargissant, des occurrences commencent à apparaitre.

Succession chevauchée

Illustration 55 - Modélisation appliquée et simplifiée de la succession chevauchée

Recherche appliquée

Cette forme est la plus complexe des successions. Elle a demandé plus de réflexion afin de définir sa recherche qui présente 4 possibilités distinctes.

1 – Regroupement il + elle

Si le dernier maillon_elle et le premier maillon_il sont regroupés **et s'il** existe un maillon_il strictement après le premier maillon_les **et s'il** existe un maillon_les strictement après le dernier maillon_il **et si** le premier maillon_les ne commence pas strictement avant la fin du premier maillon_il
alors il y a **succession chevauchée**

2 – Regroupement il + les

Si le dernier maillon_il et le premier maillon_les sont regroupés **et s'il** existe un maillon_elle strictement après le premier maillon_il **et s'il** existe un maillon_elle strictement avant le dernier maillon_il **et si** le premier maillon_il ne commence pas strictement avant la fin du premier maillon_elle
alors il y a **succession chevauchée**

3 – Regroupement elle + il et il + les

Si le dernier maillon_elle est regroupé avec le premier maillon_il **et si** le dernier maillon_il est regroupé avec le premier maillon_les
alors il y a **succession chevauchée <u>minimale groupée</u>** (se reporter à la forme d'avant)

4 – Pas de regroupement des premiers et derniers maillons

Si le premier maillon est un maillon_elle **et s'il** existe strictement avant le premier maillon_il **et s'il** existe un maillon_il entre le premier et le dernier maillon_elle **et s'il** existe un maillon_il strictement après le dernier maillon_elle **et si** le premier maillon_les est regroupé au dernier maillon_elle **OU** existe strictement après **et s'il** existe un maillon_les strictement avant le dernier maillon_il **et s'il** existe un maillon_les strictement après le dernier maillon_il **et s'il** existe un maillon_il entre le premier et le dernier maillon_les
alors il y a **succession chevauchée**

<u>Résultats</u>

Niveau	Nombre de fichiers	Occurrences	Proportion occurrences
CE2	32	1	3,13%
CM1	23	2	8,70%
CM2	32	1	3,13%
6e	62	0	0,00%
5e + 4e	36	1	2,78%
3e	31	0	0,00%
M2	11	0	0,00%
Total	**227**	**5**	**2,20%**

Tableau 15 - Nombre d'occurrences de la succession chevauchée

Le Tableau 15 nous indique que la succession chevauchée n'est trouvée que 5 fois, soit une occurrence de moins que pour la succession stricte. Les 4/5 des occurrences se trouvent en primaire et la dernière dans le groupement 5^{e} + 4^{e}, des résultats similaires à ceux de la succession stricte donc.

<u>Fusion</u>

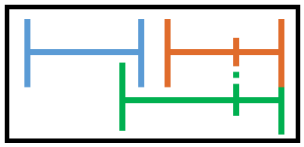


Illustration 56 - Modélisation appliquée et simplifiée de la fusion

<u>Recherche appliquée</u>

En partant de la fin de la liste de tuples de la CR_elle ou de la CR_il et de la CR_les enfants, **si** les derniers maillons respectifs sont superposés, **alors** il y a **fusion** de longueur 1. Nous remontons les listes en regardant si chaque maillon antérieur d'une CR est superposé au maillon antérieur de l'autre. **Tant que** c'est le cas, la longueur de la fusion s'incrémente de 1. Lorsque les maillons ne sont plus superposés, la longueur de la fusion est enregistrée.

Résultats

Niveau	Nombre de fichiers	Occurrences fusion il+les	Proportion occurrences	Occurrences fusion elle+les	Proportion occurrences
CE2	32	21	65,63%	16	50,00%
CM1	23	17	73,91%	13	56,52%
CM2	32	24	75,00%	10	31,25%
6e	62	40	65,52%	25	40,33%
5e + 4e	36	23	63,89%	5	13,89%
3e	31	17	54,84%	6	19,35%
M2	11	2	18,18%	1	9,09%
Total	**227**	**144**	**63,72%**	76	33,63%

Tableau 16 - Nombre d'occurrences des fusions « simples »

Dans le Tableau 16, nous pouvons remarquer que la fusion la plus fréquente est celle du référent singulier « il » avec le référent pluriel « les enfants », présente dans près de deux tiers des copies étudiées. Cette fusion est plus présente dans les faibles niveaux scolaires comme l'illustrent les cellules en vert. Ce schéma se répète, de façon atténuée, pour la fusion du référent « elle » dans « les enfants ». Les CE2 et les CM1 sont toujours présents dans les trois niveaux scolaires l'exploitant le plus mais on retrouve les 6e à la place des CM2 dans ce type de fusion. Si la fusion de « elle » et « les enfants » ne se trouve que dans un tiers des copies, elle reste majoritairement utilisée par les primaires.

De plus, s'il existe une fusion elle+les et il+les dans le même fichier il y a alors **double fusion** de longueur minimale 1. Afin de connaitre la longueur de la double fusion, il faut alors prendre la longueur minimale entre les deux fusions relevées.

Résultats

Niveau	Nombre de fichiers	Occurrences double fusion	Proportion occurrences
CE2	32	14	43,75%
CM1	23	12	52,17%
CM2	32	10	31,25%
6e	62	20	32,26%
5e + 4e	36	5	13,89%
3e	31	5	16,13%
M2	11	0	0,00%
Total	**227**	**66**	**29,20%**

Tableau 17 - Nombre d'occurrences de la double fusion

Les copies présentes dans le Tableau 17 sont les copies qui contiennent une fusion il+les et une fusion elle+les. Ces fusions peuvent se produire à des moments différents dans la CR_les enfants mais cela veut dire qu'au moins le dernier maillon de cette CR est une superposition d'un maillon « elle », d'un maillon « il » et d'un maillon « les enfants », d'où la double fusion. Cette forme spéciale de la fusion se trouve dans près d'un tiers des copies. Elle est plébiscitée par les mêmes niveaux scolaires que la fusion elle+les, et suit la même tendance. Cette tendance similaire est cohérente puisque le total de double fusion ne peut dépasser le total de fusion

elle+les mais elle révèle également que s'il existe une fusion elle+les dans une copie, cela implique très probablement la présence d'une double fusion.

Association/dissociation

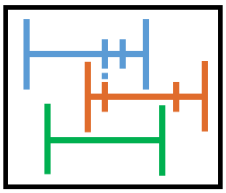


Illustration 57 - Modélisation appliquée et simplifiée de l'association/dissociation

Recherche appliquée

Depuis la liste des tuples de la CR_elle ou de la CR_il et de la CR_les enfants, **si** des maillons entre les deux CR sont superposés, le nombre de maillons superposés sans interruption donne une longueur d'association. **Si** les maillons se dissocient, cette longueur est conservée. Il y a **alors association/dissociation**. L'opération est répétée jusqu'à la fin des deux listes. Si la dernière association se poursuit jusqu'au dernier maillon des deux CR confrontées, elle n'est pas prise en compte puisqu'il s'agit alors d'une fusion.

Résultats

Niveau	Nombre de fichiers	Occurrences A/D il+les	Proportion occurrences	Occurrences A/D elle+les	Proportion occurrences
CE2	32	12	37,50%	6	18,75%
CM1	23	8	34,78%	4	17,39%
CM2	32	19	59,38%	9	28,13%
6e	62	39	62,90%	25	40,32%
5e + 4e	36	19	52,78%	18	50,00%
3e	31	22	70,97%	9	29,03%
M2	11	4	36,36%	1	9,09%
Total	**227**	**123**	**54,42%**	72	31,86%

Tableau 18 - Nombre d'occurrences des A/D

Dans le Tableau 18, nous pouvons observer que l'association/dissociation la plus fréquente est celle du référent singulier « il » avec le référent pluriel « les enfants », présente dans un peu plus de la moitié des copies étudiées. Si la fusion, manipulation plus « simple » des référents, devait se trouver plus fréquemment chez les élèves de primaire, les niveaux scolaires utilisant l'association/dissociation devraient être plus élevés. Les résultats présentés vont dans ce sens. En effet, un décalage marqué s'effectue vers les collégiens avec les plus grandes proportions de copies contenant des A/D il+les en CM2, 6e et 3e. Des résultats confirmés avec les plus grandes proportions de copies contenant des A/D elle+les se trouvant sur la période collège complète.

Au même titre que la double fusion, s'il existe au moins une A/D elle+les et une A/D il+les dans le même fichier il y a alors **double A/D**.

Résultats

Niveau	Nombre de fichiers	Occurrences double A/D	Proportion occurrences
CE2	32	4	12,50%
CM1	23	3	13,04%
CM2	32	9	28,13%
6e	62	23	37,10%
5e + 4e	36	16	44,44%
3e	31	8	25,81%
M2	11	1	9,09%
Total	**227**	**64**	**28,32%**

Tableau 19 - Nombre d'occurrences de la double A/D

Le Tableau 19 présente les résultats de cette recherche. La double A/D se trouve dans un peu moins d'un tiers des copies. Contrairement à la double fusion, la présence des deux types d'A/D n'implique pas forcément une superposition momentanée des trois types de maillons étudiés. La double A/D indique seulement qu'au moins une association/dissociation de chaque type est réalisée dans les copies. Comme dans les résultats de la double fusion, la double A/D suit la tendance de l'A/D elle+les et se trouve donc le plus fréquemment dans les copies de collégiens.

Association/dissociation **puis** fusion

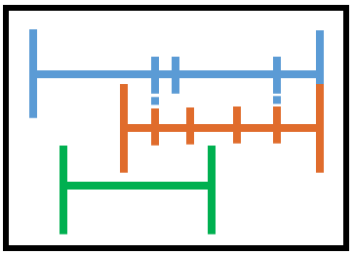


Illustration 58 - Modélisation appliquée et simplifiée de l'association/dissociation puis fusion

Recherche appliquée

Pour déterminer les fichiers qui possèdent une A/D puis fusion, il faut simplement croiser les ensembles contenant les fichiers positifs des deux programmes précédents.

Résultats

Niveau	Nombre de fichiers	Occurrences A/D il+les + fusion il+les	Proportion occurrences	Occurrences A/D elle+les + fusion elle+les	Proportion occurrences
CE2	32	12	37,50%	6	18,75%
CM1	23	8	34,78%	4	17,39%
CM2	32	16	50,00%	7	21,88%
6e	62	34	54,84%	18	29,03%
5e + 4e	36	17	47,22%	3	8,33%
3e	31	16	51,61%	4	12,90%
M2	11	2	18,18%	0	0,00%
Total	**227**	**105**	**46,46%**	42	18,58%

Tableau 20 - Nombre d'occurrences des A/D puis fusion « simples »

Le Tableau 20 présente les résultats des A/D puis fusion avec les mêmes référents, autrement dit, « simples ». Sur les 123 copies contenant une A/D il+les, 105 (soit 85,37%) contiennent donc également une fusion de il+les. La répartition de ces copies est identique à celle de l'A/D. Sur les 72 copies contenant une A/D elle+les, 42 (soit 58,33%) contiennent également une fusion elle+les. La répartition de ces copies est cependant identique à celle de la fusion elle+les, et non pas à celle de l'A/D. Ces résultats nous indiquent que l'A/D et la fusion sont liées d'une certaine façon car la présence de l'une implique souvent la présence de l'autre, surtout dans le cas du référent « il ». Ces résultats semblent cohérents puisque la fusion peut être exploitée comme une suite logique après une ou plusieurs A/D.

Niveau	Nombre de fichiers	Occurrences double A/D + double fusion	Proportion occurrences
CE2	32	4	12,50%
CM1	23	3	13,04%
CM2	32	7	21,88%
6e	62	15	24,19%
5e + 4e	36	3	8,33%
3e	31	3	9,68%
M2	11	0	0,00%
Total	**227**	**35**	**15,49%**

Tableau 21 - Nombre d'occurrences de la double A/D puis double fusion

Le Tableau 21 présente les résultats obtenus lors de la recherche de la forme double A/D puis double fusion. Seules 15,49% des copies contiennent cette forme spécifique. Comme dans les résultats de la double fusion, la double A/D puis double fusion suit la tendance de l'A/D elle+les puis fusion elle+les puisqu'elle ne peut pas la dépasser. Cette forme se fait plus présente dès la fin de primaire avec un pic en première année de collège avant de s'effacer rapidement dans les années suivantes.

Il est également possible de mesurer la répartition des A/D puis fusion « croisées » en fonction du niveau scolaire, c'est-à-dire que le référent singulier de l'A/D n'est pas le même que celui de la fusion. Nous pouvons également mesurer la répartition des A/D puis double fusion.

Résultats

Niveau	Nombre de fichiers	Occurrences A/D il+les + fusion elle+les	Proportion occurrences	Occurrences A/D elle+les + fusion il+les	Proportion occurrences
CE2	32	6	18,75%	4	12,50%
CM1	23	5	21,74%	4	17,39%
CM2	32	8	25,00%	7	21,88%
6e	62	17	27,42%	20	32,26%
5e + 4e	36	3	8,33%	15	41,67%
3e	31	6	19,35%	4	12,90%
M2	11	0	0,00%	0	0,00%
Total	**227**	45	19,91%	**54**	**23,89%**

Tableau 22 - Nombre d'occurrences des A/D puis fusion « croisées »

Le Tableau 22 présente les résultats issus de la recherche de ces formes très spécifiques. Elles se trouvent moins fréquemment que les A/D « simples » et les A/D puis fusion. L'A/D il+les puis fusion elle+les se trouve un peu plus fréquemment en fin de primaire que sa forme opposée qui, elle, se trouve plus fréquemment et de façon plus marquée en milieu de collège.

Niveau	Nombre de fichiers	Occurrences A/D il+les + double fusion	Proportion occurrences	Occurrences A/D elle+les + double fusion	Proportion occurrences
CE2	32	6	18,75%	4	12,50%
CM1	23	5	21,74%	4	17,39%
CM2	32	8	25,00%	7	21,88%
6e	62	16	25,81%	16	25,81%
5e + 4e	36	3	8,33%	3	8,33%
3e	31	5	16,13%	3	9,68%
M2	11	0	0,00%	0	0,00%
Total	**227**	**43**	**19,03%**	37	16,37%

Tableau 23 - Nombre d'occurrences des A/D puis double fusion

Le dernier tableau de cette section, le Tableau 23, nous indique que l'A/D il+les puis double fusion suit de manière très rapprochée la tendance de l'A/D il+les puis fusion elle+les. Nous pourrions attendre la même chose pour la forme opposée mais ce n'est pas le cas. Au contraire de l'A/D elle+les puis fusion il+les se trouve plutôt en fin de primaire et pendant la première année de collège, de façon similaire à la double A/D puis double fusion.

3 – Critiques sur les programmes

Si ces programmes fonctionnent bien, ils ne sont pas parfaits. Premièrement, aucun des programmes ne prend en compte les incertitudes de délimitation ou de rattachement liées aux maillons. La prise en compte de ces caractéristiques pourrait permettre de s'assurer que le maillon est bien présent dans la CR concernée et d'écarter, ou de traiter de façon spécifique, ceux qui le ne sont pas de façon certaine. Également, les commentaires, hors « Membre de lesEnfants », ne sont pas traités. Si le traitement sémantique de ces commentaires s'avèrerait complexe, la notification de leur présence pourrait peut-être permettre d'améliorer légèrement les détections de cohabitations. Cependant, les annotations filtrées, issues des annotations « gold », réduisent considérablement la nécessité de prendre en compte les incertitudes et les commentaires car la version « gold » de ces annotations est une version dans laquelle tous les cas de désaccords, et donc d'incertitudes, ont été résolus.

Deuxièmement, la caractéristique de groupe est conservée dans les tuples représentants les maillons de chaque CR mais n'est pas utilisée dans les programmes. La prise en compte de cette caractéristique dans les conditions de cohabitations, notamment pour les associations/dissociations avec l'association d'un référent singulier dans un groupé composé au minimum de lui-même et du référent « les enfants », pourrait permettre de consolider leurs recherches. Cependant, nous n'avons pas considéré qu'il était nécessaire de nous en servir car la superposition est déjà un indicateur de cette formation groupée.

Dernièrement, les programmes ne prennent pas en compte la superposition à +- 1 caractère. En effet, lors de l'annotation, il est assez facile de ne pas superposer parfaitement les maillons même si Glozz permet de visualiser rapidement s'ils le sont ou non. Si l'annotateur place deux

maillons au même endroit pour désigner la même unité linguistique mais qu'un des deux commence un caractère avant, ou finit un caractère après le second, à cause d'une inattention, les programmes présentés ne considèrent pas ces maillons comme superposés. Par conséquent, l'association n'existe pas ou est rompue quand elle devrait être réalisée ou continue.

VI – Analyses sur le corpus filtré

Après avoir relevé des données caractérisant le corpus filtré, y avoir appliqué nos modélisations théoriques et mesuré les occurrences par niveau scolaire pour chacune des cohabitations présentées, nous allons maintenant tenter de mettre en évidence l'existence de liens entre la complexité des copies du corpus, le niveau scolaire et les cohabitations des continuités référentielles.

1 – Lien entre nombre de maillons et longueur de la copie

Dans la partie IV-2, nous avons présenté la répartition du nombre de maillons, ainsi que leurs moyennes pour chaque niveau scolaire. Si le dénombrement des maillons pour chaque copie peut donner un aperçu de sa longueur, la mesure n'équivaut pas celle de la longueur réelle des copies. Comme nous l'avons indiqué dans la partie V-2, nous récupérons cette information lors de l'exécution des programmes et nous proposons dans cette partie de vérifier l'existence d'un lien entre le nombre de maillons par copie et sa longueur[16].

Niveau scolaire	Longueur minimale	Longueur maximale	Longueur moyenne	Augmentation
CE2	276	1153	613,47	
CM1	244	1162	664,52	8,32%
CM2	381	2728	960,19	44,49%
6e	296	2391	961,94	0,18%
5e + 4e	473	2684	1355,89	40,95%
3e	532	**4984**	1534,35	13,16%
M2	1031	3740	2223,73	44,93%

Tableau 24 - Longueurs des copies sur le corpus filtré

Le Tableau 24 présente la répartition des différentes longueurs relevées en fonction du niveau scolaire. Pour les trois premières colonnes, les cellules en vert représentent les trois valeurs les plus élevées de la colonne. La copie la plus courte se trouve en CE2 avec 276 caractères et la plus longue en 3e avec 4984 caractères. L'étendue de la série est de 4708. Elle exprime un écart notable entre la plus courte et la plus longue copie. Concernant les longueurs moyennes, les trois valeurs les plus élevées correspondent aux trois niveaux scolaires les plus élevés, dans l'ordre croissant. À partir des longueurs moyennes, nous pouvons calculer le taux d'augmentation d'un niveau à l'autre, les résultats se trouvent dans la dernière colonne. Nous observons une augmentation de la longueur moyenne des copies entre chaque niveau scolaire. L'augmentation la plus forte est celle qui correspond au passage de la 3ème au Master 2. Un

[16] L'unité utilisée pour les mesures des longueurs des copies est celle du caractère.

résultat cohérent puisque c'est le plus grand écart entre deux niveaux dans notre jeu de données. Cependant, l'augmentation entre le CM1 et le CM2 est quasiment aussi forte et l'augmentation entre la 6ème et le groupe 5ème + 4ème n'est pas très éloignée. Nous n'avons pas réussi à émettre une hypothèse valable pour expliquer cette observation.
Concernant l'augmentation maximale entre le CE2 et le M2, elle est de 262,48%. Les copies de Master sont donc, en moyenne, 2,5 fois plus longues que celles des CE2. Cette augmentation de la longueur explique donc en partie pourquoi le temps traitement nécessaire à l'annotation de ces deux niveaux est aussi différent.

Nous avons également relevé la distribution des longueurs des copies afin d'avoir un aperçu différent de ces données. Nous avons arbitrairement délimité des segments de longueur de 400 caractères. En complément de la longueur moyenne, nous avons également calculé la médiane des longueurs qui est de 891.

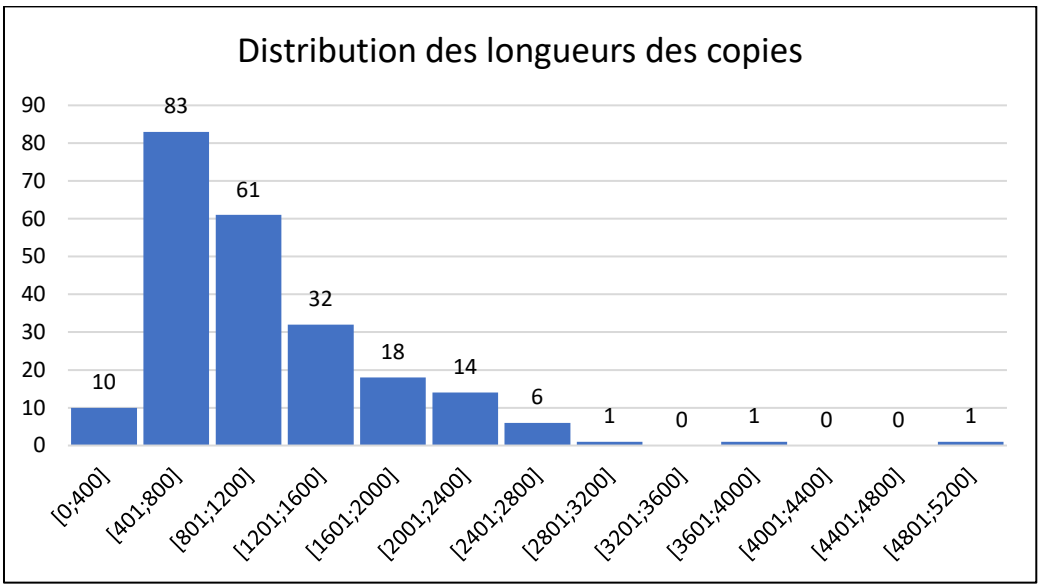


Illustration 59 - Distribution des longueurs des copies sur le corpus filtré

L'Illustration 59 représente la distribution des longueurs et nous pouvons observer qu'elle n'est pas uniforme. Peu de copies font moins de 400 caractères et nous y trouvons des copies de primaire et de 6e. Un peu plus d'un tiers des copies (36,56%) se trouvent dans le segment le plus fourni [401;800]. Le second segment le plus volumineux est le suivant [801;1200] avec un peu plus d'un quart (27,31%) des copies. Ces deux segments cumulent 63,44% du total des copies, soit plus de la majorité. La décroissance se poursuit ainsi jusqu'au segment [2401,2800]. La suite entre 2801 et 5200 caractères comporte 3 copies exceptionnellement longues.

Maintenant que nous avons présenté les données concernant les longueurs relevées, nous pouvons les croiser avec les dénombrements des maillons.

Niveau scolaire	Nombre moyen de maillons par copie	Longueur moyenne par copie
CE2	24,63	613,47
CM1	27,30	664,52
CM2	36,06	960,19
6e	38,42	961,94
5e + 4e	46,86	1355,89
3e	46,42	1534,35
M2	54,27	2223,73

Tableau 25 - Comparaison nombre moyen de maillons et longueur moyenne sur le corpus filtré

Le Tableau 25 compare le nombre moyen de maillons par niveau avec la longueur moyenne des copies par niveau. Le nombre moyen de maillons est obtenu en additionnant les moyennes de chaque type de maillon présentes dans l'Illustration 25. Pour vérifier s'il existe une corrélation entre ces variables, nous avons réalisé un nuage de points représenté sur l'Illustration 60.

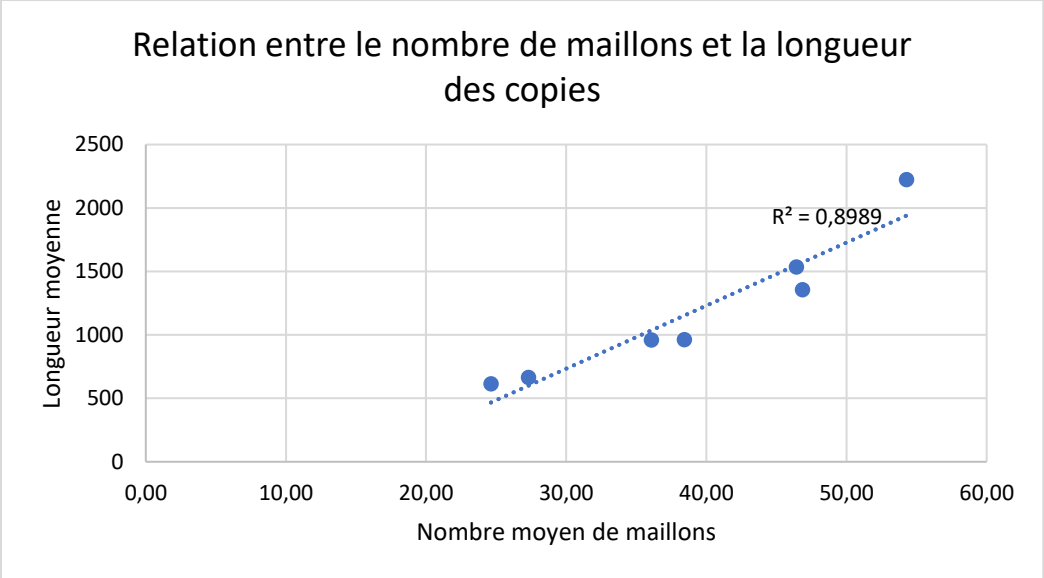


Illustration 60 - Relation entre le nombre de maillons et la longueur des copies

Chaque niveau scolaire y est représenté par un point. À partir de ces points, à l'aide d'Excel, nous avons réalisé un modèle de régression linéaire appelé « ajustement affine » (ou droite de régression). Ce modèle nous permet d'établir la relation linéaire entre nos données en traçant une droite approchant au mieux notre nuage de points. Cette droite est calculée via la méthode des moindres carrés. L'inclinaison de la pente nous indique le type de corrélation. Dans notre cas, la pente est montante. La corrélation entre nos variables est donc **positive**. Afin de vérifier la qualité de la prédiction de notre modèle de régression linéaire, nous avons calculé le coefficient de détermination linéaire de Pearson (R^2). Ce coefficient varie entre 0 et 1. Dans notre cas, et comme indiqué sur l'Illustration 60, il est d'environ 0,90. Le pouvoir prédictif de notre modèle est donc fort et nous pouvons nous y fier. Afin de mesurer la qualité de la corrélation, nous avons calculé le coefficient de corrélation. Ce coefficient varie entre -1 et +1. Dans notre cas, il est d'environ 0.95, ce qui nous indique que la corrélation est **très forte**. Il y

a donc une **corrélation positive très forte** entre le nombre moyen de maillons par copie et sa longueur.

D'après nos calculs, nous avons prouvé la corrélation suivante : plus le niveau scolaire augmente, plus les copies sont longues, et plus elles contiennent de maillons.

3 – Lien entre complexité des copies et cohabitations

Nous savons maintenant que la longueur des copies et le nombre de maillons sont corrélés. Une plus grande quantité de maillons peut amener à une plus grande complexité dans la gestion des référents. Ainsi les copies les plus longues, et donc les plus complexes, devraient favoriser les cohabitations complexes comme l'association/dissociation. C'est ce que nous allons vérifier dans cette partie.

Type de cohabitation	Moyenne des maillons elle	Moyenne des maillons il	Moyenne des maillons les enfants	Moyenne totale des maillons	Écart avec la moyenne globale
Succession stricte (7)	3,71	5,00	4,43	13,14	-25,08 points
Succession chevauchée (5)	10,40	10,20	4,40	25,00	-13,22 points
Global (227)	**13,89**	**15,48**	**8,85**	**38,22**	**0 points**
Fusion il (144)	13,89	17,81	9,20	40,90	2,68 points
Fusion elle (76)	19,17	15,51	7,88	42,57	4,35 points
Double fusion (66)	18,97	16,59	7,41	42,97	4,75 points
A/D il (123)	15,39	18,85	11,30	45,54	7,32 points
A/D elle (72)	19,83	19,38	11,11	50,32	12,1 points
Double A/D (64)	20,61	20,36	11,61	52,58	14,36 points
Double A/D puis double fusion (35)	23,91	20,26	8,80	52,97	14,75 points

Tableau 26 - Moyennes des nombres de maillons par type en fonction des cohabitations

Comme nous savons que la longueur de la copie et les nombres de maillons sont liés, nous avons choisi de présenter le Tableau 26 sous l'angle des maillons car nous pouvons détailler un peu plus les résultats qu'en présentant les longueurs. Le tableau présente uniquement les principaux types de cohabitations qui n'ont pas de résultat nul[17].

La première colonne à gauche indique le type de cohabitation et son nombre d'occurrences. Les quatre suivantes représentent les moyennes du nombre de maillons en fonction de leurs types, ou du total, mesurées sur les copies qui présentent le type de cohabitation de la première colonne. La dernière colonne, la plus à droite, représente l'écart entre la moyenne totale du nombre de maillons par copie mesurée pour chaque cohabitation et celle sur la totalité du corpus filtré. Les données correspondant aux résultats issus du corpus filtré complet sont mises en évidence dans la ligne en orange. Les lignes du tableau sont ordonnées en fonction de leur écart mesuré dans la dernière colonne.

[17] Le tableau complet est présent en Annexe 1 – Tableau de la relation entre cohabitations et longueurs des copies.

Cet écart, en nombre de points, s'avère correspondre à la complexité que nous avons estimée pour chaque cohabitation. En effet, la cohabitation estimée la plus simple, la succession stricte se trouve en première position car elle possède l'écart le plus faible. La double association/dissociation, ainsi que la double A/D puis double fusion, les formes de cohabitations estimées les plus complexes se trouvent en dernières positions car leurs écarts sont les plus élevés. Ce tableau met donc en évidence la relation entre le nombre moyen de maillons par copie et la complexité de la cohabitation utilisée. En effet, plus l'écart avec la moyenne globale augmente, plus il y a de maillons par copies et plus ces copies présentent des cohabitations complexes.

Ce tableau permet également de vérifier la présence d'un type de maillon en fonction d'une cohabitation. Par exemple, les cohabitations avec l'inclusion du référent singulier « il » dans le référent pluriel « les enfants » présentent, en moyenne, plus de maillons associés au référent « il » que de maillons associés au référent « elle ». L'inverse est également vrai pour l'inclusion du référent singulier « elle » dans le référent pluriel « les enfants » mais ce qui est intéressant, c'est que c'est également ce type de maillon qui est le plus présent quand les cohabitations présentent les deux inclusions. C'est-à-dire que la double fusion, la double A/D et la double A/D puis double fusion présentent, en moyenne, toutes les 3 un nombre de maillons « elle » plus important que ceux des deux autres types. Cette forte présence du type de maillon « elle » pourrait s'expliquer par le fait que le début de la CR se trouvant plus tôt que celui des autres, le nombre de maillons entre le début et la fin de la CR_elle est régulièrement plus élevé car la CR parcourt souvent la grande majorité du texte, surtout dans le cas de la double fusion.
Un résultat corroboré par la position inférieure des maillons « il », sauf dans ses propres cohabitations, ainsi que la position des maillons « les enfants », toujours derniers. La CR_les enfants est souvent plus courte du fait de sa position dans la consigne mais aussi à cause de la tournure de la phrase qui incite à la placer en phrase finale du récit. Des CR plus courtes impliquent donc des dénombrements de maillons plus faibles.

Ces observations sont différentes pour les successions. En gardant à l'esprit que leurs occurrences sont bien plus faibles, nous pouvons tout de même remarquer que pour la succession stricte, et donc le traitement « pas à pas », le nombre de maillons « elle » est inférieur au nombre de maillons « il », et même au nombre de maillons « les enfants ». Ainsi la première CR traitée est plus courte que les deux suivantes. Le Tableau 26 nous apprend également que ce type de cohabitation ne se trouve que dans des copies très courtes avec une moyenne de maillons dépassant légèrement 13, quand la moyenne du corpus filtré est à un peu plus de 38. Pour la succession chevauchée, les résultats évoluent un peu. Si les copies sont toujours plus courtes que la moyenne, elles le sont moins. Le nombre de maillons « les enfants » reste toujours inférieur aux deux autres types mais les nombres de maillons associés aux référents « elle » et « il » se rapprochent au point de presque s'égaliser.

4 – Lien entre niveau scolaire et cohabitations

Avec les échantillons et les résultats présentés, il est en l'état difficile de répondre définitivement à la question : **existe-t-il un lien entre le niveau scolaire et le type de cohabitation utilisé ?** Même si nos résultats nous donnent certains indices sur de potentielles préférences de cohabitations des continuités référentielles en fonction du niveau scolaire, les variables sont peut-être indépendantes. Alors pour confirmer ou infirmer cette hypothèse, nous

avons décidé d'utiliser le test du χ^2 (khi-deux). Nous avons donc repris les résultats des répartitions des cohabitations en fonction des niveaux scolaires présentés précédemment pour construire la table des effectifs observés (table de contingence). À partir de cette table, nous avons réalisé la table des effectifs attendus. Petite précision cependant, les valeurs minimales conseillées des effectifs attendus pour calculer le khi-deux doivent être d'au moins 5. Les effectifs attendus du niveau Master 2 étant trop faibles (entre 0 et 2 env.), nous avons décidé d'écarter les Master 2 de ce calcul. Pour les mêmes raisons, nous avons également décidé d'écarter les cohabitations de successions (entre 0 et 1 env.) ainsi que la double A/D puis double fusion (plusieurs effectifs attendus inférieurs à 5). Concernant les formes « simples » et « croisées » d'A/D puis fusion, ainsi que les formes d'A/D puis double fusion, comme elles sont moins pertinentes, qu'elles contiennent plusieurs effectifs attendus inférieurs à 5 et considérant la double A/D puis double fusion comme le niveau le plus complexe possible étudié ici, nous ne les avons pas incluses dans cette observation. Le Tableau 27, ci-dessous, représente donc la table de contingence résultante.

	CE2	CM1	CM2	6e	5e + 4e	3e	Totaux
Fusion il+les	21	17	24	40	23	17	142
Fusion elle+les	16	13	10	25	5	6	75
Double fusion	14	12	10	20	5	5	66
A/D il+les	12	8	19	39	19	22	119
A/D elle+les	6	4	9	25	18	9	71
Double A/D	4	3	9	23	16	8	63
Totaux	73	57	81	172	86	67	536

Tableau 27 - Table de contingence

Les deux tables réalisées, nous avons pu calculer le khi-deux qui correspond dans notre cas, à une valeur *p-value = 1,36E-12*. Cette valeur étant largement inférieure au seuil d'erreur de 0,05, l'hypothèse nulle H_0 selon laquelle les variables présentées dans le Tableau 27 sont indépendantes est rejetée. Une liaison entre les données semble donc exister. Afin d'explorer plus en détail ce potentiel lien, nous avons décidé d'effectuer le calcul des résidus de Pearson avec la formule (observé – attendu)/√(attendu), et nous avons obtenu le tableau suivant.

	CE2	CM1	CM2	6e	5e + 4e	3e
Fusion il+les	0,37757394	0,48874664	0,54853966	-0,82472285	0,04534007	-0,17801725
Fusion elle+les	1,81020312	1,77904076	-0,39623288	0,19014843	-2,02758513	-1,10227038
Double fusion	1,18442507	1,42495021	-0,39623288	-0,82904717	-2,02758513	-1,42886902
A/D il+les	1,00435614	0,3704196	2,85804454	3,87236002	2,58452251	4,78713554
A/D elle+les	-2,53541707	-2,43293815	-2,11835092	-2,13391169	-0,250203	-1,52327875
Double A/D	-1,82329819	-1,65600942	-0,52798995	0,04534007	1,36532536	-0,29371348

Tableau 28 - Résidus de Pearson

Le Tableau 28 met en avant l'attraction (les résidus positifs) ou la répulsion (les résidus négatifs) des niveaux scolaires pour les différentes formes de cohabitation de l'association/dissociation uniquement puisque la succession n'a pas pu être prise en compte à cause de ses effectifs attendus inférieurs au minimum requis pour ce calcul. Si l'on s'intéresse aux valeurs extrêmes positives (en vert) et aux valeurs extrêmes négatives (en rouge), nous pouvons observer que :

- L'attraction pour la fusion il+les est un peu plus forte pour les CE2, CM1 et CM2 que pour les collégiens, même si les écarts sont faibles
- La fusion elle+les ainsi que la double fusion marquent quant à elles un écart plus important avec une attraction plus forte chez les CE2 et CM1 que chez les collégiens en 5^e, 4^e et 3^e.
- L'attraction de l'association/dissociation il+les est, à l'inverse, plus importante chez les collégiens que chez les élèves de primaires, notamment pour les 6^e et les 3^e.
- Il y a cependant une répulsion inattendue pour l'association/dissociation elle+les à tous les niveaux. Cette répulsion est tout de même plus faible pour la fin de collège.
- La double association/dissociation marque une forte répulsion dans les niveaux CE2 et CM1 et une attraction notable au niveau $5^e + 4^e$.

Ces résultats renforcent l'idée qu'il existe un lien entre le niveau scolaire et le type de cohabitation de continuités référentielles utilisé. En effet, plus les niveaux scolaires augmentent, plus l'attractivité, et donc la présence, de cohabitations complexes est forte. En parallèle, la répulsion des formes les plus simples augmente.
Et au contraire, plus les niveaux scolaires se réduisent, plus l'attractivité pour les formes de cohabitations simple se renforce, tandis que la répulsion pour les formes complexes augmente.

Cependant, si la faiblesse extrême de la *p-value* mesurée permet d'affirmer avec certitude qu'il existe un lien entre les différents types de cohabitations de CR et le niveau scolaire, l'absence des formes de la succession dans ces résultats ne permet pas une observation idéale. Le jeu de données étant finalement faible, il faudrait réaliser les mêmes calculs sur de plus grands jeux pour s'assurer des conclusions avancées ici.

Conclusion et perspectives

Dans ce mémoire, nous nous sommes appliqués à décrire deux grands types de cohabitations de continuités référentielles que nous avons par la suite automatiquement détectés dans les copies issues du corpus RésolCo. Ce corpus possède la particularité d'être composé d'un ensemble de textes majoritairement produits par des apprenants qui peuvent proposer des structures diverses et variées en réponse à la consigne de RésolCo.

Cette réponse à la consigne implique la manipulation d'au moins trois référents distincts et il existe plusieurs façons de les faire cohabiter. Ce sont ces cohabitations que nous avons étudiées. Dans un premier temps, nous avons modélisé deux grands types de cohabitations en prenant deux extrêmes sur l'échelle de complexité de la manipulation des référents. Dans un deuxième temps, nous avons appliqué ces modélisations au corpus RésolCo et mis en place une détection automatisée des cohabitations dans le corpus. Dans un troisième et dernier temps, nous avons étudié les résultats avec la volonté de monter l'existence d'un lien entre cohabitations de continuités référentielles et niveau scolaire. Nos calculs nous ont alors permis de conclure en faveur de l'existence d'un tel lien.

En perspective, il serait possible d'étudier les cohabitations de continuités référentielles :

- À travers d'autres types de cohabitations
- En observant la position de la phrase consigne dans chaque CR, ce qui permettrait de vérifier si elle est présente dans les premiers maillons et ainsi appuyer la théorie du traitement des référents « pas à pas » dans le cas de la succession.

- En observant le regroupement des deux référents singuliers entre eux avant ou sans une inclusion dans le référent « les enfants » par la suite.

De plus, afin de contraster les résultats obtenus, il pourrait être intéressant d'étudier les mêmes cohabitations dans des productions répondant à différentes consignes ou dans des genres textuels différents. En effet, le genre textuel semble influer sur la composition des chaines de référence, et donc sur les continuités référentielles, comme l'ont étudié Boudreau (2004), Schnedecker (2005), Schnedecker & Longo (2012, pp. 1957-1959) ainsi que Schnedecker & Landragin (2014, pp. 5-6). Nous pouvons alors nous demander si l'impact du genre textuel sur les éléments linguistiques qui composent les continuités n'en a pas également un sur leurs cohabitations.

En définitive notre étude porte sur un corpus à la taille limitée mais a permis de présenter des résultats encourageants qui ne demandent qu'à être complétés.

Références bibliographiques

Boudreau, S. (2004). Résolution d'anaphores et identification des chaines de coréférence selon le type de texte. Mémoire de Maîtrise, Université de Montréal. https://papyrus.bib.umontreal.ca/xmlui/handle/1866/14866

Democrat. *ANR DEMOCRAT*. Consulté le 7 février 2021, à l'adresse : https://www.lattice.cnrs.fr/democrat/motivations.html

Erkü, F. & Gundel, J. K. (1987). The pragmatics of indirect anaphors. *The Pragmatic Perspective, Amsterdam*, 533-545. https://doi.org/10.1075/pbcs.5.39erk

Garcia-Debanc, C., Ho-Dac, L.-M., Bras, M., & Rebeyrolle, J. (2017). Vers l'annotation discursive de textes d'élèves. *Corpus*, *16*, Article 16. https://doi.org/10.4000/corpus.2783

Garcia-Debanc, C., Rebeyrolle, J. & Ho-Dac, L.-M. (2021). La continuité référentielle dans le corpus Résolco. *Langue française.* Version soumise.

Glozz Annotation Platform. Consulté le 7 février 2021, à l'adresse http://www.glozz.org/

Halliday, M. A., & Hasan, R. (1976). Cohesion in English. London: Longman

Kantor, B. & Globerson, A. (2019). Coreference Resolution with Entity Equalization. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019), Florence, Italy*, 673– 677. https://www.aclweb.org/anthology/P19-1066/

Kleiber, G. (1991). Anaphore-deixis : où en sommes-nous ?. *L'information grammaticale,* 51, 3-18. https://doi.org/10.3406/igram.1991.3231

Landragin, F. (2020). *Rapport final du projet ANR Democrat, "Description et modélisation des chaines de référence : Outils pour l'annotation de corpus et le traitement automatique"* [Research Report]. ANR (Agence Nationale de la Recherche - France). https://hal.archives-ouvertes.fr/hal-02533314

Longo, L. (2014). Vers des moteurs de recherche « intelligents » : un outil de détection automatique de thèmes. Méthode basée sur l'identification automatique des chaines de référence. Thèse, Université de Strasbourg, 2013. https://tel.archives-ouvertes.fr/tel-00939243

Oberle, B. (2017). ODACR (Outil de Détection Automatique des Chaines de Référence). https://hal.inria.fr/hal-01837101

Péry-Woodley, M.-P., Asher, N., Enjalbert, P., Benamara, F., Bras, M., Fabre, C., Ferrari, S., Ho-Dac, L.-M., Le Draoulec, A., Mathet, Y., Muller, P., Prévot, L., Rebeyrolle, J., Tanguy, L., Vergez-Couret, M., Vieu, L., & Widlöcher, A. (2009). ANNODIS : Une approche outillée de l'annotation de structures discursives. *Conférence sur le Traitement Automatique des Langues Naturelles (TALN 2009)*. https://hal.archives-ouvertes.fr/hal-00410590

Riegel, M., Pellat, J.-C., Rioul, R. (2009). Grammaire méthodique du français. Paris, PUF/Quadrige.

Schnedecker, C. (2005). Les chaines de référence dans les portraits journalistiques : éléments de description. *Travaux de linguistique*, 2(2), 85-133. https://doi.org/10.3917/tl.051.0085

Schnedecker, C. (2006). De l'un à l'autre et réciproquement... aspects sémantiques, discursifs et cognitifs des pronoms anaphoriques corrélés l'un-l'autre et le premier-le second (Champs linguistiques Recherches). Bruxelles [Paris: De Boeck-Duculot]. https://doi.org/10.3917/dbu.schne.2006.02

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. https://doi.org/10.4000/praxematique.5339

Schnedecker, C., & Landragin, F. (2014). Les chaines de référence : Présentation. *Langages, Armand Colin (Larousse jusqu'en 2003),* 2014, pp.3-22. https://halshs.archives-ouvertes.fr/halshs-01069451/document

Schnedecker, C., & Longo, L. (2012). Impact des genres sur la composition des chaines de référence : Le cas des faits divers. *SHS Web of Conferences*, 1, 1957-1972. https://doi.org/10.1051/shsconf/20120100061

Urieli, A. (2013). Robust French syntax analysis: reconciling statistical methods and linguistic knowledge in the Talismane toolkit. Thèse doctorale, Université de Toulouse le Mirail – Toulouse II. English. https://tel.archives-ouvertes.fr/tel-01058143/document

Widlöcher, A., & Mathet, Y. (2014). La plate-forme Glozz : Environnement d'annotation et d'exploration de corpus. *Actes de la 16e Conférence Traitement Automatique des Langues Naturelles (TALN'09), session posters, Jun 2009, Senlis, France, France*. https://hal.archives-ouvertes.fr/hal-01011969

Annexes

Annexe 1 – Tableau de la relation entre cohabitations et longueurs des copies

	Moyenne des maillons elle	Moyenne des maillons il	Moyenne des maillon les enfants	Moyenne totale des maillons	Écart avec la moyenne globale	Longueur moyenne	Écart avec la moyenne globale
Succession stricte (7)	3,71	5,00	4,43	13,14	-25,08 points	692,29	-391,94 points
Succession chevauchée (5)	10,40	10,20	4,40	25,00	-13,22 points	738,60	-345,63 points
Global (227)	**13,89**	**15,48**	**8,85**	**38,22**	**0 points**	**1084,23**	**0 points**
Fusion il (144)	13,89	17,81	9,20	40,90	2,68 points	1027,49	-56,73 points
Fusion elle (76)	19,17	15,51	7,88	42,57	4,35 points	970,49	-113,74 points
Double fusion (66)	18,97	16,59	7,41	42,97	4,75 points	940,03	-144,2 points
A/D il (123)	15,39	18,85	11,30	45,54	7,32 points	1163,50	79,28 points
A/D elle (72)	19,83	19,38	11,11	50,32	12,1 points	1238,19	153,97 points
Double A/D (64)	20,61	20,36	11,61	52,58	14,36 points	1281,38	197,15 points
Double AD puis double fusion (35)	23,91	20,26	8,80	52,97	14,75 points	1176,97	92,75 points
Formes spéciales							
A/D il+les puis fusion il+les (105)	14,93	19,50	10,75	45,19	6,97 points	1118,71	34,49 points
A/D il+les puis double fusion (43)	22,30	19,28	8,53	50,12	11,9 points	1098,88	14,66 points
A/D il+les puis fusion elle+les (45)	22,60	19,29	8,89	50,78	12,56 points	1134,71	50,49 points
A/D elle+les puis fusion elle+les (42)	23,43	18,90	9,00	51,33	13,11 points	1171,71	87,49 points
A/D elle+les puis double fusion (37)	23,51	19,78	8,73	52,03	13,81 points	1156,51	72,29 points
A/D elle+les puis fusion il+les (54)	20,59	21,13	10,54	52,26	14,04 points	1229,09	144,87 points
Formes sans occurrences							
Succession chevauchée minimale groupée (0)	0	0	0	0,00	n/a	n/a	n/a
Succession chevauchée minimale distincte (0)	0	0	0	0,00	n/a	n/a	n/a

Tableau 29 - Relations entre cohabitations et longueurs des copies

Déclaration sur l'honneur de non-plagiat

Je soussigné,

Aubertin, Lucas

Régulièrement inscrit à l'Université de Toulouse II Jean Jaurès : 22000781

Année universitaire : 2020 - 2021

certifie que le document joint à la présente déclaration est un travail original, que je n'ai ni recopié ni utilisé des idées ou des formulations tirées d'un ouvrage, article ou mémoire, en version imprimée ou électronique, sans mentionner précisément leur origine et que les citations intégrales sont signalées entre guillemets.

Fait à : Toulouse, France

Le : 21 / 06 / 2021

Signature :