

Doctorat de l'Université de Toulouse

préparé à l'Université Toulouse - Jean Jaurès

Vers une typologie des chaînes de référence à la lumière de
corpus annotés diversifiés

Thèse présentée et soutenue, le 3 octobre 2025 par

Silvia FEDERZONI

École doctorale

CLESCO - Comportement, Langage, Éducation, Socialisation, Cognition

Spécialité

Sciences du langage

Unité de recherche

CLLE - Unité Cognition, Langues, Langage, Ergonomie

Thèse dirigée par

Cécile FABRE et Lydia-Mai HO-DAC

Composition du jury

M. Frédéric LANDRAGIN, Président, CNRS Île-de-France Gif-sur-Yvette

Mme Marion FOSSARD, Rapporteuse, Université de Neuchâtel

Mme Josette REBEYROLLE, Examinatrice, Université Toulouse - Jean Jaurès

Mme Catherine SCHNEDECKER, Examinatrice, Université de Strasbourg

Mme Cécile FABRE, Directrice de thèse, Université Toulouse - Jean Jaurès

Mme Lydia-Mai HO-DAC, Co-directrice de thèse, Université Toulouse - Jean Jaurès

À Eleonora

Remerciements

On décrit souvent la thèse comme un travail solitaire : une épreuve d'endurance, de patience, de confrontation avec soi-même. Un peu comme un match de tennis : seul-e sur le terrain, face à l'adversaire, face à soi-même, face à la longueur du combat. Mais, comme au tennis, on oublie parfois que si l'on avance, ce n'est jamais vraiment seul-e. Derrière chaque coup joué, chaque effort, il y a des soutiens, visibles ou discrets, des présences qui encouragent, qui conseillent, qui tendent une main. Il y a celles et ceux qui posent les bases, relancent, soutiennent sans relâche. Alors, même si ce chemin semble solitaire, il ne l'est pas. Si aujourd'hui j'arrive à ce point du parcours, c'est grâce à toutes celles et ceux qui, à leur manière, m'ont accompagnée. C'est donc à cette « équipe », dans toutes ses dimensions, que je voudrais ici exprimer ma gratitude.

En premier lieu, je tiens à remercier mes directrices, Cécile Fabre et Lydia-Mai Ho-Dac, qui, à la manière de coachs, attentives et exigeantes, ont su guider, orienter, questionner, pousser à aller plus loin, toujours dans le souci de faire progresser le jeu et la réflexion. Elles ont su accueillir les doutes, les hésitations, les imprévus et les difficultés rencontrées. Leur regard, leurs remarques, leur rigueur et leur confiance ont été des repères précieux tout au long de ce chemin. Grâce à elles, ce travail a été possible et, au-delà de l'apprentissage scientifique, il a aussi été une expérience humaine enrichissante. Les mots peinent à dire toute la reconnaissance que je ressens envers elles pour m'avoir accompagnée jusqu'à jouer la dernière balle.

Je remercie sincèrement les rapporteur·rice·s, Marion Fossard et Frédéric Landragin, pour avoir accepté de jouer le rôle de juges dans ce match final. J'adresse également toute ma gratitude aux examinatrices, Josette Rebeyrolle et Cathérine Schne-decker, pour avoir accepté de rejoindre le terrain et de faire partie du jury. C'est un honneur de pouvoir compter sur l'expérience de ces quatre « joueur·se·s professionnel·le·s » dont l'expertise et les conseils permettront sans doute d'enrichir cette recherche. Pour leur engagement pendant les comités de suivi de thèse, je tiens à remercier Josette Rebeyrolle et Jean-Yves Antoine qui, comme les analystes d'un match, ont contribué par leurs lectures attentives, leurs remarques et leurs conseils à affiner le jeu, à pousser la réflexion plus loin et à soutenir l'aboutissement de ce travail.

Je tiens à remercier Marie-Pierre Escoubas-Benveniste. C'est sous sa direction, lors de mon mémoire de master à l'Université La Sapienza de Rome, que j'ai découvert la linguistique et la recherche. Cette première mise en jeu a largement orienté la suite du match.

Je souhaite remercier Alexandre Pous, Léa Samama et Johann Tournebize, qui ont contribué avec sérieux et engagement à l'annotation des données. Leur aide précieuse a rendu possible cette étape qui me tenait à cœur.

Je souhaite maintenant exprimer ma gratitude au laboratoire CLLE. Ce laboratoire a été, tout au long de ce parcours, un terrain de confrontation d'idées et de soutien. Par leurs conseils, leurs retours, la richesse de nos échanges — lors des séminaires, des réunions ou dans les discussions plus informelles — les membres du laboratoire ont contribué à créer un espace d'entraînement stimulant, jamais trop isolé. Je pense en particulier à Anne Condamines, grâce à qui la jeune joueuse tout juste sortie de

master a pu découvrir le laboratoire et entrevoir des perspectives de recherche qu'elle n'imaginait pas accessibles. Je la remercie pour les échanges et les collaborations toujours enrichissantes. Encore une fois, merci à Josette Rebeyrolle, pour la richesse de nos discussions, ses conseils de lecture, son écoute attentive et toujours bienveillante, son soutien sans faille — et aussi pour les conseils de théâtre, qui ont offert quelques respirations bienvenues hors du terrain. Sa présence a été précieuse tout au long de ce parcours, en particulier dans les moments difficiles.

Comme dans tout parcours sportif, il arrive qu'on change de club, pour explorer de nouveaux terrains, ou tout simplement parce que c'est la vie, avec ses contraintes et ses hasards. C'est ainsi que j'ai eu la chance de rejoindre, pendant un temps, l'équipe orléanaise du LLL. Merci à tous les membres de ce laboratoire qui m'ont accueillie avec bienveillance et ont su alimenter les échanges scientifiques tout en ouvrant d'autres perspectives, humaines et intellectuelles. Un merci particulier à Anne-Lyse Minard-Forst pour son soutien, sa compréhension et son humanité.

Dans le cadre orléanais, je souhaite aussi remercier Anaïs Lefevre-Halftermeyer. J'ai eu de la chance de la croiser sur ce chemin. Elle a su être là, avec attention, patience, respect. Un de ces soutiens rares dont la présence, discrète mais solide, fait toute la différence. À des moments où tout me semblait inaccessible, elle m'a aidée à grimper. Dans la paroi raide de mes journées, elle m'a offert quelques prises pour ne pas glisser, et surtout, pour continuer à monter, jusqu'à croire possible d'atteindre le sommet. Ces quelques lignes ne suffisent pas à exprimer l'ampleur de ma gratitude — mais elles en portent, je l'espère, l'essentiel.

À toutes celles et ceux qui, comme moi, jouent ce sport si particulier qu'est la thèse, je veux aussi adresser ces mots. Les autres doctorant-e-s, les partenaires d'entraînement, les compagnons de galère avec qui on partage les doutes, les pauses-café, les discussions tardives, les « je n'y arriverai jamais » comme les petites victoires. Parce qu'il y a, dans ces échanges, cette solidarité précieuse de celles et ceux qui savent. Merci à Lockman Aalioui, Cécile Ferlin, Killyam Forge, Julie Humbert-Droz, Mariame Maarouf, Dimitri Paisios, Daniele Sanacore, Marine Wauquier... et à tant d'autres, bien trop nombreux et nombreuses pour tou-te-s les citer. Je pense aussi à toutes celles et ceux avec qui j'ai eu la chance de m'engager au-delà de mon travail individuel : dans la représentation des doctorant-e-s, au sein du CEPEL ou encore dans l'équipe de la NewsCLLEtter. Ces aventures partagées ont été profondément formatrices, et ont donné à ce parcours une dimension collective que je n'oublierai pas.

Parmi ces compagnons de route, je pense particulièrement à Claire Ibarboure, avec qui j'ai partagé bien plus qu'un bureau. Une forme de quotidien complice, fait d'échanges, de concentration partagée, de silences respectueux et de soutien discret.

Je remercie Victoria O'Callaghan, pour sa présence attentive, pour les discussions profondes et les regards qui n'avaient pas toujours besoin de mots. Sa joie de vivre, contagieuse, a souvent été une bouffée d'air bienvenue dans le quotidien.

Je pense aussi à Chiara Minoccheri, avec qui le lien tissé autour du doctorat a très vite dépassé le seul terrain académique. Je la remercie pour le soutien mutuel,

les doutes partagés, les plaintes, les idées... et cette complicité qui, je l'espère, nous accompagnera bien au-delà de cette aventure.

Je remercie Julie Rouaud pour les conversations précieuses, les idées partagées, les moments d'échange qui ont compté. Une amitié née dans le cadre du doctorat, qui perdure et dont je mesure la chance.

À Filip, un merci tout spécial. Je ne saurai pas dire l'importance de son amitié avec ces quelques mots. Présent dans les creux comme dans les élans, il a su être là — vraiment là. Je le remercie pour les aventures toulousaines, les conversations infinies, les silences qui n'ont rien d'un vide, les rires qui font respirer, son soutien indéfectible, et bien plus encore.

Il y a aussi, autour du terrain, toutes celles et ceux sans qui ce chemin aurait été encore plus rude. Je pense tout particulièrement à Mathilde, rencontrée à la fac mais très vite devenue bien plus qu'une collègue. Une amie précieuse, avec qui un lien sincère et profond s'est tissé. Je la remercie pour les échanges vrais, l'attention discrète mais constante, son soutien et cette présence qui compte, même quand les chemins s'éloignent un peu. Son amitié compte plus que ces quelques mots ne sauraient le dire. Merci aussi à sa famille — Guillaume, Léo et Ben — pour les moments passés ensemble, empreints de leur joie de vivre.

Je voudrais aussi remercier Alexandra, Biagio, Joséphine, Morgane et Sebastian pour les discussions enrichissantes et les promenades orléanaises.

Je souhaite également remercier celles et ceux qui, loin du terrain académique, m'ont offert de vrais moments de répit par leur présence et leur légèreté, et qui ont fait de mes escapades italiennes de vraies bouffées d'oxygène. Je pense notamment à Giada, Maria et Andrea, Roxy, Lilla, Mimmo, Federica, Paola, Lorena, pour avoir cru en moi, même quand moi-même je doutais. Un grand merci à Carola, dont l'amitié fidèle défie les kilomètres.

Je pense aussi à ma famille. À mes parents, pour leur présence, même à distance, même dans la complexité des choses. À ma sœur Alessandra pour la confiance qu'elle m'a toujours témoignée. À ma sœur Serena pour notre lien unique que rien ne remplace. *A tutta la mia famiglia, grazie per avermi lasciata andare, anche quando il cuore avrebbe voluto il contrario.*

Enfin, je souhaite remercier Eleonora, à qui cette thèse est dédiée. Je lui suis reconnaissante d'avoir toujours été là, malgré la distance, malgré les épreuves, avec cette attention discrète, cette bienveillance, cette confiance sans faille qui lui appartenaient. Son amitié et sa confiance ont compté, silencieusement mais profondément, tout au long de ce chemin.

Bref, merci à toutes celles et ceux qui se reconnaîtront entre ces lignes — même s'ils ne sont pas cité·e·s, par oubli ou par manque de place — et qui, chacun·e à leur manière, ont accompagné ce parcours, offert de l'élan, de l'appui, du souffle.

Ça y est... La dernière prise est atteinte. Le sommet est là. Il ne reste qu'à poser un instant le regard sur le chemin parcouru.

La dernière balle est jouée : game, set, match. Cette victoire est celle d'une équipe.

Table des matières

Introduction	1
I Les chaînes de référence (CR) : un phénomène complexe	7
1 Les chaînes de référence : un phénomène central du discours	11
1.1 Une suite d'expressions référentielles	12
1.1.1 Position des maillons	14
1.1.2 Relation entre les maillons	16
1.2 Un indice de cohésion	17
1.3 Un héritage diversifié	20
1.3.1 Héritage cognitif	20
1.3.2 Héritage psycholinguistique	23
1.3.3 Héritage linguistique	24
1.4 Bilan	26
2 Description des chaînes de référence	29
2.1 Composition des chaînes de référence : critères <i>micro</i>	31
2.1.1 Nature grammaticale des maillons	31
2.1.2 Fonction syntaxique	33
2.1.3 Coefficient de stabilité	33
2.1.4 Autres critères	34
2.2 Dimension des chaînes de référence : critères <i>macro</i>	36
2.2.1 Longueur des chaînes	36
2.2.2 Portée des chaînes	37
2.2.3 Distance entre les maillons	38
2.2.4 Persistance	41
2.2.5 Densité référentielle des chaînes	42
2.2.6 Bilan sur les critères de description	42
2.3 Facteurs de variation des chaînes de référence	43
2.3.1 Type de texte	43
2.3.2 Genre textuel	44
2.3.3 <i>Nature</i> du référent	45
2.3.4 Cohabitation des chaînes de référence	46
2.3.5 Autres facteurs de variation	48
2.4 Bilan	51

3	Des corpus annotés en référence	53
3.1	En quoi consiste l'annotation de la référence ?	54
3.2	Guides et modèles d'annotation	58
3.3	Deux corpus de référence : AnnoDis et DEMOCRAT	60
3.3.1	Le corpus AnnoDis	61
3.3.2	Le corpus DEMOCRAT	63
3.3.3	Comparaison des modèles d'annotation	64
3.4	Impact du modèle d'annotation sur la description des chaînes de référence	66

II Caractériser les chaînes de référence par une approche linéaire 69

4	Au cœur des données : de la sélection à la description	75
4.1	Le corpus AnnoDis : harmonisation des annotations	77
4.1.1	Format xml avec annotation intégrée	78
4.2	Du corpus DEMOCRAT à la ressource DEMRC	80
4.2.1	Nettoyage des annotations et constitution de la ressource DEMRC	80
4.2.2	Format d'analyse proposé	81
4.2.3	Pré-traitement de la ressource DEMRC	84
4.3	Anatomie des chaînes dans les deux ressources exploitées	86
4.3.1	Longueur des chaînes	87
4.3.2	Distance entre les maillons	90
4.3.3	Nature des maillons	92
4.4	Bilan	95
5	Mise en œuvre de l'approche linéaire via l'analyse des séquences	97
5.1	Les chaînes de référence : des <i>séquences</i> de maillons	98
5.1.1	Qu'est-ce que l'analyse des séquences ?	99
5.2	L'analyse des séquences appliquée aux chaînes de référence	100
5.2.1	Création des séquences	101
5.2.2	Exploration et description des séquences	105
5.2.3	Comparaison des séquences	105
5.2.3.1	Mesures basées sur la distribution des états	109
5.2.3.2	Mesures basées sur le nombre d'attributs communs	110
5.2.3.3	Mesures d'édition	112
5.2.4	Classification des séquences	115
5.2.5	Interprétation des clusters obtenus	117
5.3	Des séquences aux patrons	119
6	Explorer les séquences	123
6.1	Enrichissement des ressources	124
6.1.1	La nature des référents dans AnnoDis	124
6.1.2	La nature des référents dans DEMRC	127
6.2	Paramétrage de l'analyse des séquences	129
6.2.1	Création des séquences	130

6.2.2	Comparaison et classification des séquences	133
6.2.3	Choix du nombre de clusters	135
6.3	Bilan	140
III Vers une typologie des chaînes de référence		143
7	Explorer les enchaînements : AnnoDis	147
7.1	Exploration des séquences	148
7.2	Description des clusters obtenus	148
7.2.1	Cluster C1-3c AnnoDis : SN_{def}^{++}	151
7.2.2	Cluster C2-3c AnnoDis : SN_{ind} , SN_{dem} et Pro	152
7.2.3	Cluster C3-3c AnnoDis : Np et Pro	153
7.3	Lien entre les clusters AnnoDis et les facteurs de variation	155
7.3.1	Impact de la nature du référent sur la composition des chaînes	155
7.3.2	Impact du genre textuel sur la composition des chaînes	156
7.4	Les patrons des chaînes AnnoDis	158
7.5	Bilan	162
8	Explorer les enchaînements : application de l'analyse des séquences à la ressource DEMRC	165
8.1	Projection des patrons AnnoDis sur la ressource DEMRC	166
8.2	Exploration des séquences	170
8.3	Description des clusters obtenus	173
8.3.1	Cluster C1-4c DEMRC : Np et Pro	176
8.3.2	Cluster C2-4c DEMRC : $SN_{\emptyset}$ et SN_{def}	178
8.3.3	Cluster C3-4c DEMRC : SN_{def}^{++}	182
8.3.4	Cluster C4-4c DEMRC : des séquences hétérogènes	186
8.3.5	Bilan	197
8.4	Lien entre les clusters et les facteurs de variation	197
8.4.1	Impact de la nature du référent sur la composition des chaînes	197
8.4.1.1	Référents concrets et abstraits	198
8.4.1.2	Référents humains et non humains	199
8.4.2	Impact du type de texte sur la composition des chaînes	200
8.4.3	Impact du genre textuel sur la composition des chaînes	201
8.5	Bilan	203
8.6	Les patrons des chaînes DEMRC	205
8.7	Bilan	212
Conclusions et perspectives		215
A	Balises xm1 utilisées pour signaler la structure textuelle dans chaque fichier du corpus DEMOCRAT	237
B	Exemples complets des chaînes de référence	239
B.1	diablaucorps_RC_19–21_s–CHAINE–5	239
B.2	articleswiki_RC_19–21_s–CHAINE–1353	253

C	Documentation de la ressource Democrat_RC_19-21	261
D	Guide d'annotation de la nature du référent	271
E	Séquences représentatives : exemples	285
E.1	Cluster C2a-11c DemRC	285
E.2	Cluster C2b-11c DemRC	287
E.3	Cluster C3a-11c DemRC	288
E.4	Cluster C4e-11c DemRC	291
E.5	Cluster C4f-11c DemRC	291
F	Analyse des patrons : lien avec les facteurs de variation	293
F.1	Les patrons AnnoDis et le type de référent	293
F.2	Projection des patrons AnnoDis sur DEMRC : comparaison entre AnnoDis et DEMRC _{Nar-}	294
F.3	Projection des patrons AnnoDis sur DEMRC en fonction du type de texte	295
F.4	Projection des patrons AnnoDis sur DEMRC : comparaison selon la nature du référent	296
G	Les patrons dans DEMRC	297
G.1	Distribution des patrons selon le type de texte	297
G.2	Distribution des patrons selon la nature du référent	298

Table des figures

1	The stars or the humans ?	1
1.1	Échelle d'accessibilité d'Ariel (1990)	22
2.1	Représentation de la distance et de la persistance dans les CR	42
3.1	Trois conceptions de maillons	55
3.2	Représentation schématique de deux chaînes de référence, adaptée de Schnedecker (2021, 55)	72
4.1	Première étape : sélection et pré-traitement des données	76
4.2	Aperçu des CT dans la ressource AnnoDis : version xml avec annotation intégrée	79
4.3	Aperçu des chaînes de référence dans DEMRC : format d'analyse avec annotation intégrée	84
4.4	Extrait du fichier <i>articleswiki_RC_s-CHAINE-285.csv</i>	85
4.5	Longueur des chaînes dans AnnoDis en fonction du genre textuel	87
4.6	Longueur des chaînes dans DEMRC en fonction du type de texte	89
4.7	Longueur des chaînes dans DEMRC en fonction du genre textuel	89
4.8	Visualisation de la chaîne constituée de 895 maillons (<i>s-CHAINE-5</i>) dans le texte <i>Le diable au corps</i> (sous-corpus des <i>romans</i>)	90
4.9	Visualisation de la distribution de la nature des maillons dans AnnoDis et DEMRC _{Nar-}	94
4.10	Visualisation de la distribution de la nature des maillons dans DEMRC en fonction du type de texte	95
4.11	Visualisation de la distribution de la nature des maillons dans AnnoDis en fonction du genre textuel	95
4.12	Visualisation de la distribution de la nature des maillons dans DEMRC en fonction du genre textuel	96
5.1	Les étapes de l'analyse des séquences appliquée aux chaînes de référence	100
5.2	Une chaîne de référence en tant que séquence d'états	102
5.3	Exemple de graphique de distribution obtenu pour la ressource AnnoDis	106
5.4	Exemple des <i>index plots</i> obtenus après un clustering en 4 classes sur les données AnnoDis	118
5.5	Exemple des séquences représentatives obtenues après un clustering en 4 classes sur les données DEMRC	119

6.1	Exemple de deux chaînes AnnoDis encodées en tant que séquences d'états caractérisés par la nature du maillon	130
6.2	Distribution des chaînes AnnoDis et DEMRC en fonction de leur longueur en nombre de maillons	131
6.3	Exemple de clustering avec une combinaison de paramètres insatisfaisante	134
6.4	Sauts d'inertie résultant du clustering sur AnnoDis et DemRC	136
6.5	Visualisation graphique des logiques de regroupement pour le clustering dans DEMRC	137
6.6	Visualisation graphique des logiques de regroupement pour le clustering dans AnnoDis	139
7.1	Distribution de la nature des maillons dans AnnoDis après la coupure des chaînes à sept maillons	148
7.2	Visualisation des résultats de la classification des séquences AnnoDis en 3 classes par des <i>index plots</i>	149
7.3	Entropie transversale des séquences AnnoDis selon les trois clusters obtenus	150
7.4	Clusters résultant de la division du C1-3c divisé en C1a-5c et C1b-5c dans AnnoDis	152
7.5	Clusters résultant de la division du C3-3c divisé en C3a-5c et C3b-5c	154
7.6	Attraction et répulsion entre les 3 clusters AnnoDis et la nature du référent	156
7.7	Attraction et répulsion entre les 3 clusters AnnoDis et le genre textuel	157
7.8	Répartition des sept patrons AnnoDis en fonction du caractère <i>humain</i> ou <i>non humain</i> du référent de la chaîne.	160
8.1	Couverture des patrons : comparaison entre AnnoDis, $DEMRC_{\{Nar-\}}$ et $DEMRC_{\{Nar+\}}$	168
8.2	Couverture des patrons : comparaison entre les chaînes à référent humain et les chaînes à référent non humain dans AnnoDis et dans DEMRC	169
8.3	Distribution de la nature des maillons dans DEMRC après la coupure des chaînes à sept maillons	171
8.4	Visualisation par des <i>index plots</i> des résultats de la classification des séquences DEMRC en quatre clusters	174
8.5	Entropie transversale des séquences selon les quatre clusters obtenus	175
8.6	Séquences représentatives du C1-4c	176
8.7	Clusters résultant de la division du C2-4c avec leur entropie	179
8.8	Clusters et séquences représentatives résultant de la division du C2-4c	180
8.9	Clusters résultant de la division du C3-4c avec leur entropie	183
8.10	Séquences représentatives des clusters résultant de la division du C3-4c	183
8.11	Subdivisions du C4-4c selon les différents niveaux de classification. Les lignes pointillées indiquent que le cluster ne subit pas de subdivisions ultérieures	188
8.12	Clusters résultant de la division du C4-4c avec leur entropie	189
8.13	Séquences représentatives des clusters résultant de la division du C4-4c	189
8.14	Clusters résultant de la division du C4-4c avec leur entropie	192

8.15 Séquences représentatives des clusters résultant de la division du C4-4c	193
8.16 Clusters et entropie des clusters résultants de la subdivision du C4-4c	195
8.17 Clusters et séquences représentatives résultant de la division du C4-4c	195
8.18 Corrplot : résidus de Pearson standardisés	198
8.19 Corrplot 11c et type de référent : humain vs non humain	199
8.20 Corrplot 11c et type de texte : $\text{DemRC}_{\{\text{Nar-}\}}$ vs $\text{DemRC}_{\{\text{Nar+}\}}$	200
8.21 Corrplot 4c et genres textuels	201
8.22 Corrplot 11c et genres textuels	203
8.23 Distribution des patrons selon le type de texte	210
8.24 Distribution des patrons selon le degré de concretude des référents	211
8.25 Distribution des patrons selon l'animacité des référents	211

Liste des tableaux

1	Liste des abréviations des identifiants accompagnant les exemples . . .	xx
2.1	Nature des maillons : neuf catégories retenues	33
3.1	Composition de la ressource AnnoDis annotée en chaînes topicales . .	62
3.2	Composition du corpus DEMOCRAT	64
3.3	Constitution des corpus AnnoDis et DEMOCRAT	64
4.1	Distribution de la nature des maillons dans AnnoDis après harmonisa- tion des annotations	77
4.2	Composition de la ressource DEMOCRAT_19 ^e -21 ^e	82
4.3	Statistiques descriptives des deux ressources exploitées	91
4.4	Distribution de la nature des maillons dans AnnoDis et DEMRC . . .	93
5.1	Organisation des informations pour les données utilisées pour l'ana- lyse des séquences	103
5.2	Méthodes de calcul des coûts dans l'Optimal Matching (TraMineR) .	114
6.1	Distribution de la nature du référent dans AnnoDis	127
6.2	Distribution de la nature des référents dans DEMRC	129
6.3	AnnoDis	131
6.4	DEMRC	131
6.5	Distribution des séquences selon leur longueur après cupure à sept états dans les deux ressources	131
6.6	Distribution de la nature des maillons dans AnnoDis avant et après la coupure des chaînes	132
6.7	Distribution de la nature des maillons dans DEMRC avant et après la coupure des chaînes	132
6.8	Paramètres utilisés pour l'analyse des séquences dans les ressources AnnoDis et DEMRC	141
7.1	Les patrons les plus fréquents dans la ressource AnnoDis. Les pourcen- tages sont calculés par rapport au nombre total des chaînes à référent non humain ($AD_{Hum\{-}\}$), à référent humain ($AD_{Hum\{+}\}$) et au nombre total de chaînes dans la ressource	158
8.1	Couverture des patrons : comparaison entre AnnoDis et DEMRC . . .	167
8.2	Distribution de la nature détaillée du premier maillon dans DEMRC .	171

8.3	Les patrons les plus fréquents dans DEMRC	207
A.1	Balises xml utilisées pour signaler la structure textuelle dans chaque fichier du corpus DEMOCRAT	238
F.1	P-value et degré de significativité de la distribution des patrons Anno- Dis selon la nature du référent	293
F.2	P-value et degré de significativité de la distribution des patrons Anno- Dis et DEMRC _{Nar-}	294
F.3	Résultats du test binomial entre DEMRC _{Nar+} et DEMRC _{Nar-} par patron	295
F.4	Résultats du test binomial entre DEMRC _{Hum+} et DEMRC _{Hum-} par patron	296
G.1	Distribution des patrons DEMRC dans DEMRC _{Nar+} et DEMRC _{Nar-}	298
G.2	Distribution des patrons DEMRC selon la nature du référent	299

Liste des abréviations

CR Chaîne de Référence

DEMRC Democrat RC 19-21

ER Expression Référentielle

LSF Linguistique Systémique Fonctionnelle

SN_{def} Syntagme Nominal **défini**

SN_{dem} Syntagme Nominal **démonstratif**

SN_{ind} Syntagme Nominal **indéfini**

SN_∅ Syntagme Nominal sans déterminant

∅ Sujet zéro

Np Nom propre

Poss Possessif

Pro Pronom

TAL Traitement Automatique du Langage

Guide de lecture des exemples

Tous les exemples présentés dans cette thèse sont accompagnés d'un identifiant et d'un lien permettant de les visualiser sur des plateformes dédiées à l'exploration des corpus. Ces informations figurent à la fin de chaque exemple, sur la droite, sous la forme suivante : [**Genre Textuel**_{texte} ; ]

Par souci de concision, le genre textuel ainsi que le nom du texte d'origine sont abrégés. Vous trouverez ci-dessous la liste des abréviations utilisées tout au long de cette thèse, accompagnées de leur version complète ([tableau 1](#)). Si vous lisez ce document au format PDF, le nom complet de l'identifiant ainsi que le corpus d'origine de l'exemple s'afficheront en passant la souris sur "[]".

À la suite du [tableau 1](#), vous trouverez la liste de tous les exemples avec leurs identifiants d'origine dans les corpus. Ces identifiants permettent d'effectuer des recherches sur les plateformes d'exploration des corpus, facilitant ainsi l'accès aux sources originales. Si vous consultez la version PDF de ce document, vous y serez redirigé.e.s en cliquant sur l'identifiant abrégé. Si vous consultez la thèse en version papier, vous pourrez retrouver l'identifiant d'origine à partir du numéro de l'exemple et l'utiliser pour accéder aux données correspondantes.

Abréviation	Genre Textuel	Texte
AE _{W}	Articles Encyclopédiques	Wikipédia
AP _{ER}	Articles de Presse	Est Republicain
AS _{G} AS _{L}	Articles Scientifiques	Geopolitique Linguistique
B _{ES} B _{MdH} B _{RdC}	Biographie	Élisabeth Seton Madame de Hautefort Rosalie de Constant
N _{MF} N _{MA} N _{S}	Nouvelle	Mademoiselle Fifi Morte Amoureuse Sarrasine
P _{A}	Pamphlet	Aden
R _{BeP} R _{CF} R _{P} R _{VdP} R _{DVM} R _{DaC} R _{DL} R _{JC} R _{N}	Roman	Bouvard et Pécuchet Capitaine Fracasse Pauline Ventre de Paris De la Ville au Moulin Diable au Corps Douce Lumière Jean-Christophe Némoville
TJ _{CCF} TJ _{CPP} TJ _{CIU} TJ _{CMM} TJ _{CTA} TJ _{CTT}	Texte Juridique	Code Civil des Français Code Procédure Pénale Convention Institut Universitaire Convention Milieu Marin Convention Transport Aérien Convention Thon Tropical

TABLEAU 1 – Liste des abréviations des identifiants accompagnant les exemples

Liste des identifiants des exemples

Introduction

- (1) DEMRC : jeanchristophe-1_RC_19-21_s-CHAINE-730 [] p. 2

Chapitre 1

Section 1.1

- (2) DEMRC : elisabethseton_RC_19-21_s-CHAINE-1 [] p. 13

Section 1.2

- (3) RésolCo : EC-CE2-2015-PBOX-D1-R3-V1 [] p. 18

Chapitre 2

Section 2.1.1

- (4) DEMRC : articleswiki_RC_19-21_s-CHAINE-1039 [] p. 32
 (5) DEMRC : sarrasine_RC_19-21_s-CHAINE-248 [] p. 32

Section 2.2.3

- (6) DEMRC : rosaliedeconstant_RC_19-21_s-CHAINE-406 [] p. 39

Chapitre 3

Section 3.3.3

- (13) AnnoDis : wik2_leonardDeVinciCT_coder2_1257949561835 [] ... p. 64

Section 3.4

- (16) DEMRC : aden_RC_19-21_s-CHAINE-39 [] p. 66

Chapitre 4

Section 4.3.2

- (17) DEMRC : jeanchristophe-1_RC_19-21_s-CHAINE-29 [] p. 91

Chapitre 5

Section 5.2.3

- (18) AnnoDis : wik2_titanicCT_coder2_1281096316275 [] p. 108
 (19) DEMRC : ventredeparis_RC_19-21_s-CHAINE-687 [] p. 108
 (20) DEMRC : estrepublikain_RC_19-21_s-CHAINE-1636 [] p. 109

Chapitre 6

Section 6.1.1

- (21) AnnoDis : geop_21CT_coder2_1254904687138 [] p. 125
 (22) AnnoDis : geop_28CT_coder2_1280481038709 [] p. 125
 (23a) AnnoDis : wik2_linuxCT_coder1_1255343034250 [] p. 126
 (23b) AnnoDis : wik2_bicycletteCT_coder2_1285841290521 [] p. 126

Section 6.2.1

- (24) DEMRC : madamedehautefort_RC_19-21_s-CHAINE-802 [] ... p. 133

Chapitre 7

Section 7.2.1

- (25) AnnoDis : wik2_vinDeChampagneCT_coder3_1254135673500 [] p. 151
 (26) AnnoDis : wik2_affaireDreyfussCT_coder2_1254316308199 [] .. p. 151

Section 7.2.2

- (27) AnnoDis : ling_barbazanCT_coder2_1320622605359 [] p. 152

Section 7.2.3

- (28) AnnoDis : ling_dalCT_coder3_1253024655578 [] p. 153
 (29) AnnoDis : wik2_affaireDreyfussCT_coder2_1254320843928 [] p. 153

Section 7.4

- (30) AnnoDis : wik2_affaireDreyfussCT_coder2_1254320843928 [] ..p. 159
- (31) AnnoDis : wik2_hippieCT_coder3_1254154221031 []p. 160
- (32) AnnoDis : wik2_affaireDreyfussCT_coder2_1254320731278 [] ..p. 161
- (33) AnnoDis : ling_barbazanCT_coder2_1320622234494 []p. 161
- (34) AnnoDis : wik2_albertEinsteinCT_coder1_1254213710781 [] ...p. 162

Chapitre 8

Section 8.1

- (35) DEMRC : bouvardetpecuchet_RC_19-21_s-CHAINE-808 []p. 170

Section 8.2

- (36) DEMRC : ventredeparis_RC_19-21_s-CHAINE-934 []p. 172
- (37) DEMRC : bouvardetpecuchet_RC_19-21_s-CHAINE-1664 [] ...p. 172
- (38) DEMRC : capitainefracasse_RC_19-21_s-CHAINE-656 []p. 172
- (39a) DEMRC : diableaucorps_RC_19-21_s-CHAINE-243 []p. 173
- (39b) DEMRC : estrepublicain-2_RC_19-21_s-CHAINE-1766 []p. 173

Section 8.3.1

- (40a) DEMRC : estrepublicain-2_RC_19-21_s-CHAINE-114 []p. 176
- (40b) DEMRC : rosaliedeconstant_RC_19-21_s-CHAINE-908 []p. 176
- (40c) DEMRC : diableaucorps_RC_19-21_s-CHAINE-1687 []p. 177
- (40d) DEMRC : rosaliedeconstant_RC_19-21_s-CHAINE-35 []p. 177
- (40e) DEMRC : estrepublicain_RC_19-21_s-CHAINE-135 []p. 177

Section 8.3.2

- (41a) DEMRC : codecivildesfrancais-1_RC_19-21_s-CHAINE-336 [] .p. 180
- (42a) DEMRC : jeanchristophe-1_RC_19-21_s-CHAINE-880 []p. 180
- (55e) DEMRC : convtransportaerien_RC_19-21_s-CHAINE-459 []p. 288

Section 8.3.3

- (43a) DEMRC : conventionmilieumarin_RC_19-21_s-CHAINE-187 [] .p. 182
- (43b) DEMRC : sarrasine_RC_19-21_s-CHAINE-1230 []p. 184
- (43c) DEMRC : elisabethseton_RC_19-21_s-CHAINE-564 []p. 184
- (44a) DEMRC : estrepublicain-2_RC_19-21_s-CHAINE-1766 []p. 184
- (44b) DEMRC : diableaucorps_RC_19-21_s-CHAINE-1330 []p. 185
- (44c) DEMRC : codecivildesfrancais-2_RC_19-21_s-CHAINE-1392 [] p. 185

- (44d) DEMRC : estrepublicain_RC_19-21_s-CHAINE-881 [] p. 185
 (44e) DEMRC : conventionthontropical_RC_19-21_s-CHAINE-455 [] .p. 185

Section 8.3.4

- (45a) DEMRC : codeprocedurepenale_RC_19-21_s-CHAINE-1657 [] . p. 190
 (45b) DEMRC : mademoisellefifi-3_RC_19-21_s-CHAINE-93 [] p. 190
 (45c) DEMRC : codecivildesfrancais-2_RC_19-21_s-CHAINE-1950 [] p. 190
 (45d) DEMRC : codeprocedurepenale_RC_19-21_s-CHAINE-1758 [] . p. 190
 (45e) DEMRC : jeanchristophe-1_RC_19-21_s-CHAINE-963 []p. 190
 (46a) DEMRC : articleswiki_RC_19-21_s-CHAINE-979 []p. 191
 (46a) DEMRC : articleswiki_RC_19-21_s-CHAINE-979 [] p. 191
 (47a) DEMRC : elavilleaumoulin_RC_19-21_s-CHAINE-859 [] p. 191
 (47b) DEMRC : nemoville_RC_19-21_s-CHAINE-955 []p. 191

 (48a) DEMRC : morteamoureuse_RC_19-21_s-CHAINE-1119 [] p. 193
 (48b) DEMRC : conventionthontropical_RC_19-21_s-CHAINE-1432 [] p. 193
 (49a) DEMRC : aden_RC_19-21_s-CHAINE-1624 [] p. 194
 (49b) DEMRC : nemoville_RC_19-21_s-CHAINE-83 []p. 194
 (49c) DEMRC : rosaliedeconstant_RC_19-21_s-CHAINE-1333 [] p. 194
 (49d) DEMRC : delavilleaumoulin_RC_19-21_s-CHAINE-350 [] p. 194
 (49e) DEMRC : estrepublicain_RC_19-21_s-CHAINE-619 [] p. 194

 (50b) DEMRC : jeanchristophe-2_RC_19-21_s-CHAINE-793 []p. 196

Section 8.6

- (52) DEMRC : estrepublicain-2_RC_19-21_s-CHAINE-84 [] p. 208
 (53) DEMRC : codeprocedurepenale_RC_19-21_s-CHAINE-1560 [] . p. 208
 (58a) DEMRC : conventionthontropical_RC_19-21_s-CHAINE-1549 [] p. 291
 (57a) DEMRC : mademoisellefifi-1_RC_19-21_s-CHAINE-860 []p. 291

Annexe E

Annexe E.1

- (54a) DEMRC : conventionthontropical_RC_19-21_s-CHAINE-802 [] . p. 285
 (54b) DEMRC : codecivildesfrancais-2_RC_19-21_s-CHAINE-623 [] . p. 286
 (54c) DEMRC : codecivildesfrancais-2_RC_19-21_s-CHAINE-927 [] . p. 286
 (54d) DEMRC : codecivildesfrancais-1_RC_19-21_s-CHAINE-336 [] . p. 286
 (54e) DEMRC : codeprocedurepenale_RC_19-21_s-CHAINE-1290 [] . p. 287

Annexe E.2

- (55a) DEMRC : codeprocedurepenale_RC_19-21_s-CHAINE-719 [] .. p. 287
- (55b) DEMRC : estrepublicain_RC_19-21_s-CHAINE-600 [] p. 287
- (55c) DEMRC : conventionmilieumarine_RC_19-21_s-CHAINE-1199 [] p. 287
- (55d) DEMRC : jeanchristophe-1_RC_19-21_s-CHAINE-880 [] p. 288
- (55e) DEMRC : convtransportaerien_RC_19-21_s-CHAINE-459 [] p. 288

Annexe E.3

- (56a) DEMRC : conventionmilieumarine_RC_19-21_s-CHAINE-187 [] . p. 288
- (56b) DEMRC : sarrasine_RC_19-21_s-CHAINE-1230 [] p. 289
- (56c) DEMRC : elisabethseton_RC_19-21_s-CHAINE-564 [] p. 289
- (56d) DEMRC : conventionmilieumarine_RC_19-21_s-CHAINE-313 [] . p. 289
- (56e) DEMRC : conventionthontropical_RC_19-21_s-CHAINE-19 [] .. p. 290

Annexe E.4

- (57a) DEMRC : mademoisellefifi-1_RC_19-21_s-CHAINE-860 [] p. 291
- (57b) DEMRC : codecivildesfrancais-1_RC_19-21_s-CHAINE-1028 [] p. 291
- (57c) DEMRC : bouvardetpecuchet_RC_19-21_s-CHAINE-120 [] p. 291
- (57d) DEMRC : codecivildesfrancais-1_RC_19-21_s-CHAINE-1402 [] p. 291
- (57e) DEMRC : mademoisellefifi-3_RC_19-21_s-CHAINE-2 [] p. 291

Annexe E.5

- (58a) DEMRC : conventionthontropical_RC_19-21_s-CHAINE-1549 [] p. 291
- (58b) DEMRC : codecivildesfrancais-1_RC_19-21_s-CHAINE-1367 [] p. 292
- (58c) DEMRC : articleswiki_RC_19-21_s-CHAINE-423 [] p. 292
- (58d) DEMRC : convtransportaerien_RC_19-21_s-CHAINE-142 [] p. 292
- (58e) DEMRC : codeprocedurepenale_RC_19-21_s-CHAINE-647 [] .. p. 292

Introduction

Avez-vous déjà lu ou entendu une phrase dont le sens vous échappait, au point de devoir demander une précision pour comprendre à quoi ou à qui un mot faisait référence ? C'est ce qui arrive aux protagonistes de l'image ci-dessous.



FIGURE 1 – The stars or the humans ?

Dans la [figure 1](#), un petit éléphant et un petit renard contemplant les étoiles. Au cours de leur échange, le renard demande une précision sur ce à quoi l'éléphant fait référence ¹. Cet échange illustre un phénomène fondamental du langage : la référence, c'est-à-dire la manière dont nous évoquons des objets, des concepts ou toute autre entité dont nous voulons parler. Mais comment savons-nous à quoi renvoient les mots que nous utilisons ? Comment relier-nous un pronom, un nom ou une expression à l'entité dont il est réellement question dans un texte ou une conversation ? Quelles stratégies permettent d'assurer une bonne interprétation et une continuité référentielle ?

Ces questions sont au cœur de cette thèse. En linguistique, le phénomène qui garantit cette continuité dans le discours est connu sous le nom de chaînes de référence (CR). Nous nous intéressons ici à leur fonctionnement textuel par le biais de l'exploration de ressources diversifiées. En analysant l'enchaînement des maillons qui les composent,

1. L'ambiguïté de la phrase en anglais provient du fait que le pronom 'they' est neutre et peut désigner aussi bien des humains que des objets inanimés. En français, une formulation équivalente éviterait sans doute cette confusion, car le genre grammatical imposerait un accord distinct : « ils » pour les humains, « elles » pour les étoiles.

nous cherchons à mieux comprendre les stratégies d'introduction et de maintien du référent, ainsi que leur impact sur l'organisation et l'interprétation du discours. Ce travail s'inscrit dans un contexte particulièrement dynamique, où les recherches sur les CR se multiplient dans divers domaines. En linguistique descriptive, elles permettent d'affiner notre compréhension des mécanismes référentiels et de leurs implications discursives (Cornish, 2021 ; Gardelle & Vincent-Durroux, 2022 ; Landragin, Glikman, Schnedecker, & Todirascu, 2024 ; Manuélian & Schnedecker, 2021, 2022). En linguistique contrastive, elles offrent des perspectives comparatives sur la manière dont différentes langues gèrent la continuité référentielle (Guo, 2021 ; Hou & Landragin, 2020 ; Kunz, 2010 ; Kunz, Lapshinova-Koltunski, & Martínez, 2016). Enfin, le traitement automatique du langage (TAL) s'intéresse à ces phénomènes pour développer et améliorer les systèmes de résolution de la référence (Anikina et al., 2025, 2021 ; Bevacqua, Loáiciga, Rohde, & Hardmeier, 2021 ; Chai & Strube, 2022 ; Davis & van Schijndel, 2020 ; Grobol, 2020 ; E. J. Gundel & Abbott, 2019 ; Held, Iter, & Jurafsky, 2021 ; Li, Kobayashi, & Ng, 2021 ; Quignard, Le Mené, & Landragin, 2021 ; Toshniwal, Xia, Wiseman, Livescu, & Gimpel, 2021).

Avant de définir plus précisément ce qu'est une chaîne de référence, il convient de définir ce qu'est la référence et quel est son rôle dans le discours. Dans une perspective linguistique et fonctionnelle, « [l]e processus de référenciation a pour finalité d'attirer l'attention d'un ou de plusieurs interlocuteurs sur une ou plusieurs entités du monde externe » (Charolles, 2002). Pour ce faire, tout locuteur fait appel à des éléments linguistiques (noms propres, syntagmes nominaux, pronoms, etc.), que l'on appelle expressions référentielles (ER). Lorsque plusieurs ER renvoient à une même entité — le référent (être humain, non humain, entité concrète, entité abstraite, etc.) — elles créent ce que l'on définit comme une chaîne de référence, c'est-à-dire « la suite des expressions d'une production écrite ou orale, qui renvoient à une même entité » (Schnedecker, 2021, 2). Pour illustrer ce phénomène, considérons l'exemple suivant (1). Ce passage met en évidence une chaîne de référence où plusieurs ER (en gras) renvoient à une même entité (*le grillon*).

- (1) Il cherchait **un grillon** pour **en** faire un cheval : il **lui** mettait doucement sa baguette sur le dos, et disait une formule. **L'insecte** se sauvait : il **lui** barrait le chemin. Après quelques instants, il était couché à plat ventre, près de **lui**, et il **le** regardait. Il avait oublié son rôle de magicien, et s'amusait à retourner sur le dos **la pauvre bête**, en riant de **ses** contorsions. [R_{JC1} ; 

La succession des expressions référentielles (ER) — appelées aussi maillons pour rester dans la métaphore de la chaîne — assure la continuité référentielle, créant ainsi des liens de cohésion (Halliday & Hasan, 1976). Considérées dans le cadre de la linguistique fonctionnelle systémique (Halliday, 1985), les chaînes de référence (CR) jouent un rôle crucial dans l'organisation et l'interprétation du discours. Pour cette raison, la référence et les CR ont fait l'objet d'une multitude d'études dans des domaines variés : linguistique cognitive, linguistique descriptive, psycholinguistique, didactique, traitement automatique des langues, entre autres.

Nous nous intéressons à la capacité des CR à créer des liens de cohésion. Dans cette optique, nous considérons les CR comme un tout, c'est-à-dire comme des structures signalées par des séquences d'expressions coréférentielles, également appelées mentions, plutôt que comme un ensemble d'expressions équivalentes.

Nous partons de plusieurs constats. Le premier porte sur l'absence d'études systématiques et à large échelle CR en français écrit. En effet, pendant longtemps, il a manqué pour le français un corpus de grande taille pour permettre des études systématiques ou des études TAL. La mise à disposition du corpus Democrat ([Landragin, 2016](#)) a ouvert de nombreuses perspectives, tant pour la linguistique descriptive que pour le TAL. Par ailleurs, étant donné la complexité du phénomène et la variété des aspects qui peuvent être pris en compte pour la description des CR, les études existantes se focalisent généralement sur un nombre restreint de CR et d'aspects pris en considération. Les nombreuses études existantes ont permis de sélectionner des critères de description et ont montré que la composition des CR varie en fonction de plusieurs facteurs : le canal ([Amoia, Kunz, & Lapshinova-Koltunski, 2012](#)), la langue ([Baumer, Dias, & Schnedecker, 2023](#)), l'époque ([Landragin et al., 2024](#)), le type textuel et le genre textuel ([Kunz & Lapshinova-Koltunski, 2015](#) ; [Longo, 2013](#) ; [Schnedecker & Longo, 2012](#)). À notre connaissance, les études sur le genre textuel portent sur un seul genre ou sur un échantillon restreint de données.

Un autre constat concerne la difficulté de pallier ce manque en combinant plusieurs corpus annotés. En effet, les différents corpus existants, aussi bien en français que dans d'autres langues, adoptent des modèles d'annotations différents. Selon le cadre théorique que l'on adopte et l'objectif de recherche, les choix d'annotations sont différents. De plus, tous les corpus n'annotent pas forcément le même type de phénomène, ou de référent. Alors que cette hétérogénéité est généralement perçue comme une limite à la comparabilité des résultats, nous nous demandons, au contraire, s'il serait possible d'en tirer parti pour appréhender différents aspects des CR, précisément parce que les phénomènes annotés varient. Pour essayer de répondre à cette question, nous proposons d'exploiter des corpus annotés diversifiés.

Enfin, un dernier constat concerne la description même des CR. Les études existantes décrivent les chaînes de référence selon des caractéristiques globales : longueur en nombre de maillons, distance moyenne entre les maillons, proportion de la catégorie grammaticale des maillons, etc. Ces études ont permis de mettre au jour des critères de description pertinents pour l'étude des CR et ont mis en évidence les facteurs qui peuvent influencer les caractéristiques des CR. Cependant, dès que l'on s'intéresse aux stratégies d'introduction et de maintien du référent, ce type d'approche n'est pas suffisant.

Dans ce contexte, notre objectif est de prendre en considération les CR dans leur linéarité, en nous intéressant aux enchaînements des maillons. Afin de fournir une description systématique des enchaînements des maillons, tout en prenant en considération un grand nombre de données, nous proposons une méthode outillée de corpus adoptant une approche linéaire et non plus simplement globale des chaînes de référence. Cette approche est adaptée à l'exploitation de ressources hétérogènes (en termes de type de textes, de modèles linguistiques, de langues) et permet une analyse contrastive

des résultats obtenus. En s'appuyant sur deux corpus annotés diversifiés, notre étude examine également l'influence du modèle d'annotation sur la description des **chaînes de référence (CR)**. Cela présente un intérêt d'un point de vue descriptif, mais pourrait également fournir des éléments pertinents pour le TAL, puisque les systèmes de résolution apprennent à partir de données annotées, rendant leurs performances étroitement dépendantes des types d'annotation fournis. Cette étude est donc la première, à notre connaissance, à prendre en considération un grand nombre de **CR**, sans se limiter à un échantillon, pour le français écrit. C'est également une des premières études menées sur le corpus Democrat ([Landragin, 2016](#)) qui propose une description de la totalité des chaînes annotées dans le corpus (pour les textes du 19^{ème} au 21^{ème})². Nous utilisons des méthodes du TAL pour nous donner les moyens d'analyser de manière systématique une grande quantité de données tout en considérant les chaînes et les maillons dans leur linéarité.

La thèse s'articule en trois grandes parties, suivies d'une discussion et des perspectives. La première partie (**partie I**) propose un état de l'art sur les chaînes de référence ainsi que sur les éléments nécessaires à leur étude. Dans le **chapitre 1**, nous montrons que les chaînes de référence constituent un phénomène central du discours. Après avoir introduit les notions fondamentales d'anaphore, de coréférence et de cohésion, nous dressons un panorama des différentes disciplines qui s'intéressent à la référence, offrant ainsi un héritage diversifié sur lequel s'appuyer pour décrire et comprendre les chaînes de référence. Le **chapitre 2** recense les différents critères de description utilisés dans les travaux sur les chaînes de référence et les facteurs qui ont un impact sur leurs caractéristiques. Dans le **chapitre 3**, après avoir expliqué en quoi consiste la tâche d'annotation de la référence et les choix qu'elle implique, nous présentons les corpus annotés en référence. Nous mettons en avant les différences qui existent entre les modèles d'annotation. Nous réfléchissons ensuite à l'impact du modèle d'annotation adopté, en soulignant comment les différences entre les corpus peuvent influencer la description des chaînes de référence.

Dans la **partie II**, nous présentons la méthodologie proposée pour caractériser les chaînes de référence. Dans le **chapitre 4**, nous décrivons les corpus utilisés dans le cadre de cette étude : AnnoDis ([Péry-Woodley, Afantenos, Ho-Dac, & Asher, 2011](#)) et Democrat ([Landragin, 2016](#)). Pour chaque corpus, dans un premier temps, nous explicitons les traitements effectués afin de les préparer pour l'analyse. Dans un deuxième temps, nous présentons l'anatomie des chaînes de référence, ce qui nous offre une vision précise des données que nous allons manipuler par la suite. Dans le **chapitre 5**, nous présentons les fondements de l'approche linéaire en détaillant l'analyse des séquences ([Abbott, 1983](#)), qui constitue le cadre de référence pour notre méthodologie. Dans le **chapitre 6**, nous expliquons l'application de notre méthode aux chaînes de référence, en justifiant nos choix et en fixant le paramétrage utilisé dans le cadre de ce travail.

Dans la **partie III**, nous rendons compte des résultats obtenus afin de nous diriger vers une typologie des chaînes de référence. Nous explorons les enchaînements des

2. [Landragin et al. \(2024\)](#) ont récemment proposé une description de l'ensemble du corpus en prenant en compte certains critères, tels la longueur des chaînes, la distance entre les maillons, la nature des maillons et la stabilité référentielle.

maillons dans les chaînes du corpus AnnoDis, dans le [chapitre 7](#), et Democrat, dans le [chapitre 8](#).

Première partie

Les chaînes de référence (CR) : un phénomène complexe

Cette première partie de la thèse nous permet de définir les chaînes de référence (CR) et de poser la notion de cohésion (**chapitre 1**). Dans le **chapitre 2**, nous présentons les différents critères de description et les facteurs de variation que les études sur les CR ont mis au jour. En adoptant une approche bottom-up, nous nous appuyons sur l'analyse de corpus. Or, les corpus annotés en référence présentent une certaine hétérogénéité, ce qui est souvent perçu comme une limite à la comparabilité des résultats obtenus. Pourtant, plutôt que de la voir comme un obstacle, nous nous interrogeons sur la manière dont cette hétérogénéité pourrait être exploitée pour explorer différents aspects des chaînes de référence et en dégager de nouvelles perspectives. Il est crucial, pour pouvoir croiser des ressources hétérogènes, de comprendre en quoi elles diffèrent. Cette étape permet de proposer des comparaisons entre les résultats obtenus et d'analyser les CR en tenant compte des nuances nécessaires. Dans le **chapitre 3**, nous présentons donc les différentes ressources qui existent et nous décrivons les principales différences entre les modèles d'annotation. Enfin, nous soulignons l'impact que les modèles d'annotation peuvent avoir pour la description des CR en nous focalisant sur les critères de description les plus utilisés dans la littérature.

Chapitre 1

Les chaînes de référence : un phénomène central du discours

Contents

1.1	Une suite d'expressions référentielles	12
1.1.1	Position des maillons	14
1.1.2	Relation entre les maillons	16
1.2	Un indice de cohésion	17
1.3	Un héritage diversifié	20
1.3.1	Héritage cognitif	20
1.3.2	Héritage psycholinguistique	23
1.3.3	Héritage linguistique	24
1.4	Bilan	26

Dans toute situation de communication, orale ou écrite, les locuteurs mobilisent des expressions référentielles (noms propres, pronoms, syntagmes nominaux, etc.) afin d'évoquer des référents (personnes, animaux, idées, etc.). Cet acte, connu sous le nom de *référence*, implique un double processus : celui du choix et de la production d'une expression référentielle par un locuteur et celui de l'interprétation par un ou plusieurs interlocuteurs. Pour que ce processus aboutisse, les expressions référentielles jouent un rôle crucial : elles ne se limitent pas à désigner des référents, mais fournissent des indices sur leur statut et leur place dans le discours. Dans une perspective fonctionnelle et cognitive (Charolles, 2002), il est admis que les expressions référentielles ont une valeur instructionnelle : en fonction de leurs caractéristiques (catégorie grammaticale, fonction syntaxique, contenu lexical), elles apportent des instructions distinctes. Certaines, plus riches et informatives, sont préférées pour introduire un nouveau référent dans le discours. D'autres, plus réduites, permettent de maintenir ou de réactiver un référent déjà mentionné. C'est cette dimension instructionnelle des expressions référentielles, au sens de Charolles (2002), qui guide notre analyse.

Lorsque plusieurs expressions signalent un même référent, elles créent ce que l'on appelle, depuis Chastain (1975), une *chaîne de référence*. Composées d'expressions

référentielles, appelées *maillons*, les chaînes permettent d'assurer la continuité référentielle dans le discours. À la suite de [Schneidecker \(2019\)](#), nous retenons qu'une chaîne de référence implique au moins trois maillons, faute de quoi il est plus pertinent de parler d'anaphore ou de coréférence. Considérée dans le cadre de la linguistique fonctionnelle systémique ([Halliday, 1985](#)), il est admis que la *référence*, avec d'autres processus, contribue à la création des liens de cohésion ([Halliday & Hasan, 1976](#)). Suivant cette perspective, nous considérons que les chaînes de référence contribuent à la création de liens de cohésion. Parce qu'elles sont constituées d'expressions référentielles liées entre elles, les chaînes participent à la structuration du discours en assurant la liaison entre ses segments. Elles permettent ainsi au lecteur ou à l'interlocuteur de suivre les évolutions des référents dans l'espace textuel, jouant un rôle organisationnel fondamental. C'est dans cette optique que nous envisageons notre objet d'étude : les enchaînements de maillons. Ces enchaînements sont au cœur des mécanismes de référence et de cohésion, et leur observation fine permet d'accéder aux stratégies discursives mises en place pour introduire, maintenir ou réactiver les référents. Notre démarche vise ainsi à dégager des patrons d'enchaînements récurrents et à interroger leur contribution à la structuration du discours. À terme, cette réflexion s'inscrit dans une perspective plus large de typologie des chaînes de référence, en fonction de leurs caractéristiques et de leur rôle dans l'organisation textuelle.

Ce chapitre introductif a pour objectif de poser les fondements théoriques nécessaires à cette analyse. Nous commençons par caractériser les expressions référentielles et leurs fonctions (cf. 1.1). Nous présentons ensuite les relations qui peuvent s'établir entre elles, notamment à travers les notions d'anaphore et de coréférence, et la manière dont elles s'articulent dans les chaînes de référence. Nous abordons ensuite la notion de cohésion textuelle (cf. 1.2), avant de montrer que la référence et les chaînes de référence ont été abordées dans des cadres disciplinaires variés, chacun apportant un éclairage complémentaire sur les procédés mis en œuvre (cf. 1.3). En gardant pour fil directeur notre objectif d'analyse des stratégies d'introduction et de maintien des référents, nous mettons en évidence les concepts déterminants pour la suite de notre travail.

1.1 Une suite d'expressions référentielles

Comme anticipé dans l'introduction, lorsque nous communiquons, que ce soit à l'oral ou à l'écrit, nous évoquons plusieurs entités : personnes, animaux, objets, événements, idées, concepts, etc. En linguistique, on appelle ces entités *référents*. L'acte de mentionner ces référents est connu sous le nom de *référence*. La référence constitue un phénomène essentiel et constamment présent dans le discours, à savoir la production d'énoncés en contexte ([Péry-Woodley, 2000](#)). Comme la citation de [Kibrik \(2011\)](#) l'atteste, la référence est un aspect central du fonctionnement des langues naturelles :

« As people talk, they constantly refer, that is, they mention persons or objects. The process of reference is among the most central and ubiquitous phenomena in the functioning of natural languages » ([Kibrik, 2011](#), xxvi).

Sans *référer*, le locuteur — celui qui produit un message — ne serait pas en mesure de poser "ce dont il veut parler", et son interlocuteur — celui qui reçoit ce message —

ne pourrait pas comprendre "ce dont il est question" dans le discours. En effet, « [l]e processus de référenciation a pour finalité d'attirer l'attention d'un ou de plusieurs interlocuteurs sur une ou plusieurs entités du monde externe » (Charolles, 2002, 5). Autrement dit, le processus de référence permet au locuteur de signifier les entités dont il va parler à son interlocuteur. À son tour, ce dernier focalisera son attention sur ces entités et pourra correctement interpréter le message qui lui est destiné.

Afin de désigner les référents, quelle que soit leur nature (être humain, non humain, entité concrète, entité abstraite, etc.), tout locuteur fait appel à des éléments linguistiques (noms propres, syntagmes nominaux, pronoms, etc.), que l'on appelle *expressions référentielles* (ER). Dans l'exemple (2), extrait de la biographie *Élisabeth Seton*, l'entité principale sur laquelle le locuteur attire notre attention est Élisabeth Seton¹, un référent humain. Les ER qui portent sur ce référent sont signalées en gras dans l'exemple.

- (2) **Élisabeth Seton**. Il y a déjà quatre-vingts ans que **Madame Seton** s'est éteinte à Emmetsburg, mais, loin de plonger dans l'ombre, **sa** mémoire grandit. D'après quelques journaux, les évêques des États-Unis songeraient à demander la canonisation de **cette illustre convertie**. On a même annoncé que le président de l'Université de Niagara était chargé de recueillir les preuves de l'héroïcité de **ses** vertus. Quoi qu'il en soit, **Élisabeth Seton** est **l'une des gloires de l'Amérique** et **son** histoire offre un grand intérêt. **Cette femme comblée de tous les dons** a éprouvé toutes les vicissitudes humaines ; **elle** a connu les ardent tendresses, les joies qui transportent et aussi toutes les angoisses, tous les déchirements de la douleur. Encore protestante, **elle** s'est élevée à une admirable vertu : « **Je** ne regarde ni en arrière, ni en avant, **je** regarde en haut, » disait-**elle** dans l'écroulement de **son** bonheur, l'un des plus complets qu'une créature mortelle ait jamais possédés [...] [B_{ES} ; ↻]

Le référent est mentionné pour la première fois par le biais d'un nom propre, *Élisabeth Seton*. Il est ensuite repris dans le texte par le biais d'autres expressions référentielles. Bien que ces ER soient diversifiées (noms propres, pronoms, syntagmes nominaux définis ou démonstratifs), elles permettent d'attirer l'attention de l'interlocuteur sur un référent spécifique, qu'il pourra suivre tout au long du discours (Charolles, 2002 ; Schnedecker, 2021). Dans ce sens, les ER peuvent être considérées comme des marques de *continuité référentielle* : elles guident les interlocuteurs dans l'interprétation du discours, signalant que le locuteur continue de parler de la même entité. En lisant l'extrait (2), le lecteur peut percevoir une continuité référentielle réalisée par la succession des différentes expressions référentielles qui renvoient à un même référent (*Elisabeth Seton*). Les réalisations linguistiques de cette continuité référentielle donnent lieu à ce que l'on appelle une *chaîne de référence* (CR), c'est-à-dire « la suite des expressions d'un texte entre lesquelles l'interprétation construit une relation d'identité référentielle » (Corblin, 1995, 123).

Les ER à disposition des locuteurs sont de forme très variée, comme nous pouvons le constater dans l'exemple (2). Un même référent peut être mentionné et repris

1. De nombreuses entités sont mentionnées dans cet extrait, nous avons souligné uniquement les expressions qui renvoient au référent principal du discours.

avec des expressions différentes. Ces expressions ne sont pas choisies au hasard : dans une perspective qualifiée par Charolles (2002) de « fonctionnalisme teinté de cognitivisme », il est admis que les expressions référentielles, selon leurs caractéristiques, codent des instructions spécifiques pour l'interprétation du discours (Charolles, 2002). Dans le discours écrit, où la communication est distanciée, les interlocuteurs ne partagent pas le contexte et il n'y a pas d'interaction possible permettant une négociation du sens. Le lecteur doit alors identifier les *indices* qui lui permettent d'interpréter les intentions du locuteur. Ainsi, les ER sont considérées comme des *marques* ou des *indices* que l'on trouve à la surface des textes pour *signaler* à l'interlocuteur que l'on continue à parler de la même entité. Ces *signalisations* guident alors l'interprétation du discours (Ho-Dac, 2007).

Dans cette optique, les ER ont donc une valeur instructionnelle : selon qu'il souhaite introduire un nouveau référent ou en maintenir un déjà mentionné, le locuteur mobilise des formes linguistiques différentes, plus ou moins aptes à remplir ces fonctions. En effet, les ER se réalisent sous des formes grammaticales variées (syntagmes nominaux définis ou indéfinis, pronoms, syntagmes démonstratifs, etc.), peuvent occuper diverses fonctions syntaxiques, et selon leur contenu lexical, être plus ou moins informatives. D'un point de vue cognitif, selon leur forme, les expressions référentielles signalent le statut cognitif du référent et son accessibilité (cf. 1.3.1). Certaines expressions sont autonomes d'un point de vue référentiel, dans le sens où elles permettent d'identifier directement le référent, sans avoir recours au contexte². Les expressions autonomes du point de vue référentiel sont les noms propres complets, les descriptions définies complètes qui donnent des informations sur les référents, les noms qui renvoient à des entités uniques (*le soleil, la terre*) (Charolles, 2002), par exemple. Ces formes — souvent plus riches en informations lexicales — sont ainsi privilégiées pour l'introduction de nouveaux référents, pour le réactiver ou pour éviter tout risque d'ambiguïté. D'autres expressions, comme les pronoms, plus réduites ou moins informatives, nécessitent la prise en compte du contexte pour pouvoir être interprétées correctement. Elles sont alors préférées pour assurer le maintien d'un référent déjà établi dans la mémoire discursive de l'interlocuteur.

La forme des ER est donc étroitement liée à leur rôle dans la dynamique référentielle du discours. Leur complexité syntaxique, leur degré d'informativité et leur autonomie référentielle influencent directement leur aptitude à introduire un nouveau référent, à maintenir un référent déjà mentionné ou à le réactiver dans le discours. Ainsi, selon le rôle que ces ER jouent dans le discours, elles sont susceptibles d'occuper des positions différentes dans les chaînes de référence.

1.1.1 Position des maillons

Cornish (1998) et Charolles (1988) distinguent trois positions possibles des maillons dans les CR en fonction de leur forme. La position de *premier maillon*, de *deuxième maillon* et la position de *troisième maillon* ou *maillon-n*.

2. Ici *contexte* est utilisé pour signaler le co-texte, mais également le contexte d'un point de vue plus pragmatique.

Premier maillon : Le premier maillon de la CR a pour fonction d'introduire un référent pour la première fois dans le discours ou de réactiver un référent qui avait été mentionné, mais qui a ensuite cédé la place à d'autres référents. Dans cette position, on trouve principalement des expressions référentielles autonomes, telles que les noms propres, ou des syntagmes définis. D'après [Charolles \(1988\)](#) les SN définis complets sont utilisés principalement en première position car ils permettent d'introduire le référent pour la première fois dans un discours. Dans cette position, on peut aussi trouver des SN indéfinis, dont la fonction est d'introduire des nouveaux référents. Pour ce qui concerne les SN indéfinis, [Schnedecker et Landragin \(2014\)](#) soulignent que ceux-ci sont moins aptes à déclencher une CR. Les auteurs soutiennent en effet que ce type de syntagme introduit un référent qui n'est pas très important, qui ne constitue pas, de leur point de vue, un "focus d'attention". Cependant, la mise en chaîne n'est pas impossible. En raison de leur nature, les syntagmes indéfinis ne peuvent cependant pas occuper une autre position que celle de premier maillon : ils ne co-réfèrent pas ([Corblin, 1987](#)).

Deuxième maillon : cette position est remplie principalement par des expressions anaphoriques qui ont pour fonction de maintenir le référent activé. Dans cette position, on retrouve principalement des pronoms ou des SN démonstratifs. [Corblin \(1987\)](#) affirme que les SN démonstratifs « instaurent une relation contextuelle forte avec quelque élément antécédent », d'où l'impossibilité de les interpréter indépendamment de leur antécédent, et donc de les rencontrer comme les premiers maillons de CR. Les démonstratifs seraient utilisés, par exemple, dans la reprise recatégorisante (e.g. *cette illustre convertie* (2)) ou pour apporter une information supplémentaire sur le référent (e.g. *Cette femme comblée de tous les dons* (2)), mais aussi comme reprise immédiate d'un SN indéfini ([Corblin, 1995](#)). Selon [Corblin \(1987\)](#), on pourrait également trouver dans cette position des SN définis. Il affirme que ce type de syntagme peut occuper aussi bien la position de premier maillon que de maillons successifs (e.g. *L'insecte* (1)). Les syntagmes définis figureraient en première position quand ils peuvent être interprétés indépendamment du contexte, dans ce cas le syntagme est complet et porteur de toute l'information nécessaire pour son interprétation.

Troisième maillon ou maillon-n : cette position n'est pas unique, c'est la raison pour laquelle [Cornish \(1998\)](#) l'appelle maillon-n. Dans cette position, on peut trouver des formes de haute accessibilité comme le sujet zéro et les pronoms clitiques. Toutefois, les SN définis et démonstratifs ne sont pas exclus de cette position (e.g. *La pauvre bête* (1)). Au contraire, l'utilisation de ces formes est préférée lorsqu'une compétition des référents risque d'apporter une confusion. Cependant, dans ces cas, c'est la reprise par le biais d'un syntagme nominal défini qui va être préférée, comme le montrent [Corblin \(1987\)](#) et [Kleiber \(1994\)](#). Ils montrent que le choix entre défini et démonstratif est influencé par la présence d'une compétition référentielle (e.g. *L'insecte* (1)). En effet, quand, dans la phrase précédente, deux référents sont instanciés, la reprise via un démonstratif risque d'être ambiguë. D'après [Corblin \(1995\)](#), les SN démonstratifs seraient préférés quand il s'agit de comparer entre eux deux référents. Ils permettraient alors de situer un référent par rapport à l'autre. [Salles \(2015b\)](#) souligne que les tendances que nous venons d'illustrer proviennent des études qui se fondent

sur une approche linéaire, où la première mention correspond à la première position, et donc au premier maillon de la chaîne. De plus, la description de ces tendances reste générale et ne prend en considération ni la saillance des référents, ni la distance entre les maillons.

Pour résumer, il est donc plus probable d’avoir des SN ou des noms propres en début de chaîne, car ceux-ci ont pour fonction d’introduire pour la première fois un référent qui n’est pas connu. Par la suite, l’utilisation des pronoms est possible, car le référent a été activé. Cette vision s’accorde avec l’échelle d’accessibilité (cf. 1.3.1). Le choix des expressions référentielles dépend d’autres facteurs que leur position dans la chaîne. Salles (2015b) met l’accent sur le fait que la saillance du référent et les relations de cohérence qui s’établissent entre les phrases jouent un rôle important dans le choix des expressions référentielles. Elle affirme que la nature de la première mention et donc du premier maillon influence la forme de la deuxième mention : « si le référent est, dès sa première mention, présenté comme (relativement) actif, sa deuxième mention se fera probablement au moyen d’une expression codant un statut actif, tels les pronoms (haute accessibilité pour Ariel (2001)) ». Nous reviendrons sur l’impact d’autres facteurs, par exemple, la nature ontologique des référents, le type de texte ou le genre textuel, dans le chapitre 2 (cf. 2.3).

La forme des expressions référentielles et leur degré d’autonomie ne conditionnent pas seulement la position que celles-ci peuvent occuper dans les chaînes, mais également la nature des relations qu’elles entretiennent entre elles.

1.1.2 Relation entre les maillons

Les maillons qui constituent une chaîne entretiennent des relations de coréférence. La notion de coréférence coïncide avec une relation d’identité, en effet les mentions réfèrent à la même entité.

Selon le degré d’autonomie des expressions référentielles, ces relations peuvent être anaphoriques ou non anaphoriques. Lorsque les ER peuvent être interprétées *directement*, c’est-à-dire indépendamment de l’antécédent, la relation entre maillons est dite *coréférentielle non anaphorique*. C’est le cas, par exemple, de la relation entre *Elisabeth Seton* et *Madame Seton* dans l’exemple (2). Ces deux expressions sont liées par une relation de coréférence non-anaphorique, car elles désignent le même référent, mais l’interprétation de la deuxième ne repose pas sur la présence de la première : elles « permettent d’identifier le référent indépendamment de toute information sur le contexte où elles sont employées » (Charolles, 2002, 76).

À l’inverse, si une expression référentielle nécessite la prise en compte du contexte, et donc des maillons précédents, pour pouvoir être interprétée, la relation qui s’établit entre les maillons est dite *coréférentielle anaphorique*. Cette relation est instanciée par les formes pronominales, les SN définis ou démonstratifs non complets ou les formes possessives (Schnecker & Landragin, 2014). Les expressions anaphoriques nécessitent la présence d’un antécédent explicite pour pouvoir être interprétées.

Plusieurs types d'anaphore peuvent être distingués. À ce propos, [Schneidecker et Landragin \(2014, 4\)](#) expliquent que seuls certains types d'anaphores instaurent une relation de coréférence :

- les anaphores fidèles : elles se caractérisent par le fait de « réinstancier la tête lexicale d'un syntagme nominal antérieur », par exemple *Élisabeth Seton ... Madame Seton ... Élisabeth Seton* ou *un grillon ... le/ce grillon* ;
- les anaphores infidèles : la tête lexicale varie par rapport au syntagme nominal précédent. Ce type d'anaphore inclut les anaphores hyperonymiques (*un grillon... l'insecte*), les anaphores recatégorisantes (*Madame Seton... cette illustre convertie*), et les anaphores pronominales (*Élisabeth Seton... elle...*)³.

En fonction du type de reprise anaphorique, [Manuélian \(2003\)](#) propose une distinction entre reprise directe et reprise indirecte. La reprise directe se manifeste quand la reprise de l'antécédent est faite par le biais d'une anaphore fidèle, donc avec la reprise de la tête lexicale du syntagme nominal antérieur. En revanche, si la reprise est faite par une anaphore infidèle, on parle de reprise indirecte. La reprise indirecte peut être décomposée en fonction des différentes relations sémantiques qui lient les maillons. [Manuélian \(2003\)](#) propose une distinction entre reprise par un hyperonyme, reprise par un hyponyme et reprise par un synonyme. On peut également rencontrer des cas où la reprise est faite en utilisant la « grille thématique d'un verbe » ([Manuélian, 2003](#)), par exemple « ils revendiquaient $\Rightarrow$ cette revendication⁴ ». Il existe également des cas où l'antécédent et l'anaphore n'entretiennent pas de relation lexicale identifiée. C'est le cas des anaphores recatégorisantes (e.g. *Cette illustre convertie... Une des gloires de l'Amérique (2)*). Ce type d'anaphore permet au locuteur d'apporter une nouvelle information sur le référent. D'autres types d'anaphore, telles les anaphores associatives ou résomptives, n'établissent pas de lien coréférentiel. Pour cette raison, elles ne sont pas considérées dans les chaînes de référence. En effet, bien que ces anaphores contribuent à la continuité référentielle, l'identité du référent n'est pas la même. [Recasens, Hovy, et Martí \(2010\)](#) parle à ce propos de quasi-identité et propose une typologie des différents cas de figure que cette notion couvre.

Les relations entre les maillons, et notamment les liens anaphoriques, jouent un rôle fondamental dans l'organisation discursive. La succession des ER coréférentes (anaphoriques ou pas) créent une continuité référentielle au sein du discours. Nous considérons donc que les chaînes de référence jouent un rôle fondamental dans l'organisation et l'interprétation du discours, en contribuant à la création des liens de cohésion (*cohesive ties*) ([Halliday & Hasan, 1976](#)). C'est cette fonction cohésive des chaînes de référence que nous allons à présent explorer plus en détail.

1.2 Un indice de cohésion

[Charolles \(1988\)](#) considère les CR comme l'un des quatre plans d'organisation textuelle dont l'interaction réciproque permet de doter le discours de continuité :

3. Cette différenciation est tirée de ([Schneidecker & Landragin, 2014, 4](#)). Nous illustrons avec les exemples (1) et (2).

4. Cet exemple est tiré du corpus AnnoDis ([Péry-Woodley et al., 2011](#)).

« les langues possèdent des marques (ou systèmes de marques) destinées à indiquer les relations qu’entretiennent les unités composant un discours, donc des marques permettant de fournir à un destinataire potentiel des instructions interprétatives propres à favoriser la compréhension » (Charolles, 1988, 4)

Les CR jouent ainsi un rôle fondamental dans le discours, comme le souligne Schne-decker (2021) :

« [...] [L]es chaînes de référence sont au cœur de l’activité langagière : dans une situation de production « normale », on ne parle ni n’écrit sans continuité et celle-ci passe par les éléments qui émergent d’un discours, par leur suivi et leur portée. » (Schne-decker, 2021, 241)

Afin de comprendre l’importance des CR dans la structuration et l’interprétation du discours, prenons le texte suivant (ex. (3)), rédigé par un enfant de CE2⁵, où les maillons sont singalés en gras et en couleur pour les différencier.

- (3) **Cette maison** est hanté.
Elle habitait dans **cette maison** depuis longtemps..
Il marcha marcha. **Il** se retourna en entendant ce grand bruit.
Maintenant ça fait très peur à tout le monde. Dans la rue il y a presque personne qui ose sortir, il y a que les gens courageux qui sortent dehors.
Depuis cette aventure, les enfants ne sortent plus la nuit. [ResolCo_{CE2} ; 9]

De quelle maison est-il question dans ce texte ? Qui sont *Elle* et *Il* ? De quel bruit s’agit-il ? Qui ou quoi fait peur à tout le monde ? De quelle aventure parle ce texte ? Et qui sont les enfants qui ne sortent plus la nuit ?

Ces questions montrent que le texte manque, entre autres, d’accessibilité et de continuité référentielle, considérée dans la systémique fonctionnelle comme un des procédés de construction des liens de cohésion (Halliday & Hasan, 1976). Le problème principal dans ce texte porte sur l’introduction des référents : les ER utilisées ne semblent pas être adéquates. Certaines expressions sont plus à même d’introduire un nouveau référent dans le discours. Ce sont celles qui sont autonomes d’un point de vue référentiel. D’un point de vue fonctionnel, les expressions référentielles ont une valeur instructionnelle qui guide le lecteur dans l’interprétation des différentes unités textuelles. Ainsi, leur forme est plus ou moins adéquate pour l’introduction des référents et pour leur maintien.

D’un point de vue cognitif, selon leur forme, les expressions référentielles signalent le statut cognitif du référent et son accessibilité (cf. 1.3.1). Ainsi, la forme des ER doit être conforme au statut du référent, en termes d’accessibilité. Ainsi, dans le texte (3) l’élève utilise un SN démonstratif pour introduire le premier référent, *cette maison*. Le démonstratif porte le lecteur à penser que le référent est accessible et qu’il devrait

5. Le texte est tiré du corpus RésolCo (Garcia-Debanc, Ho-Dac, Bras, & Rebeyrolle, 2017). Nous avons conservé l’orthographe du texte d’origine consultable à l’adresse suivante : http://redac.univ-tlse2.fr/corpus/resolco/copies/html_normalise/EC-CE2-2015-PBOX-D1-R3-V1_N.html. Soulignons également que dans le texte, il n’existe pas vraiment de CR car il n’y a pas trois maillons d’un même référent.

connaître la maison dont il est question, soit parce que cette entité a déjà été introduite et est donc accessible, soit parce que les interlocuteurs partagent le même contexte. Or, aucune de ces deux situations n'est avérée ici. Il en va de même pour les deux personnages introduits dans les phrases suivantes. L'élève utilise des pronoms personnels. Or, les pronoms n'ont pas d'autonomie référentielle et pour cette raison, ils ne sont pas adaptés pour introduire de nouveaux référents. L'interprétation correcte des pronoms dépend de la présence d'une mention du référent qui soit autonome d'un point de vue référentiel (cf. 1.1). Les pronoms peuvent être utilisés lorsque le référent est accessible et actif. Nous ne sommes donc pas en mesure de comprendre qui sont les personnages du texte. Dans l'exemple (3), le manque de continuité est dû au fait que les référents sont mentionnés successivement dans le texte : un référent est mentionné dans une ou deux phrases maximum et il est ensuite abandonné, pour laisser la place à un nouveau référent. Cela crée des continuités référentielles « à court terme ». Il manque une continuité référentielle « à long terme » permettant de doter le texte d'une certaine *texture*.

Si nous comparons le texte d'élève avec l'exemple (2), où la gestion de la continuité référentielle est bien maîtrisée, nous comprenons l'importance des chaînes de référence dans la création des liens de cohésion. Dans ce texte⁶, le référent a été introduit avec un nom propre, ce qui, d'une part, permet d'identifier correctement le référent et, d'autre part, donne une idée de l'importance du référent : nous comprenons qu'il s'agit du personnage principal, ou du thème. Le référent est ensuite maintenu dans le texte par le biais d'expressions référentielles distinctes, toutes permettant de comprendre que l'on continue à parler du même référent. La succession des différentes expressions référentielles crée ici les liens de continuité référentielle qui, avec d'autres indices de cohésion, permettent de donner au texte une certaine *texture*.

La notion de *texture* indique la qualité propre à une production, qui permet au lecteur de l'appréhender comme un texte et non pas comme une simple succession de phrases isolées (Halliday & Hasan, 1976) :

« The concept of TEXTURE is entirely appropriate to express the property of "being a text". A text has texture, and this is what distinguishes it from something that is not a text. It derives this texture from the fact that it functions as a unit with respect to its environment » (Halliday & Hasan, 1976, 2)

D'après les auteurs, la texture se construit principalement grâce aux liens de cohésion, définis comme des relations sémantiques qui existent entre les différentes unités du texte. La notion de *cohésion* indique alors qu'il existe, à l'intérieur d'un texte, des relations sémantiques, qui permettent de le considérer comme une unité.

« The concept of cohesion is a semantic one ; it refers to relations of meaning that exist within the text, and that define it as a text » (Halliday & Hasan, 1976, 4)

6. L'exemple 2, contrairement au texte d'enfant, instancie un seul référent principal. Nous soulignons également qu'il ne s'agit pas seulement d'une question de maîtrise : le corpus RésolCo est constitué de textes rédigés à partir d'une consigne précise et complexe qui confronte volontairement les enfants à des problèmes référentiels (Garcia-Debanc et al., 2017).

La cohésion d'un texte signifie que les différentes parties de ce texte sont reliées entre elles par des liens de sens. Les auteurs soulignent que la cohésion se manifeste quand l'interprétation d'un élément du discours dépend de la présence d'un autre :

« Cohesion occurs where the INTERPRETATION of some element in the discourse is dependent on that of another. The one PRESUPPOSES the other, in the sense that it cannot be effectively decoded except by recourse to it. When this happens, a relation of cohesion is set up, and the two elements, the presupposing and the presupposed, are thereby at least potentially integrated into a text. » (Halliday & Hasan, 1976, 4)

Les CR contribuent donc à conférer au texte une cohésion particulière, une texture bien définie.

1.3 Un héritage diversifié

Bien que la plupart des auteurs adoptent la définition qui décrit une CR comme « la suite des expressions d'un texte entre lesquelles l'interprétation construit une relation d'identité référentielle » (Corblin, 1995), la notion de CR hérite de la diversité des points de vue concernant la notion même de *référence*. Étant donné l'importance de ce phénomène dans l'organisation et l'interprétation du discours, plusieurs disciplines se sont intéressées à la référence, apportant chacune des éclairages sur les phénomènes de coréférence. Les sections qui suivent présentent les principaux domaines qui se sont intéressés à la référence et aux chaînes de référence. Si notre travail se situe dans une perspective linguistique, il s'appuie également sur les apports des différentes disciplines. L'organisation des sections suit une progression allant des approches plus générales du traitement de la référence vers les analyses plus ciblées des chaînes de référence et des outils descriptifs propres au domaine de la linguistique et du TAL.

1.3.1 Héritage cognitif

En linguistique cognitive, les études s'intéressent aux principes qui régissent l'interprétation des expressions référentielles. Ces études ont donné lieu à des théories qui expliquent les choix des expressions référentielles et leur interprétation sur la base du statut cognitif du référent, de son accessibilité ou de son degré d'activation dans la mémoire discursive (Ariel, 2001 ; Grosz, Joshi, & Weinstein, 1995 ; J. K. Gundel, Hegarty, & Borthen, 2003). Ces études ont proposé des théories axées sur le traitement cognitif des référents. Ici, nous nous limitons à détailler la théorie de l'accessibilité, sur laquelle nous nous appuyons pour la description des résultats en corpus. Néanmoins, il convient de mentionner la théorie du centrage (Grosz et al., 1995). Il s'agit d'une théorie à visée informatique axée sur le traitement cognitif du discours qui vise à prédire la forme d'ER d'une proposition à l'autre, en exploitant la fonction syntaxique des expressions référentielles, en lien avec leur saillance. De nombreuses études en linguistique et en TAL ont repris ces théories pour les tester en corpus, pour décrire les CR et les ER ou encore pour les opérationnaliser et intégrer leurs prédictions dans des systèmes de résolution de la référence (Longo, 2013 ; Strube & Hahn, 1996).

Théorie de l'accessibilité (Ariel, 2001)

D'un point de vue cognitif, les expressions référentielles fournissent des instructions permettant aux interlocuteurs de récupérer l'information pertinente dans leur mémoire. En d'autres termes, les expressions référentielles fournissent une information sur le degré d'accessibilité mentale d'un référent à un moment donné de la communication. D'après Ariel (2001), chaque expression référentielle manifeste un degré spécifique et différent d'accessibilité mentale. L'accessibilité mentale concerne le degré d'activation d'un référent dans la mémoire des interlocuteurs, c'est-à-dire la facilité cognitive avec laquelle un locuteur accède à une information qu'il a stockée précédemment en mémoire.

Tout locuteur, en lisant un texte, reçoit des informations nouvelles au fur et à mesure qu'il avance dans sa lecture. Ces informations sont stockées dans une mémoire temporaire que l'on appelle *mémoire tampon*. Lorsqu'une information n'est pas très importante pour la suite du texte, elle n'est plus mentionnée, et elle est « effacée » de la mémoire temporaire, qui continue à stocker les nouvelles informations ou à réactiver des informations déjà présentes. La réactivation de l'information se met en place chaque fois que cette information est mentionnée dans le texte. C'est le cas pour les référents : chaque nouvelle mention du même référent contribue à le maintenir activé dans la mémoire des locuteurs, et il sera plus facilement repérable, donc très accessible. En revanche, si le référent n'a pas été mentionné depuis un moment, le coût cognitif pour le récupérer est plus élevé. Dans ce cas, le référent est considéré comme moins accessible.

La théorie de l'accessibilité (Ariel, 2001) propose une échelle qui classe les expressions référentielles selon leur degré d'accessibilité, de haute à faible accessibilité. Pour simplifier, nous parlons ici d'expressions de basse, moyenne ou haute accessibilité, mais pour être exacte, il faudrait parler systématiquement d'expressions renvoyant à un référent de basse, moyenne ou haute accessibilité. Selon l'échelle d'accessibilité proposée par Ariel (2001) (figure 1.1) en fonction du degré d'accessibilité du référent, on peut distinguer :

- les expressions de basse accessibilité : il s'agit d'expressions référentielles très informatives, comme les noms propres ou les descriptions définies (Charolles, 1988 ; Cornish, 1998). *Emmanuel Macron* et *le Président de la République française* peuvent être interprétés indépendamment l'un de l'autre parce qu'elles apportent à elles seules l'information suffisante pour permettre aux interlocuteurs de comprendre à qui le locuteur fait référence. Du fait de leur autonomie référentielle (cf. 1.1), les expressions dites de basse accessibilité permettent de référer à une entité peu ou pas accessible au moment du discours. Ainsi, elles permettent d'introduire un nouveau référent ou de le réactiver. Si le référent a déjà été instancié dans le texte, mais que d'autres référents ont été ensuite mentionnés et entrent en compétition avec lui, le recours aux expressions de basse accessibilité permet de le réactiver.
- les expressions d'accessibilité moyenne : ces expressions renvoient à des référents qui ne constituent pas le focus d'attention du discours (Longo, 2013), mais qui ont déjà été instanciés dans le texte. Selon Ariel (2001), la forme de reprise

privilegiée est le syntagme nominal démonstratif, par exemple : *Élisabeth Seton... cette femme.*

- les expressions de haute accessibilité : elles sont utilisées lorsque le référent est actif et très accessible. Il s'agit des référents saillants ou dans le focus d'attention, c'est-à-dire qu'ils se trouvent au « premier plan de la représentation du discours et qui vont attirer l'attention du lecteur » (Longo, 2013) (le focus selon Longo (2013)). Le référent étant actif dans la mémoire du lecteur, il peut être repris avec des expressions peu informatives et n'ayant pas d'autonomie référentielle, comme les pronoms. Selon Ariel (2001), les expressions de haute accessibilité ne sont pas autonomes d'un point de vue de la référence. De ce fait, leur fonction est de maintenir actif ou dans le focus d'attention le référent dont il est question.

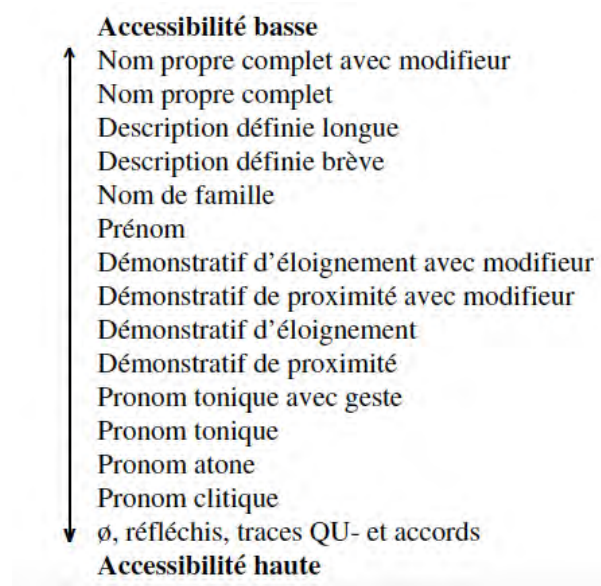


FIGURE 1.1 – Échelle d'accessibilité d'Ariel (1990)

Selon Ariel (1990) (citée par Longo (2013, 77)), le degré d'accessibilité du référent est influencé par quatre facteurs, que nous proposons de regrouper en deux :

- La distance entre les expressions et l'environnement d'identification du référent. Concernant la distance entre l'antécédent et sa reprise référentielle : plus cette distance est grande plus l'accessibilité du référent diminue ; en effet, comme nous l'avons dit plus haut, si le référent n'est pas réactivé, le coût cognitif pour récupérer l'information pertinente sera plus élevé ; l'environnement d'identification des référents : si l'antécédent et sa reprise figurent dans la même phrase ou dans le même paragraphe, le degré d'accessibilité est plus élevé. Longo (2013, 77) souligne que cette caractéristique met en évidence que la structure du discours peut influencer la forme de la référence. Ces deux facteurs, d'après nous, se recoupent : la présence de l'antécédent et de sa reprise référentielle dans la même phrase correspond à une faible distance entre les deux (la distance entre les expressions référentielles pouvant être appréhendée en nombre de phrases ou de paragraphes, comme nous le verrons dans le chapitre 2) ;

- la présence d'autres antécédents candidats fait diminuer le degré d'accessibilité du référent ; ce facteur est, d'après nous, lié à celui de la saillance du référent : plus un référent est saillant, plus il est accessible ; la notion de saillance recoupe celle de focus. Un référent est saillant quand il constitue l'élément central sur lequel se focalise l'attention du lecteur. Cela signifie qu'il est fréquemment mentionné dans le texte et qu'il reste actif dans la mémoire du lecteur, qui pourra accéder à l'information pertinente avec un coût cognitif bas.

Enfin, [Ariel \(2001, 22\)](#) affirme que la corrélation qui existe entre formes et fonctions des référents et le codage du degré d'accessibilité n'est pas arbitraire ; trois critères sont, en effet, impliqués :

- *l' informativité* qui concerne le niveau d'information lexicale porté par l'expression de reprise référentielle ; Nous pouvons faire un parallèle avec la distinction entre ER autonomes et non autonomes d'un point de vue référentiel ;
- *la rigidité* qui se caractérise par la capacité de l'expression référentielle à être unique ;
- *l'atténuation* qui correspond à la « brièveté de l'expression référentielle » ([Longo, 2013](#)).

La combinaison de ces trois paramètres permet de prédire le degré d'accessibilité d'un référent donné. Plus l'expression référentielle est informative, rigide et peu atténuée (SN définis complets, Np), moins le référent est accessible. Inversement, une expression référentielle peu informative, peu rigide et plus atténuée (pronoms), indique que le référent a un degré d'accessibilité plus élevé.

[Schneidecker \(2005\)](#), en étudiant les caractéristiques des chaînes de référence dans les portraits journalistiques, montre que l'application de cette théorie sur des données linguistiques réelles est limitée car elle ne prend pas en considération le genre textuel, qui constitue un paramètre important du fonctionnement des CR. Dans ses études, elle montre, en effet, que les chaînes de référence sont tributaires du genre textuel. [Longo \(2013\)](#) utilise, elle aussi, le calcul de l'accessibilité comme complément à d'autres critères, comme la prise en compte du genre textuel, pour l'identification automatique des chaînes de référence. Dans notre étude, nous mobilisons cette théorie pour la description des stratégies d'introduction et de maintien du référent. Sur la base des études citées ci-dessus, nous nous attendons à identifier des cas qui diffèrent de ce qui est prédit par cette théorie et identifier des stratégies liées à des facteurs tels les genres textuels ou la nature des référents.

1.3.2 Héritage psycholinguistique

En psycholinguistique, les études portent principalement sur les facteurs qui influencent les choix des expressions référentielles (ER). Les études menées dans ce domaine construisent des protocoles expérimentaux permettant de contrôler les facteurs pouvant influencer le choix des expressions référentielles : type de référent, saillance visuelle, compétition des référents, etc. Ces protocoles sont construits ad hoc pour répondre à des objectifs spécifiques. Certaines études, par exemple [Wolna, Durlík, et Wodniecka \(2022\)](#), s'intéressent aux processus d'interprétation de l'anaphore. En utilisant la technique de l'eye-tracking, ces derniers étudient le processus de récupération de l'antécédent en temps réel.

Deux études sont particulièrement intéressantes pour notre travail, dans la mesure où elles apportent un éclairage sur les stratégies d'introduction et de maintien des référents et permettent de confronter les résultats issus de notre analyse de corpus et ceux obtenus dans des expérimentations impliquant des participants. Fossard et al. (2018) s'intéressent à la compétition des référents et à leur saillance visuelle. Les auteures étudient comment les locuteurs ajustent leurs choix référentiels lors d'une tâche de narration collaborative. Les chercheurs ont analysé les expressions référentielles utilisées par les participants à trois étapes discursives : introduction du personnage, maintien du personnage en focus et changement vers un autre personnage en focus. Les résultats montrent que pour l'introduction d'un personnage, les locuteurs utilisent principalement des syntagmes indéfinis. Concernant le maintien, l'utilisation de pronoms est plus fréquente lorsque l'histoire ne comporte qu'un seul personnage. En présence de deux personnages, les locuteurs utilisent davantage d'expressions définies pour éviter toute ambiguïté. Dans la phase de changement, lorsque les deux personnages sont du même genre, les locuteurs enrichissent leurs expressions référentielles avec des informations supplémentaires (par exemple, « la fille blonde ») pour clarifier le référent.

Une autre étude (Achim, Fossard, Couture, & Achim, 2015) a examiné comment les locuteurs adaptent leurs expressions référentielles en fonction des connaissances probables de leur interlocuteur, et comment cette adaptation est liée aux capacités de théorie de l'esprit (ToM). Les résultats obtenus montrent que les participants fournissent des descriptions plus détaillées, notamment en ajoutant des descripteurs physiques, pour les personnages probablement inconnus de l'interlocuteur, même lorsque aucune demande explicite de clarification n'a été formulée. Ces résultats indiquent que les locuteurs adaptent spontanément leurs expressions référentielles en fonction des connaissances probables de leur interlocuteur.

Les deux études mettent en évidence que les locuteurs adaptent les stratégies d'introduction et de maintien du référent selon le contexte et selon le degré de familiarité supposé de l'interlocuteur. Pour ce qui concerne l'introduction des référents, nous retenons que les résultats convergent vers une préférence des syntagmes nominaux indéfinis. Nous retrouvons ici les tendances décrites par Corblin (1987) et Ariel (2001). Concernant le maintien d'un référent en focus, les locuteurs utilisent des pronoms lorsque le contexte est peu ambigu (par exemple quand un seul personnage est présent) ; ils optent par des syntagmes nominaux définis dans les cas de compétition référentielle. Enfin, lorsque les locuteurs doivent procéder à un changement⁷ de référent, ils enrichissent les expressions référentielles avec des modificateurs pour éviter toute ambiguïté. Nous pouvons ici retrouver les tendances décrites par la théorie de l'accessibilité.

1.3.3 Héritage linguistique

Pour ce qui concerne les études en linguistique, plusieurs directions ont été suivies. La description des CR repose sur un certain nombre de critères (cf. chapitre 2), comme le nombre de maillons qui constituent une chaîne ou la catégorie grammaticale des maillons. Dans les études linguistiques, on ne parle de CR qu'à partir de

7. Cet aspect ne sera pas abordé dans le cadre de cette thèse : les annotations que nous exploitons ne permettent pas en l'état d'analyser les changements de référent.

trois maillons. Cela permet de distinguer cette notion de celle d'anaphore et de coréférence (Schneidecker, 2021). La notion de CR permet, en effet, d'appréhender certains phénomènes référentiels qui pourraient être biaisés par l'analyse par paire typique de l'anaphore ou de la coréférence. C'est le cas par exemple, des discordances de genre entre les expressions référentielles (e.g. *un grillon... La pauvre bête* (1)) (Schneidecker, 2021) (cf. 2.1.4). De plus, le fait de dépasser le traitement par paire permet de suivre le référent tout le long du discours, et suivre ainsi sont *devenir discursif* (Schneidecker, 2021).

En linguistique contrastive, les études se sont focalisées sur les variations des caractéristiques des chaînes en fonction de la langue (Baumer et al., 2023 ; Kunz & Lapshinova-Koltunski, 2015). En linguistique descriptive, les études ont porté sur la description linguistique des caractéristiques des expressions référentielles (ER) et des CR. Parmi les travaux les plus influents, on peut citer ceux de Corblin (1987, 1995) ; Salles (2015a) qui analysent les fonctions des ER selon leur forme (catégorie grammaticale) et leur position dans la CR (cf. 1.1.1). Nombreuses sont les études qui se sont intéressées aux variations liées au genre textuel, à la langue ou à l'époque (Baumer et al., 2023 ; Longo, 2013). Nous reviendrons en détail sur les apports de ces études dans le chapitre 2, où nous présentons les critères de description et les facteurs de variation.

Dans une perspective fonctionnelle, les travaux de Givón (1983) sur la continuité thématique offrent un cadre d'analyse intéressant, bien qu'ils portent sur la gestion du topique plutôt que sur la notion de chaîne de référence. Givón propose de distinguer les *topiques continus*, si la distance entre les expressions référentielles ne dépasse pas les vingt phrases, les *topiques persistants*, si les référents sont réinstanciés immédiatement dans la phrase successive et les *topiques discontinus* lorsqu'il y a une compétition avec d'autres référents ou une distance supérieure à 20 phrases. Cette typologie repose sur plusieurs critères, notamment la distance entre les expressions référentielles (cf. 2.2.3), la persistance (cf. 2.2.4) et la présence d'autres référents concurrents. Selon cette approche, la forme de l'ER mobilisée pour maintenir le référent varie en fonction du type de topique : un topique continu est généralement repris par un pronom, tandis qu'un référent moins saillant ou en concurrence est plus souvent réintroduit par un syntagme nominal.

Comme le souligne Schneidecker (2021), en appliquant la théorie de Givón (1983) aux chaînes de référence, on s'attendrait à distinguer deux types de stratégies différentes : dans le cas d'un référent saillant (topique continu), la CR est généralement longue (plus de 10 phrases consécutives d'après Givón) et elle est composée principalement de pronoms. À l'inverse, lorsque le référent est moins saillant ou entre en compétition avec d'autres référents, la chaîne serait plus courte et le référent régulièrement réinstancié par le biais de syntagmes nominaux. Cependant, ces hypothèses sont à nuancer. Schneidecker (2021) montre que ces patrons ne s'observent pas systématiquement, en particulier lorsque le référent n'est pas humain. Elle souligne également que le type de texte a un impact sur les modes de composition des chaînes (cf. 2.3.1). Ces apports permettent de commencer à dégager des types de chaînes différents, comme des CR longues, constituées de formes peu variées d'ER, ou des CR plus courtes qui alternent différentes formes d'ER.

1.4 Bilan

Ce premier chapitre avait pour objectif de poser le cadre de notre travail. Nous avons abordé les chaînes de référence comme des structures discursives complexes, composées de maillons aux caractéristiques variées et entretenant des relations de coréférence. En nous inscrivant dans une perspective fonctionnelle, qui envisage les formes linguistiques selon leur rôle dans l’organisation du discours, nous considérons les chaînes de référence comme des manifestations de stratégies discursives. L’enchaînement de maillons assure la continuité référentielle, participant ainsi à la cohésion du discours. C’est cette capacité des chaînes à structurer le discours qui nous intéresse. Notre objectif est d’identifier les stratégies mises en place pour introduire et maintenir un référent. Pour ce faire, nous adoptons une définition exigeant au moins trois maillons dans une chaîne (cf. [Schnedecker 2021](#)), seuil minimal qui permet de distinguer les chaînes des cas de coréférence ou d’anaphore. Par ailleurs, les relations entre maillons sont ici considérées comme des relations de coréférence stricte, conformément à l’annotation des données que nous exploitons, ce qui cadre notre analyse.

Nous nous appuyons également sur les apports d’autres disciplines. Les travaux psycholinguistiques, fondés sur des données expérimentales, offrent un éclairage précieux sur les choix référentiels en contexte. Nous avons montré que les résultats nourrissent notre réflexion sur les stratégies d’introduction et de maintien, bien que notre analyse ne permette pas de traiter la question des changements de référent (shifts), qui nécessiterait une annotation spécifique.

Nous mobilisons également plusieurs cadres théoriques pour guider notre analyse :

- La théorie de l’accessibilité ([Ariel, 2001](#)) nous permet d’interpréter le choix de formes référentielles selon le degré d’activation du référent dans la mémoire du lecteur ou de l’interlocuteur ;
- La typologie de [Givón \(1983\)](#) sur la continuité thématique nous permet de distinguer différents types de chaînes — continues, discontinues ou persistantes — en fonction de la distance, de la persistance ou de la compétition référentielle. Ainsi, selon le type de chaîne, leur composition (forme des expressions référentielles) donne lieu à des enchaînements différents.

Nous nous appuyons sur ces cadres pour l’analyse des chaînes et pour discuter les cas qui s’écartent des prédictions attendues, comme l’ont déjà signalé certains travaux, notamment ceux de [Schnedecker \(2021\)](#).

En ce qui concerne le TAL, bien que nous ne mobilisions pas directement les systèmes de résolution de la coréférence dans cette étude, leur prise en compte ouvre des perspectives intéressantes. Nos résultats descriptifs pourraient, en effet, alimenter une réflexion sur la typologie des erreurs et sur l’évaluation des systèmes ([Grobol, 2020](#)) ou pour mutualiser des ressources pour l’apprentissage des systèmes ([Nedoluzhko, Novák, Popel, Žabokrtský, & Zeman, 2021](#)). Ces pistes seront abordées dans la conclusion de la thèse. Enfin, les descriptions linguistiques existantes — notamment celles concernant la position des maillons dans la chaîne et leur fonction — serviront également de référence pour détecter des usages atypiques ou marqués dans nos corpus.

Pour poursuivre vers notre objectif, il est nécessaire de définir les critères permettant de caractériser les maillons et les chaînes de référence. Le chapitre suivant est ainsi consacré à la présentation des critères de description mobilisés dans la littérature et aux facteurs qui ont un impact sur les caractéristiques des CR.

Chapitre 2

Description des chaînes de référence

Contents

2.1	Composition des chaînes de référence : critères <i>micro</i>	31
2.1.1	Nature grammaticale des maillons	31
2.1.2	Fonction syntaxique	33
2.1.3	Coefficient de stabilité	33
2.1.4	Autres critères	34
2.2	Dimension des chaînes de référence : critères <i>macro</i>	36
2.2.1	Longueur des chaînes	36
2.2.2	Portée des chaînes	37
2.2.3	Distance entre les maillons	38
2.2.4	Persistance	41
2.2.5	Densité référentielle des chaînes	42
2.2.6	Bilan sur les critères de description	42
2.3	Facteurs de variation des chaînes de référence	43
2.3.1	Type de texte	43
2.3.2	Genre textuel	44
2.3.3	<i>Nature</i> du référent	45
2.3.4	Cohabitation des chaînes de référence	46
2.3.5	Autres facteurs de variation	48
2.4	Bilan	51

Nous avons montré dans le [chapitre 1](#) que les chaînes de référence sont des structures complexes. Les analyser, en vue d'en dresser une typologie, requiert la prise en compte d'un système d'indices — ou critères de description — qui permette de caractériser les chaînes selon différents aspects : morpho-syntaxique, syntaxique, sémantique et discursif. Comme l'explique [Schneidecker \(2021\)](#), les indices en question correspondent à des critères formels et opérationnels qui favorisent une analyse objective des CR. Utiliser ces critères pour caractériser les CR permet de faciliter la comparaison des chaînes, ce qui, en retour, contribue significativement à discerner leurs différences et similitudes respectives.

Plusieurs critères ont été mis au jour dans les travaux s'intéressant aux chaînes de référence (Kibrik, 2011 ; Kunz & Steiner, 2012 ; Recasens & Hovy, 2010 ; Rousier-Vercruyssen & Landragin, 2019 ; Schnedecker, 2021). Ces travaux ont également montré que les caractéristiques des CR peuvent varier en fonction de plusieurs facteurs (cf. 2.3). Analyser la complexité et la variété de ces structures complexes signifie alors aussi les considérer dans différents types et genres textuels, prenant en compte plusieurs types de référents. Ces facteurs correspondent à ceux que Schnedecker (2021) appelle contraintes *externes*. L'auteure souligne que la composition des CR peut également dépendre de facteurs (ou contraintes) *internes*. Ces derniers sont « liés au texte lui-même et à la manière dont le locuteur entend présenter les référents » (Schnedecker, 2021, 121). Ils concernent notamment l'interaction des CR avec les autres plans d'organisation textuelle, comme par exemple le découpage en paragraphes, la présence de titres, les découpages liés au discours rapporté ou encore les découpages sémantiques. Les facteurs liés à la manière dont le locuteur entend présenter les référents portent notamment sur le sémantisme des expressions référentielles et sur les instructions interprétatives qu'elles véhiculent.

Dans ce chapitre, nous dressons, tout d'abord, un panorama des critères formels qui ont été mis au jour dans les travaux caractérisant les CR, aussi bien en linguistique descriptive qu'en TAL, où certains critères sont exploités comme contraintes pour les systèmes de résolution automatique. Pour certains critères, lorsque des méthodes de calcul alternatives ont été proposées, nous explicitons les variantes possibles. Pour chaque critère, nous précisons également quel aspect des chaînes celui-ci permet de révéler. Enfin, nous discutons des précautions qu'il faudrait prendre au moment de l'interprétation des résultats obtenus par le biais de ces critères. Ensuite, nous nous intéressons aux facteurs de variation que plusieurs travaux ont analysés afin de comprendre si et comment ces facteurs peuvent influencer sur les caractéristiques des chaînes de référence.

Le chapitre est organisé en trois sections principales : les deux premières portent sur les critères de description. En nous appuyant sur l'ouvrage de Schnedecker (2021), nous faisons une première distinction entre deux types de critères :

- la section 2.1 présente les critères permettant de caractériser la *composition* des chaînes, tels la nature des maillons ou leur fonction syntaxique. Ces critères se situent au niveau des maillons, considérés comme les unités minimales constituant les CR ;
- la section 2.2 porte sur les critères permettant de caractériser la *dimension* des chaînes de référence, comme par exemple la longueur de la CR ou sa portée. Il s'agit de critères qui se situent au niveau de la chaîne dans sa *totalité*, c'est-à-dire en tant qu'unité d'analyse à part entière.

La section 2.3 se focalise sur les facteurs qui peuvent avoir un impact sur la variation de la *composition* et de la *dimension* des chaînes.

Par souci de clarté et de lisibilité, nous présentons les critères de description ainsi que les facteurs de variation en les abordant chacun successivement et de façon distincte. Nous tenons toutefois à souligner que, pris isolément, ces critères ne permettent pas de saisir toute la complexité des chaînes : c'est leur prise en compte conjointe qui rend possible une caractérisation fine des CR, bien qu'elle puisse s'avérer complexe à

mettre en œuvre. Nous évoquons cette complémentarité dans les sections suivantes, et y revenons de façon plus approfondie dans le bilan final (cf. 2.4). De la même manière, les facteurs de variation n’agissent pas de manière indépendante : plusieurs d’entre eux simultanément peuvent avoir un impact sur la dimension et la composition des chaînes.

2.1 Composition des chaînes de référence : critères *micro*

La *composition* des chaînes concerne leur structure interne, telle qu’elle se manifeste à travers l’enchaînement des maillons qui les composent. Décrire cette composition revient ainsi à décrire les caractéristiques des maillons au moyen de différents critères, tels que leur catégorie grammaticale, leur fonction syntaxique, etc. La description peut s’effectuer en prenant en compte l’ensemble des maillons qui composent les chaînes, ce qui revient à s’intéresser à la fréquence et à la distribution, dans le corpus, des valeurs associées au critère de description considéré. Autrement dit, il s’agit de caractériser les CR en analysant la répartition des différentes valeurs de chaque critère appliqué aux maillons. Cette approche, la plus répandue dans les études existantes, permet de donner une idée de la composition des CR de manière globale. Certains auteurs se focalisent également sur le premier maillon des CR (Longo & Todirascu, 2014 ; Manuélian & Schnedecker, 2021), ce qui permet d’étudier les modalités d’introduction des référents dans le corpus et de les comparer en fonction de différents facteurs. Parfois, les descriptions distinguent le premier maillon, le deuxième et ensuite tous les autres considérés globalement (Federzoni, Ho-Dac, & Rebeyrolle, 2020). Les sections qui suivent portent sur les différents critères et sur l’aspect des chaînes ou des maillons qu’ils permettent de révéler.

2.1.1 Nature grammaticale des maillons

Un des critères les plus utilisés pour décrire la composition des CR est la nature des maillons. Il s’agit d’un trait de surface qui correspond à la catégorie grammaticale des maillons. Ce critère fournit des informations riches quant à la composition des chaînes de référence, permettant par exemple, d’analyser le degré d’homogénéité des chaînes de référence (Obry, Glikman, Guillot-Barbance, & Pincemin, 2017). Les auteurs soulignent également que la nature des maillons favorise l’étude des référents au regard de la théorie de l’accessibilité, en considérant la fréquence des expressions qui marquent une accessibilité faible, moyenne ou élevée. Ils précisent par ailleurs, que « [l]a distribution de ces catégories dépend du caractère mono- ou pluri-référentiel du texte et des risques d’ambiguïté ; les marques de faible accessibilité étant moins attendues en cas de compétition référentielle. » (Obry et al., 2017).

Obry et al. (2017) affirment que la nature des maillons est un des traits qui présente les variations les plus fortes et qui fournit des informations riches sur la typologie des chaînes de référence. En fonction des formes qui composent une CR, Schnedecker (2005), dans une étude sur les CR à référent humain dans les portraits journalistiques, distingue deux types de chaînes :

- les chaînes *homogènes*, caractérisées par une alternance de noms propres et pronoms (4);

(4) En 1812, **Étienne Geoffroy Saint-Hilaire** entérine la division entre Ancien et Nouveau Monde en proposant la séparation entre les Catarrhini (« singes de l’Ancien continent ») et les Platyrrhini (« singes d’Amérique »).

Il crée également plusieurs nouveaux genres : *Lagothrix* (les singes laineux), *Cercocebus* (les mangabeyes), *Nasalis* (les nasiques) et *Pygathrix* (les doux). [...] Les deux premières familles sont classées dans un groupe nommé « anthropomorphes » et correspondent à la distinction Catarrhini / Platyrrhini de **Geoffroy**. [AE_{w}; 

- les chaînes *hétérogènes*, construites non seulement avec des noms propres et des pronoms, mais aussi de SN définis, démonstratifs et/ou possessifs (5).

(5) **Filippo**, frère de Marianina, tenait, comme **sa** sœur, de la beauté merveilleuse de la comtesse. Pour tout dire en un mot, **ce jeune homme** était une image vivante de l’Antinoüs, avec des formes plus grêles. [...] Si **Filippo** restait, dans tous les cœurs de jeunes filles, comme un type, **il** demeurerait également dans le souvenir de toutes les mères, comme le meilleur parti de France. [...]

Filippo, Marianina, madame de Lanty et un vieux domestique avaient seuls le privilège d’aider l’inconnu à marcher, à se lever, à s’asseoir. [...]

Ou bien **Filippo** s’élançait en se glissant à travers la foule, pour le joindre, et **restait** auprès de lui, tendre et attentif, comme si le contact des hommes ou le moindre souffle dût briser cette créature bizarre. [...]

Si madame de Lanty n’était pas là, **le comte** employait mille stratagèmes pour arriver à lui; mais **il** avait l’air de s’en faire écouter difficilement, et le traitait comme un enfant gâté dont la mère écoute les caprices ou redoute la mutinerie. Quelques indiscrets s’étant hasardés à questionner étourdiment **le comte de Lanty**, **cet homme froid et réservé** n’avait jamais paru comprendre l’interrogation des curieux. [...]

A ce bruit, Marianina, **Filippo** et madame de Lanty jetèrent les yeux sur nous, et leurs regards furent comme des éclairs. [N_{s}; 

Dans les études existantes, la nature des maillons est souvent mise en lien avec la position des maillons dans la chaîne, montrant que les différentes formes fournissent au lecteur des instructions distinctes pour une interprétation correcte du discours (cf. [chapitre 1](#)). Cette idée rejoint la théorie de l’accessibilité : selon le statut cognitif du référent et le rôle de l’expression dans la chaîne, la forme choisie varie (cf. [chapitre 1](#)). Par exemple, d’après [Corblin \(1995\)](#), le premier maillon de la chaîne correspond à des syntagmes nominaux définis ou indéfinis, plus ou moins complets, pour permettre d’introduire correctement un nouveau référent. Récemment, [Manuélian et Schnedeker \(2021\)](#) ont mis en évidence que la nature des premiers maillons peut différer par rapport à ce qui a été traditionnellement décrit : leurs résultats montrent, par exemple, que les SN indéfinis et les noms propres en position initiale sont relativement rares, contrairement à ce que l’on pourrait attendre d’un point de vue théorique.

Comme nous l’avons dit, d’un point de vue opérationnel, la nature des maillons correspond à la catégorie grammaticale de ceux-ci. Nous retenons neuf catégories, présentées dans le [tableau 2.1](#), dont la catégorie *Autre* dans laquelle nous classons tous les cas qui ne sont pas couverts par les autres catégories.

Nature du maillon	Abréviation	Exemple
Noms propres	Np	<i>Filippo</i> (5)
Possessifs	Poss	<i>ses</i> (1); <i>sa</i> (5)
Pronoms	Pro	<i>lui/le</i> (1); <i>Elle</i> (2) / <i>il</i> (5)
Syntagmes nominaux définis	SN _{def}	<i>le comte de Lanty</i> (5)
Syntagmes nominaux démonstratifs	SN _{dem}	<i>ce jeune homme</i> (5)
Syntagmes nominaux indéfinis	SN _{ind}	<i>un grillon</i> (1)
Syntagmes nominaux sans déterminant	SN _∅	<i>procéder</i> (54);
Sujets zéro	∅	<i>restait</i> (5)
Autre	Autre	<i>153/3°</i> (E.5)

TABLEAU 2.1 – Nature des maillons : neuf catégories retenues

2.1.2 Fonction syntaxique

La fonction syntaxique des maillons, combinée avec d’autres critères, est un critère révélateur de la *saillance* des référents. La notion de *saillance*, de nature multifactorielle, demeure difficile à appréhender ([Landragin, 2012](#)). La fonction syntaxique peut contribuer dans ce sens. En effet, il est généralement admis qu’un référent est saillant s’il se trouve dans le focus d’attention, ce qui implique des positions syntaxiques favorables à sa mise en avant. Ainsi, si le maillon occupe la fonction de sujet de la phrase, il est considéré comme plus saillant que celui en position de complément d’objet direct, indirect ou complément du nom.

La fonction syntaxique des maillons constitue un des critères utilisés à la fois en linguistique descriptive et en TAL. La prise en compte de la fonction syntaxique permet une application simple et directe des principes évoqués dans la théorie du centrage ([Chai & Strube, 2022](#) ; [Strube & Hahn, 1996](#)). Cela facilite la modélisation de la saillance référentielle, notamment pour l’identification de l’antécédent d’un anaphorique dans les systèmes de résolution automatique de l’anaphore. Ainsi, dans ces systèmes, les candidats antécédents sont classés en fonction d’un score qui leur est attribué : les mentions occupant la position sujet dans la phrase précédente sont généralement privilégiées, recevant un score plus élevé que celles en position de complément.

2.1.3 Coefficient de stabilité

Le coefficient de stabilité est un critère quantitatif qui se situe au niveau de la chaîne, mais qui prend en compte des caractéristiques des maillons, et notamment la variation des choix lexicaux entre les maillons des CR. Ce critère révèle, comme son nom l’indique, la stabilité (ou homogénéité) lexicale des CR. Moins il y a de variation dans les désignations d’un référent, plus la CR est *stable*. Inversement, de nombreuses

variations dans les désignations du référent indiquent que la chaîne est *instable*. Le coefficient de stabilité a été utilisé pour la première fois par [Perret \(2000\)](#). Le coefficient, tel que proposé par [Perret \(2000\)](#), est calculé en nombre d'anaphores nominales divisé par le nombre de désignations différentes. Sont exclus du calcul les noms propres et les syntagmes nominaux indéfinis. Ce coefficient, calculé en nombre de désignations différentes à l'intérieur d'une même chaîne, permet de rendre compte de l'homogénéité lexicale des CR.

[Oberlé et al. \(2018\)](#) proposent d'effectuer le calcul du coefficient de stabilité en ne prenant en considération que la tête lexicale des maillons. Néanmoins, ils proposent de ne pas lemmatiser les têtes lexicales, afin de conserver la distinction entre le pluriel et le singulier. La modalité de calcul reste la même. Une des limites de ce coefficient concerne sa forte dépendance à la longueur de la chaîne. En réponse à cette limite, [Obry et al. \(2017\)](#) proposent d'utiliser la mesure de diversité proposée par [Pincemin et Malrieu \(2014\)](#). Cette mesure, consistant à faire le décompte de désignations différentes d'un référent, serait alors plus robuste que le coefficient de stabilité.

[Rousier-Vercruyssen et Landragin \(2019\)](#) soulignent, eux aussi, que le coefficient de stabilité est trop dépendant de la longueur de la CR et, donc, il peut difficilement être exploité pour une comparaison des CR. Pour pallier ce problème, ils proposent une adaptation du calcul : ils proposent de diviser le nombre de désignations différentes par le nombre total d'anaphores nominales et le transformer en pourcentage. Un pourcentage élevé est un indice de la présence de nombreuses désignations différentes et la CR est donc *peu stable* ou *instable*. Au contraire, si le pourcentage est bas, cela signifie qu'il y a peu de désignations différentes et donc la chaîne peut être considérée comme *stable*. Dans leur calcul, [Rousier-Vercruyssen et Landragin \(2019\)](#) considèrent la coréférence stricte et le maillon en entier, à la différence d'Oberlé, dont le calcul était effectué en prenant en compte uniquement la tête lexicale.

Le coefficient de stabilité permet de modéliser la composition d'une CR avec un score formel. Toutefois, à l'heure actuelle, les calculs proposés ne sont pas complètement satisfaisants. Comme le souligne [Schneidecker \(2021\)](#), le coefficient pose encore un certain nombre de questions : « limité aux anaphores nominales, il laisse de côté le nom propre (et ses sous-types), qui n'est pas anaphorique, les pronoms personnels (et autres). Par ailleurs, il occulte la question des expansions : certains SN répètent la tête lexicale, mais varient les expansions : *la jeune femme... la jeune femme triste et désespérée*. Bref, il devrait être reconsidéré pour être encore plus opérationnel ([Schneidecker, 2021, 97-98](#)) ».

2.1.4 Autres critères

Dans cette section, nous présentons des critères qui, bien que moins explorés dans les études existantes ainsi que dans notre étude, contribuent également à décrire la composition des chaînes de référence.

Longueur et complexité syntaxique des maillons La longueur est calculée en nombre de mots ou tokens qui composent le maillon. La complexité syntaxique est appréhendée par le nombre et le type de modificateurs inclus dans le maillon. Comme

le montre [Schneidecker \(2021, 94\)](#), ces deux critères « fournissent des indices sur la teneur et/ou le genre de discours dans lequel ils se manifestent ». Ces critères donnent également une idée du degré d’informativité des maillons et peuvent être mis en correspondance avec la position des maillons dans la chaîne et leur fonction. Dans cette perspective, on peut émettre l’hypothèse que le premier maillon d’une chaîne, dont la fonction est d’introduire un nouveau référent dans le discours, présente une longueur supérieure à celle des maillons successifs, qui opèrent des reprises anaphoriques d’un référent déjà activé ou accessible.

En prenant en compte les modificateurs inclus dans les maillons d’un point de vue lexical et sémantique, il est possible d’identifier quel est le positionnement du locuteur ou de caractériser le style de l’auteur ([Kunz, 2010](#) ; [Schneidecker, 2021](#)). Cela permet par ailleurs de décrire la diversité des choix lexicaux. Ce trait peut alors être mis en correspondance avec le coefficient de stabilité (cf. 2.1.3) et avec les types de reprise (cf. 1.1.2). Nous renvoyons à l’ouvrage de [Schneidecker \(2021\)](#) pour une présentation détaillée de ces aspects.

Genre et nombre L’accord en genre et en nombre est un critère de surface, qui relève de l’aspect formel des maillons. Ce critère est principalement utilisé en TAL, notamment dans les premiers travaux centrés sur le développement des systèmes automatiques de résolution de l’anaphore qui abordent la question de la référence en s’intéressant à la paire anaphorique-antécédent. L’accord en genre et nombre est utilisé comme contrainte permettant de classer les antécédents candidats : si l’antécédent s’accorde en genre et nombre avec l’anaphorique, alors il a plus de probabilité d’être le bon antécédent ([Mitkov, 2002, 2005](#)). Cette contrainte est également utilisée dans les systèmes plus récents, basés sur l’apprentissage automatique ([Ng, 2007](#) ; [Sukthanker, Poria, Cambria, & Thirunavukarasu, 2020](#)). Toutefois, [Sukthanker et al. \(2020\)](#) mettent en évidence la limite de cette approche, notamment lorsque plusieurs antécédents possibles s’accordent en genre et en nombre avec l’anaphorique. Ils soulignent également que la seule contrainte de genre et nombre n’est pas suffisante : la résolution de l’anaphore dépend aussi de l’application d’autres contraintes. Par ailleurs, plusieurs études ont montré la limite de l’accord en genre et nombre, notamment dans les cas où la reprise est faite via une anaphore infidèle dont le genre ne correspond pas ([Schneidecker, 2021](#))¹. Étant donné ces limites du point de vue linguistique, ce critère n’est pas vraiment pris en compte dans les travaux descriptifs pour caractériser la composition des chaînes de référence.

Les critères que nous avons examinés jusqu’ici permettent de caractériser les chaînes de référence à travers l’analyse de leurs maillons, en s’intéressant notamment à leur nature, leur fonction syntaxique, ou encore leur stabilité lexicale. Ces critères sont essentiels pour décrire la composition des CR. Toutefois, ils ne suffisent pas à saisir l’ensemble de leur fonctionnement dans l’organisation du discours. En les croisant avec d’autres critères portant sur la dimension globale des chaînes, il devient possible d’en proposer une caractérisation plus fine et plus fidèle à leur complexité, comme nous

1. Nous renvoyons à [Schneidecker \(2021, 58\)](#) qui présente en détail les différents critères et met en évidence comment la notion de chaîne permet de surmonter certaines difficultés liées au traitement par paire typique de l’anaphore.

l'avons souligné en introduction. C'est pourquoi nous nous intéressons à présent aux critères permettant de décrire la dimension de ces structures complexes.

2.2 Dimension des chaînes de référence : critères

macro

Un certain nombre de critères permettent de décrire les caractéristiques des CR à l'échelle *macro*, c'est-à-dire en prenant en considération les chaînes dans leur intégralité, en tant qu'unité d'analyse. Caractériser les CR à ce niveau correspond à décrire ce que [Schneidecker \(2021\)](#) appelle la *dimension* des chaînes de référence.

2.2.1 Longueur des chaînes

Un des critères les plus utilisés dans les études descriptives sur les CR est la longueur, mesure qui reflète la taille de ces structures discursives. D'un point de vue formel, ce critère permet de distinguer les chaînes *courtes*, *moyennes* ou *longues*. Au-delà du simple aspect formel, ce critère permet d'appréhender l'importance des référents dans le texte, ce qui en fait un critère important à prendre en considération. Comme le souligne [Schneidecker \(2021\)](#),

« [l]a longueur des chaînes est très importante, car elle permet, avec d'autres paramètres, de hiérarchiser l'importance des référents les uns par rapport aux autres. Ainsi la chaîne la plus longue coïncide-t-elle avec le référent qui apparaît comme central » ([Schneidecker, 2021](#), 89-90).

Selon l'importance du référent, il est alors possible de distinguer les chaînes *principales* — celles dont le référent correspond à l'entité centrale dans le discours, comme dans (2) et (5) — des chaînes *secondaires* — celles qui correspondent aux référents de moindre importance, comme dans (1). Les chaînes des référents principaux sont potentiellement plus longues que les chaînes secondaires. L'idée sous-jacente étant qu'un référent central et important aura plus de probabilités d'être repris, et cela, jusqu'à la fin du texte. Inversement, un référent secondaire engendrera des CR plus courtes, car il sera mentionné uniquement à un moment donné du discours ou mentionné sporadiquement dans le texte. Cependant, l'intérêt de ce critère ne se limite pas à la comparaison des chaînes d'un même texte. Considérer la longueur des chaînes peut également apporter des informations précieuses pour caractériser différents types ou genres textuels (cf. 2.3).

La longueur des CR est calculée en nombre de maillons. Une autre manière de calculer la longueur des CR est de considérer le nombre de mots qui constituent la chaîne ([Landragin et al., 2024](#)). Dans ce cas, on calcule le nombre de mots inclus dans le segment de texte entre le premier et le dernier maillon de la chaîne, y compris les mots correspondant aux maillons. Néanmoins, cette modalité de calcul n'est pas très répandue dans les études sur les chaînes de référence. Par ailleurs, la longueur telle qu'elle est utilisée dans la grande majorité des travaux est une mesure brute et par conséquent, elle ne rend pas compte des effets de longueur des textes. Utiliser une longueur normalisée permettrait de mieux prendre en considération ces effets et

effectuer des comparaisons en dépit de la différence de longueur des textes. Ainsi, [Baumer, Dias, Gardelle, et Prak-Derrington \(2021\)](#) utilisent une longueur normalisée, consistant à diviser le nombre de maillons de la chaîne par le nombre de mots des articles étudiés. Concernant le calcul de la longueur en nombre de mots et sa version normalisée, il vaut mieux parler de portée des CR. La portée (ou étendue) des CR est un critère complémentaire à la longueur de celles-ci, calculée en nombre de maillons.

2.2.2 Portée des chaînes

La portée d'une CR permet d'analyser l'étendue de la chaîne de référence dans le texte ([Boudreau & Kittredge, 2005](#), 54), c'est-à-dire la portion de texte couverte par la chaîne. Combinée avec la longueur, la portée (ou étendue) des CR permet également de connaître l'importance des référents dans les textes. Théoriquement, plus la portion de texte couverte par la chaîne est importante, plus il est probable que le référent soit central dans le discours. Les CR des référents secondaires couvriraient alors une partie moins importante de texte.

Si la notion de portée semble être un concept simple et intuitif, son calcul requiert néanmoins une certaine attention. Comment mesure-t-on réellement cette « portée » ? Plusieurs unités de mesure peuvent être prises en compte. Mais alors, laquelle faut-il privilégier ?

[Schneidecker \(2021, 92\)](#) souligne que la modalité de calcul la plus pertinente consiste à prendre en considération le découpage en paragraphes ou en chapitres pour les textes longs. Il s'agirait alors de « compter le nombre de paragraphes (ou de chapitres) couverts par une même CR ([Oberlé, 2016](#) ; [Oberlé et al., 2018](#)) ainsi que par la relation de contiguïté que ces paragraphes entretiennent les uns avec les autres ; » ([Schneidecker, 2019](#)). Dans ce sens, selon la portée des CR, il est possible de distinguer les chaînes globales (ou trans-paragraphiques), à savoir les chaînes qui couvrent plusieurs paragraphes plus ou moins distants, des chaînes locales (ou intra-paragraphiques), c'est-à-dire les CR qui ne dépassent pas le paragraphe ([Schneidecker, 2021](#)). Par ailleurs, [Landragin \(2017\)](#) propose de qualifier la couverture des CR comme *dense* lorsque le nombre de mentions par paragraphe est élevé, *régulière* dans les cas où le référent est mentionné dans chaque paragraphe et *éparse* quand la mention du référent de fait dans certains passages alternant avec d'autres où le référent n'est pas mentionné. Cette distinction ne peut cependant pas se faire sans prendre en compte simultanément le nombre de paragraphes dans laquelle la CR apparaît, la longueur de la CR en nombre de maillons, la contiguïté des paragraphes couverts par la chaîne dans les cas où plusieurs paragraphes sont couverts, on peut alors parler de distance entre les mentions, qui serait calculée en nombre de paragraphes (cf. 2.2.3).

La prise en compte des paragraphes dans le calcul peut toutefois être complexe à mettre en œuvre dans certains cas, comme le souligne [Schneidecker \(2021\)](#) à propos des textes monolithiques. Nous ajoutons à cela la difficulté posée par l'absence d'annotation de la structure textuelle dans certains corpus, ce qui empêche d'automatiser le calcul du nombre de paragraphes couverts par une chaîne de référence. Il serait alors nécessaire d'ajouter une annotation de la structure du texte et son découpage en paragraphes en amont de l'analyse. Toujours est-il qu'une telle annotation n'est pas

toujours envisageable, notamment lorsqu’il s’agit de traiter un volume important de données. À cela s’ajoute la difficulté liée à la variation de la notion de paragraphe selon les types de texte.

Pour pallier cette difficulté et obtenir les informations sur la portion de texte couverte par la chaîne, nous proposons de prendre en compte le nombre de mots et le nombre de phrases. Afin de distinguer cette approche du mode de calcul de la portée, nous employons le terme de *couverture* des CR.

Couverture des chaînes de référence La couverture est ici entendue comme le nombre de mots ou de phrases couverts par une CR. Tout comme la portée, ce critère permet d’analyser l’étendue des chaînes de références au sein des textes. On peut alors différencier les chaînes *locales*, qui couvrent une portion limitée de texte, des chaînes *globales*, qui au contraire couvrent une plus large portion de texte.

Calculée en nombre de mots, la couverture correspond au pourcentage obtenu en divisant le nombre de mots constituant la chaîne par le nombre total de mots du texte. Nous retrouvons ce mode de calcul dans les données mises à disposition par [Wilkens, Oberle, Landragin, et Todirascu \(2020\)](#)². Nous proposons de déterminer le nombre de mots de la chaîne en comptant tous les mots depuis le premier mot de la première phrase contenant le premier maillon, jusqu’au dernier mot de la phrase contenant le dernier maillon. Les mots formant les maillons sont inclus dans le décompte. La ponctuation est exclue du calcul.

Le calcul en nombre de phrases correspond au pourcentage obtenu en divisant le nombre de phrases couvertes par la chaîne par le nombre total de phrases du texte. Le nombre de phrases couvertes par la chaîne prend en compte le nombre de phrases qui se trouvent à l’intérieur de la CR entre le premier et le dernier maillon. Le décompte commence dès le premier maillon, peu importe que le maillon corresponde ou pas au premier mot de la phrase. Le décompte s’arrête au dernier maillon, donc la phrase où il apparaît est prise en compte.

Il est intéressant de souligner que l’interprétation de la portée et de la couverture des CR doit se faire au regard de la longueur : si la chaîne couvre une importante partie de texte et qu’elle est constituée d’un grand nombre de maillons, il est probable que la chaîne porte sur un référent central. Inversement, si une chaîne couvre une très grande portion de texte, mais qu’elle est constituée d’un faible nombre de maillons, la chaîne peut difficilement être interprétée comme principale. Un autre critère qui peut aider à la bonne interprétation de la longueur, de la portée et de la couverture des CR est la distance entre les maillons.

2.2.3 Distance entre les maillons

La distance entre les maillons, appelée aussi *interdistance* ou *distance inter—maillonnaire* par certains auteurs ([Landragin et al., 2024](#) ; [Rousier-Vercruyssen & Landragin, 2019](#)), permet d’analyser la répartition des maillons au sein d’une CR,

2. Page GitHub mettant à disposition les données : https://github.com/boberle/coreference_databases

autrement dit, comment ils sont espacés dans la CR. Ce critère peut révéler l'aspect textuel des chaînes. En effet, la distance entre les maillons est un indice de la régularité d'apparition des maillons dans le texte. Cela signifie que plus les maillons sont proches entre eux, plus il y a cohésion au sein même de la chaîne. Inversement, plus les maillons sont espacés, moins il y a de cohésion dans la CR. De ce fait, nous pouvons faire l'hypothèse qu'elle participe en moindre mesure à la création des liens de cohésion et à la texture. Si la distance entre les maillons est vraiment très importante, c'est la notion même de chaîne qui peut être remise en question. En effet, on peut se demander si le référent est vraiment encore actif dans la mémoire du lecteur et accessible, ce dont nous pouvons douter en nous appuyant sur la théorie de l'accessibilité d'Ariel (cf. 1.3.1). Prenons par exemple la chaîne dans (6), où 149 phrases séparent le deuxième et le troisième maillon. Avec une si grande distance, le référent est difficilement accessible. Cette observation se trouve également étayée par la forme du dernier maillon : l'emploi d'un syntagme nominal complet indique la nécessité de réactiver le référent (Ariel, 2001). Dès lors, la question qui se pose concerne les frontières des CR (Schnedecker, 2021) : on pourrait envisager qu'une nouvelle chaîne redémarre à partir du troisième maillon. Si tel était le cas, la chaîne initiale disparaîtrait, laissant place à deux expressions référentielles distinctes et un singleton.

- (6) On peut penser, parler et rire sans faire mille mouvemens, et, pour la clignoterie, je voudrais que, — puisque tu as le malheur d'avoir **de mauvais yeux**, et que tu ne peux voir et juger qu'en faisant un mouvement disgracieux, — tu renonçasses un peu à la curiosité, que tu te traitasses en aveugle, que tu attendisses que les objets vinssent à toi, et qu'en attendant tu pensasses et parlasses de ce que tu connais.

[5 phrases]

Et puis, déjà elle devait avoir cet esprit de répartie, cette saveur d'originalité, cette naïveté dans les jugements que ses amis admirèrent en elle plus tard ; déjà son goût pour les fleurs et la musique devait amener des sourires sur ses lèvres, des éclairs dans **ses yeux**

[149 phrases]

Malgré ce que dit M. de Constant sur **les mauvais yeux de Rosalie**, celle-ci y voyait encore assez clair pour une jeune fille de quatorze ans et demi. [B_{RdC} ; 

Cet exemple montre que la distance entre les maillons peut expliquer le choix d'une forme d'expressions référentielles plutôt qu'une autre forme.

Pour résumer, cette mesure se révèle utile pour comprendre la composition des chaînes ainsi que leur comportement textuel, surtout lorsque son interprétation est combinée avec la longueur, la portée ou la couverture et la densité de la chaîne. En effet, si la chaîne est composée d'un faible nombre de maillons, mais qu'elle couvre une grande partie du texte en nombre de mots ou de phrases, la chaîne peut difficilement être interprétée comme *principale / importante*. De plus, si les maillons sont très espacés, le risque de compétition est plus élevé, d'autres référents étant présents dans

le texte. Au contraire, si la chaîne couvre une importante partie de texte, est constituée d'un grand nombre de maillons, s'étend sur plusieurs paragraphes, et présente une distance relativement faible entre les maillons, il est probable que la chaîne porte sur un référent central.

Comme la longueur, ce critère est appréhendé de manière différente selon les auteurs. Tous les calculs sont pertinents et permettent de capter des nuances des différents aspects des CR. Les différentes modalités de calcul sont les suivantes :

- en nombre de mots (Kibrik, 2011 ; Rousier-Vercruyssen & Landragin, 2019) : il s'agit de calculer le nombre de mots qui séparent deux expressions référentielles. Ce mode de calcul peut varier dans les détails - prise en compte de tous les mots ou seulement des têtes lexicales, par exemple.
- en nombre de phrases : il s'agit de calculer le nombre de phrases qui séparent deux maillons. Généralement, si deux maillons sont dans la même phrase, la distance est égale à 0. Cette modalité de calcul est utilisée notamment en traitement automatique. Kunz (2010) propose une échelle de distance. La distance est calculée en nombre de phrases entre deux maillons. (*same* au sein de la même phrase, *next* pour un lien inter-phrastique contigu, *several* pour un intervalle entre 1 et 4 phrases, *text* pour un intervalle supérieur à 4 phrases) (Capin, Glikman, Schne-decker, & Todira, 2021).
- en nombre de paragraphes (Ng & Cardie, 2002) : comme pour les phrases, il s'agit de calculer le nombre de paragraphes qui séparent deux maillons d'une même chaîne ;

« La mesure de la distance intermaillonnaire donne une indication sur l'importance, pour chaque texte, des chaînes inter-paragraphiques et de leur empan mais, là encore, cette mesure n'est pas assez fine pour mesurer leur statut intra-paragraphique (Capin et al., 2021) ».
- en nombre d'expressions référentielles (Wolter, 2010) : il s'agit de calculer le nombre d'expressions référentielles présentes dans le texte — qu'elles soient des singletons ou qu'elles appartiennent à d'autres chaînes — entre deux maillons d'une même chaîne ;
- en nombre de syntagmes nominaux ;
- en nombre de caractères.
- en nombre d'unités élémentaires du discours : Kibrik (2011) suggère que cette distance peut être calculée soit en suivant une structure linéaire (*linear distance*), soit en tenant compte de la structure hiérarchique, cette dernière correspondant à la notion de *rhetorical distance*.

Les différentes variantes que nous venons de présenter ont leurs avantages et inconvénients, comme le souligne Schne-decker (2021) :

« Chacune de ces mesures présente ses avantages et ses inconvénients, étant entendu que plus les unités considérées sont construites syntaxiquement comme la phrase ou le SN, plus elles prêtent à discussion » (Schne-decker, 2021, 91).

La distance entre les maillons est généralement appréhendée de manière globale et les études donnent comme information la distance moyenne entre les maillons. Appréhendée de manière globale, ce critère peut permettre d'analyser la *tête* de la chaîne : plus

les maillons sont distants entre eux, plus la chaîne est *éparpillée*, comme dans (6), vice-versa, plus faible est la distance entre les maillons plus la chaîne est *compacte*, comme dans (1). Si l'on appréhende cette information de manière linéaire, en combinaison avec d'autres critères, cette distance pourrait aider à comprendre en partie le choix des expressions référentielles. Dans les travaux en TAL, la distance entre les maillons est utilisée comme contrainte dans les systèmes de résolution automatique (Grobol, 2020 ; Ng & Cardie, 2002 ; Soon, Ng, & Lim, 2001). Elle sert d'aide à la prédiction de l'antécédent d'un maillon donné. Si deux expressions référentielles apparaissent dans la même phrase, la probabilité qu'elles soient coréférentes est plus élevée. Cette contrainte, combinée avec d'autres (tels l'accord en genre et nombre, la fonction syntaxique) permet de classer par ordre de probabilité quels sont les antécédents les plus probables. Par ailleurs, la contrainte de distance entre les maillons permet de trancher les cas où plusieurs antécédents seraient disponibles. On choisirait alors l'antécédent le plus proche de l'anaphorique, comme proposé par Soon et al. (2001).

2.2.4 Persistance

Un critère qui suit naturellement la distance est la persistance. À l'origine, la notion de persistance a été proposée et utilisée par Givón (1983) pour classer les topiques du discours. Il parle de *persistance des topiques*. La notion de persistance a ensuite été reprise par Schnedecker (2021, 91), d'après qui cette notion peut s'appliquer directement aux CR, permettant de calculer l'importance d'un référent. Selon Givón (1983), un référent qui est réinstancié immédiatement dans la phrase successive, est *persistant*. Givón (1983) propose une classification du niveau du référent en fonction de la persistance : un topique est persistant s'il est repris dans la phrase immédiatement successive à la phrase en cours de traitement. Si le référent n'est pas réinstancié, on parle de « persistance minimale » ou « désintégration immédiate ». Nous pouvons faire le lien avec l'activation du référent et son accessibilité : si le référent est persistant, alors il reste activé et accessible dans la mémoire des locuteurs. Ainsi, combiné avec la longueur de la chaîne, sa portée ou couverture, ce critère peut permettre de hiérarchiser les référents selon leur importance. De plus, la persistance peut fournir des indications quant à la cohésion, et peut permettre de comparer aussi l'importance relative de deux référents dans un même paragraphe.

La persistance se mesure en nombre de phrases successives pendant lesquelles le référent reste activé avant de ne plus être mentionné. Pour comprendre, la figure 2.1 illustre la différence entre la distance et la persistance. La distance est considérée comme le nombre de phrases qui séparent une mention dans une phrase donnée (entourée par un cercle rouge dans la figure 2.1) et le maillon précédent. En d'autres termes, la distance correspond au nombre de phrases sans mention du référent qui séparent deux mentions d'un même référent. Inversement, la persistance correspond au nombre de phrases consécutives dans lesquelles le référent est mentionné sans interruption. Plus le score obtenu est élevé, plus il y a de continuité référentielle. Toutefois, des questions se posent pour l'opérationnalisation de ce calcul pour les CR. Calculer la persistance du premier maillon de la chaîne ne pose pas trop de problème. Mais, s'il existe une distance d'une phrase entre deux mentions, nous pouvons nous demander comment continuer le calcul. Doit-on considérer que le calcul s'arrête même si la

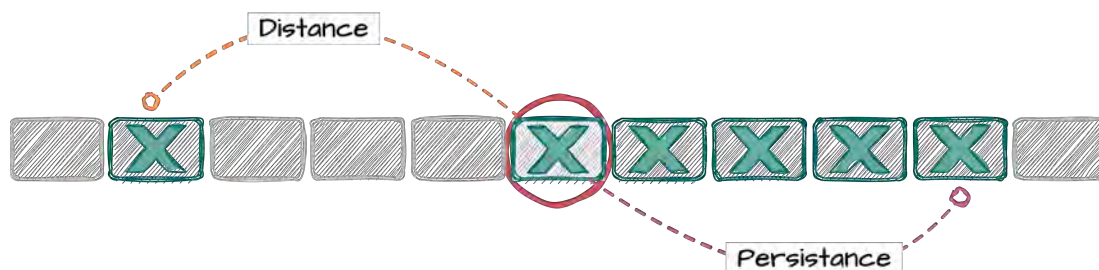


FIGURE 2.1 – Représentation de la distance et de la persistance dans les CR. Chaque case représente une phrase. Les cases en gris illustrent des phrases sans mention du référent.

chaîne continue ? Doit-on redémarrer le décompte des phrases successives à partir de la mention qui se trouve à une phrase de distance et ainsi de suite pour toute la chaîne et en proposer une moyenne, ou alors relativiser sur le nombre de phrases ?

2.2.5 Densité référentielle des chaînes

Nous parlons ici de la densité au niveau d'une même chaîne, ce qui n'est pas à confondre avec le calcul de la densité référentielle à l'échelle du texte. Au niveau du texte, [Schneidecker \(2021\)](#), propose de mesurer le ratio entre le nombre des ER et le nombre de mots du texte. [Landragin \(2017\)](#) parle de *densité référentielle*, entendue comme le ratio entre le nombre de mentions référentielles et le nombre de mots d'un paragraphe. Cela permet de rendre compte de la densité référentielle locale.

Nous proposons d'appliquer un calcul similaire, mais à l'échelle de la chaîne elle-même. Ainsi, il s'agirait de calculer le ratio entre le nombre de mentions de la chaîne et le nombre de mots, de phrases ou de paragraphes couverts par la chaîne. Cette mesure donnerait alors des indications sur la densité la chaîne, permettant de distinguer les chaînes *denses* de chaînes *éparpillées*. Ce qui pourrait venir compléter les observations sur la régularité des maillons et la portée des CR.

2.2.6 Bilan sur les critères de description

Dans cette première partie, nous avons présenté les critères quantitatifs permettant de décrire la *dimension* des CR. Nous avons vu que les modalités de calcul peuvent varier selon les auteurs, ce qui ne facilite pas la comparaison des résultats obtenus.

À plusieurs reprises, nous avons insisté sur le fait que l'interprétation des différents critères, pour être vraiment complète, devrait se faire à l'égard des autres critères et pas de manière indépendante. Les critères que nous avons présentés sont en effet complémentaires. Par exemple, nous avons vu que la prise en compte de la portée de la chaîne permet d'avoir une visibilité sur la portion de texte que la CR couvre, pouvant ainsi compléter l'interprétation et la distinction entre CR *principales* et *secondaires*.

Mais ce critère n'est pas suffisant, à cela va s'ajouter la distance entre les maillons. La prise en compte de ce critère permet de différencier les chaînes dont la longueur est identique ou similaire, mais dont la répartition des maillons dans les textes n'est pas la même.

La prise en compte de ces critères permet de proposer une première description de l'anatomie des CR et d'observer les variations de la composition et de la dimension des CR en fonction de différents facteurs de variation.

2.3 Facteurs de variation des chaînes de référence

Des nombreuses études ont montré que différents facteurs peuvent influencer la manière dont la référence est exprimée, ayant ainsi un impact sur la dimension et la composition des CR. Suivant ces travaux, les caractéristiques des chaînes de référence peuvent varier en fonction de l'époque (Glikman, Guillot-Barbance, & Obry, 2014 ; Landragin et al., 2024), la langue (Baumer, 2019 ; Kunz, 2010), du mode ou du registre (Kunz & Lapshinova-Koltunski, 2015), du type de texte (narratif ou non narratif), du genre textuel (Longo, 2013 ; Schnedecker, 2017 ; Schnedecker & Longo, 2012) ou de la nature sémantique du référent (Fossard et al., 2018 ; Longo & Todirascu, 2014). Ces facteurs ne sont pas les seuls à avoir un impact sur les caractéristiques des chaînes de référence. Comme le souligne Schnedecker (2021), des contraintes « internes » peuvent également impacter la composition des chaînes de référence.

« La composition des chaînes de référence peut être très variable. Elle dépend [...] de contraintes externes (famille de langues, époque de composition du texte, genre discursif, ...) ainsi que de contraintes « internes », c'est-à-dire liées au texte lui-même et à la manière dont le locuteur entend présenter les référents » (Schnedecker, 2021, 121)

Dans cette section, sans prétendre à l'exhaustivité, nous présentons quelques travaux ayant mis au jour les variations caractéristiques des chaînes de référence selon différents facteurs. Les descriptions des chaînes que nous allons présenter s'appuient sur les critères que nous avons présentés dans la section précédente, bien que ceux-ci puissent varier selon les études. Cela nous permettra de mettre en évidence différents aspects des chaînes de référence en fonction des facteurs de variation. Enfin, les travaux mentionnés ne relèvent pas uniquement du domaine de la linguistique descriptive et du TAL : nous mobilisons également les recherches en psycholinguistique, qui, comme nous l'avons vu, ont permis de mettre en évidence les stratégies mises en place pour l'introduction et le maintien du référent.

Nous présentons tout d'abord les facteurs de variation que nous analysons dans la thèse, à savoir le type et le genre textuels et la nature du référent. Pour ce qui concerne les autres facteurs, nous en donnons brièvement les grandes lignes et nous renvoyons aux travaux existants.

2.3.1 Type de texte

Les caractéristiques des CR peuvent varier selon le type de texte, que nous simplifions ici en narratif ou non narratif. À l'échelle du texte, la longueur globale et la

densité référentielle influencent directement la structure et la composition des chaînes de référence. Par ailleurs, plusieurs études ont montré que, dans les textes narratifs, les chaînes de référence sont en moyenne plus longues que dans les textes non narratifs (Longo, 2013). Ce phénomène s'explique notamment par la présence de passages descriptifs dans les textes narratifs, qui favorisent le maintien du référent et aboutissent à des chaînes de référence plus longues, particulièrement pour le référent principal (Schneidecker, 2021). Les variations observées en fonction du type de texte concernent également la composition des chaînes. D'après Salles (2015b), le fait que dans les textes non narratifs la reprise par un pronom soit plus rare s'expliquerait en raison de la distance entre les deux mentions du référent. Plus les maillons sont éloignés, plus le risque de confusion augmente. On retrouve ici la vision d'Ariel (2001) qui considère la distance entre les expressions référentielles comme l'un des facteurs influençant le choix d'expressions dites de haute, moyenne ou faible accessibilité.

Guo (2021) montre comment les CR des textes encyclopédiques soient différentes par rapport à celles des textes narratifs et de celles d'autres textes informatifs. Ils mettent en évidence l'impact du découpage en paragraphes et de la présence des sous-titres, paragraphes mono ou pluri-référentiels ayant une influence sur les CR et leur cohabitation dans le texte.

2.3.2 Genre textuel

Un des facteurs de variation les plus étudiés est le genre textuel. Que ce soit en se concentrant sur un seul genre ou en comparant plusieurs genres, les travaux existants ont permis de mettre en lumière l'impact du genre (parfois en lien avec d'autres facteurs) sur les caractéristiques des chaînes. Ces travaux montrent aussi bien comment les caractéristiques des CR varient en fonction des différents genres textuels et comment s'intéresser aux CR peut aider à l'identification des genres textuels (Schneidecker, 2017). Certaines études ont montré que la longueur des CR varie en fonction du genre textuel, par exemple dans les recettes de cuisine les chaînes de référence sont en moyenne plus courtes par rapport aux introductions des textes encyclopédiques, sans que le référent soit pour autant moins central (Schneidecker, 2014). Il peut aussi arriver que deux chaînes de la même longueur ne correspondent pas forcément à un référent central ou qu'elles présentent différentes configurations en termes de couverture et régularité des maillons.

L'étude de Longo et Todirascu (2010) s'intéresse, elle aussi, aux variations des CR selon le genre. L'étude s'inscrit dans une visée TAL, et son objectif est de proposer un système de résolution automatique qui prenne en compte les variations selon le genre. Pour ce faire, Longo et Todirascu (2010) exploitent un corpus multigenres. Les autrices montrent des variations au niveau de la longueur des chaînes : les CR dans les lois européennes (corpus *Acquis Communautaire*) sont trois fois plus courtes que les CR des textes narratifs, comme les *Trois Mousquetaires* ; selon Longo et Todirascu (2010), cette différence est due au fait que les lois européennes font intervenir plusieurs référents, ce qui crée une compétition référentielle. À ce propos, elles adoptent le point de vue de Schneidecker (2005) et considèrent qu'« en cas de compétition référentielle, la redénomination d'un nom propre est considérée comme une fermeture du référent en cours (donc une ouverture d'une nouvelle chaîne) », d'où des chaînes plus courtes.

En revanche, dans les textes narratifs, on trouve plusieurs passages descriptifs, qui favorisent le développement de CR longues.

Toujours dans cette étude, [Longo et Todirascu \(2010\)](#) montrent que la distance entre les maillons varie également selon le genre textuel. La distance moyenne, calculée dans cette étude en nombre de phrases, est plus courte pour les CR dans *Le Monde* et plus importante pour les CR annotées dans le corpus *Documentation française*. Une autre variation concerne la nature des maillons. Les auteures montrent que dans les textes de *Le Monde* la nature la plus fréquente est le Np (nom propre) alors que *Le Monde diplomatique* favorise les SN définis. Dans l'*Acquis communautaire* les maillons sont le plus fréquemment réalisés par des SN indéfinis. En se focalisant sur la nature du premier maillon, Longo montre dans cette étude que les catégories privilégiées suivent la même tendance : Np dans *Le Monde*, SN défini dans *Le Monde diplomatique* et SN indéfinis dans l'*Acquis communautaire*.

Par ailleurs, un autre critère pertinent analysé par cette même étude est la correspondance entre le premier maillon d'une CR et le thème phrastique. [Longo et Todirascu \(2010\)](#) montrent que dans le *Monde*, dans 80% des cas le premier maillon coïncide avec le thème phrastique ; ce pourcentage descend à 40% dans les rapports de la *Documentation française*.

Dans une perspective complémentaire, une étude visant à évaluer l'impact des chaînes de référence sur la difficulté et la lisibilité des textes a été menée par [Todi-rascu et al. \(2017\)](#). Les auteurs comparent les caractéristiques des CR annotées dans des articles encyclopédiques tirés de Wikipédia et des CR annotées dans des articles encyclopédiques tirés de Wikidia, cette dernière visant un public de jeunes lecteurs. Les auteurs montrent que dans les textes Wikipédia les CR sont en moyenne légèrement plus longues que dans les textes Wikidia. Concernant la nature des maillons, les textes Wikipédia privilégient les SN indéfinis.

Enfin, [Salles \(2015b\)](#) montre, elle aussi, que le genre textuel est un facteur pouvant influencer le choix des expressions référentielles, comme en témoignent les nombreuses études décrivant la variation des CR et des maillons en fonction du genre. Elle souligne aussi que la nature ontologique du référent (humain ou non humain) peut influencer le choix des formes de reprise. C'est à ce facteur que nous nous intéressons dans la section suivante.

2.3.3 Nature du référent

Un facteur qui peut avoir un impact sur la composition des chaînes de référence est la *nature* du référent. Nous proposons le terme de *nature* du référent pour couvrir plusieurs aspects : le degré de concrétude, c'est-à-dire si le référent est une entité concrète (tel un objet) ou une entité abstraite (comme une idée ou un concept) ; l'animacité des référents, en faisant une distinction entre les référents humains et non humains.

D'après [Salles \(2015b\)](#), la nature ontologique des référents et leur topicalité influenceraient la présence ou l'absence de la reprise pronominale. Elle affirme en effet qu'« un référent non humain occupe plus difficilement seul la scène qu'un référent humain et autorise ainsi moins de reprises pronominales ». Cette observation se recoupe

avec celles de [Longo et Todirascu \(2014\)](#), qui comparent les chaînes à référent humain et non-humain dans des textes relevant du même genre textuel et montrent que, en fonction du caractère humain ou non humain du référent, la reprise se fait différemment sans qu'il y ait variation de genre textuel. [Schneidecker \(2014\)](#) a également montré que les formes pronominales sont moins utilisées avec les référents inanimés, tels les référents instanciés dans les recettes de cuisine et les textes encyclopédiques consacrés à des pays. En revanche, les pronoms sont plus fréquents si les référents sont humains, comme c'est le cas des CR dans les portraits journalistiques ([Schneidecker, 2005](#)) ou dans les faits divers ([Schneidecker & Longo, 2012](#)).

Des études en psycholinguistique ont montré que le type de référent dont on parle peut avoir un impact sur le choix des expressions référentielles. Selon le protocole expérimental, les études se sont focalisées sur le choix des expressions référentielles selon que le référent est animé ou inanimé. Il s'agit de facteurs qui peuvent être plus facilement contrôlés pour les expérimentations en psycholinguistique. Les travaux qui se sont intéressés à ce facteur ont montré que la nature des référents peut avoir un impact aussi bien sur les choix du référent que l'on choisit de mettre en avant (saillance) que sur les choix des types des expressions référentielles que l'on utilise pour mentionner le référent en question ([Vogels, Krahmer, & Maes, 2013](#)).

2.3.4 Cohabitation des chaînes de référence

Un autre facteur qui peut avoir un impact sur la composition des CR d'un texte concerne les modalités de **cohabitation des CR**. Il s'agit d'un aspect qui permet d'analyser comment les CR s'articulent dans un texte. Pour appréhender les différentes modalités, on s'intéresse à la répartition des mentions des CR les unes par rapport aux autres, c'est-à-dire que l'on s'intéresse au comment les maillons appartenant à des chaînes différentes s'enchaînent ou coexistent dans le texte ([Schneidecker, 2021](#), 109).

« Certaines chaînes de référence se créent et se développent en toute autonomie, c'est-à-dire sans que leurs maillons entretiennent de relations avec d'autres chaînes de référence ; par contraste, le démarrage et/ou le déroulement de certaines chaînes de référence s'effectue dans la dépendance de maillons d'autres chaînes. » ([Schneidecker, 2021](#), 99)

La cohabitation des chaînes de référence est un critère qui permet d'étudier le déroulement des CR dans le texte. La modalité de coexistence dans le texte peut avoir un impact sur la composition des CR. [Schneidecker et Landragin \(2014\)](#) proposent de classer les modes de cohabitation en distinguant : a) la succession, b) l'entrecroisement, c) la dérivation, d) la partition et la fusion, e) le déroulement en parallèle.

- a) la succession : elle se réalise quand une nouvelle chaîne apparaît dans le texte et la chaîne précédemment active disparaît. Dans l'exemple (7) ci-dessous (emprunté à [Longo \(2013, 69\)](#)), nous signalons en gras les expressions appartenant à la première chaîne. La deuxième chaîne, dont les expressions sont encadrées en jaune, s'ouvre quand la première se ferme. Il en va de même pour la troisième chaîne qui est simplement encadrée.

- (7) **Un professeur d'histoire** vit une vie de famille sans histoire, entouré de l'affection des **siens** et aveugle aux problèmes politiques et

sociaux engendrés par l'"apartheid", régime ségrégationniste qui sévit en Union sud-africaine. **Cet homme paisible** a **un jardinier noir**, paisible **lui** aussi, et soumis à la fatalité. **Ce jardinier** a **un fils** et **ce fils** participe à une manifestation d'écologistes violemment réprimée par la police. On apprendra que **l'enfant** est arrêté, puis qu'il est mort.

On retrouve ici l'idée de la transition par rétention de la théorie du centrage, avec un changement de topique qui détermine l'ouverture d'une nouvelle CR.

- b) l'entrecroisement : deux référents ou plus sont activés simultanément ou l'un après l'autre, et leur fonction syntaxique dans la phrase est modifiée. Comme le montre l'exemple (8) (emprunté à (Longo, 2013, p. 69)), l'expression référentielle de la première CR (en gras dans l'exemple), ayant fonction de sujet, passe à la fonction d'objet direct dès que la deuxième CR (encadrée) s'ouvre.

- (8) Cependant **Milady**, ivre de colère, rugissant sur le pont du bâtiment, comme une lionne qu'on embarque, avait été tentée de **se** jeter à la mer pour regagner la côte, car **elle** ne pouvait se faire à l'idée qu'**elle** avait été insultée par d'Artagnan, menacée par Athos, et qu'**elle** quittait la France sans **se** venger d'eux. Bientôt, cette idée était devenue pour **elle** tellement insupportable, qu'au risque de ce qui pouvait arriver de terrible pour **elle-même**, **elle** avait supplié **le capitaine** de **la** jeter sur la côte; mais **le capitaine**, pressé d'échapper à **sa** fausse position, placé entre les croiseurs français et anglais, comme la chauve-souris entre les rats et les oiseaux, avait grande hâte de regagner l'Angleterre [...]

Contrairement à l'exemple (7), l'ouverture de la nouvelle CR ne coïncide pas avec la fermeture de la précédente et les deux référents continuent d'être mentionnés, mais l'attention du lecteur est attirée sur les actions du deuxième référent, et non plus sur le premier.

- c) la dérivation : le nouveau référent n'est pas activé via un nom propre ou un syntagme nominal, mais dépend d'une autre chaîne, c'est-à-dire que l'interprétation du référent est faite en activant nos connaissances du monde ou par inférence. Dans l'exemple qui suit (9) (emprunté à Schnedeker et Landragin (2014)) le référent est activé via une anaphore associative. *L'homme au volant* peut être interprété parce que dans l'énoncé précédent on présente le fait qu'il y a eu un accident. Nous avons signalé en gras et encadré les deux référents qui dépendent de la mention d'*accident*.

- (9) Jack est ingénieur du son. Un soir, alors qu'il se promène près d'une rivière pour enregistrer les bruits de la nature, il assiste à **un accident**. Il sauve **la passagère**, mais **l'homme au volant** est tué. Il était candidat à l'investiture de **son** parti pour la magistrature suprême. Il faut absolument taire le fait qu'il voyageait en compagnie d'**une femme, aussi publique que lui**.

- d) la partition et la fusion : cette catégorie est une sous-catégorie de la dérivation. La partition consiste à extraire deux référents individués d'un ensemble précédemment posé (c'est le cas des reprises pronominales d'un SN collectif par *l'un / l'autre*). Inversement, la fusion consiste à ressembler des référents précédemment posés isolément. Dans l'exemple (10) (emprunté à [Schneidecker et Landragin \(2014\)](#)), nous encadrons le référent collectif et signalons en gras les référents posés isolément.

- (10) **Deux hommes** parurent. **L'un** venait de la Bastille, **l'autre** du Jardin des Plantes. **Le plus grand**, vêtu de toile, marchait le chapeau en arrière, le gilet déboutonné et sa cravate à la main. **Le plus petit**, dont le corps disparaissait dans une redingote marron baissait la tête sous une casquette à visière jaune. Quand **ils** furent arrivés au milieu du boulevard, ils s'assirent à la même minute, sur le même banc.

- e) le déroulement en parallèle : deux référents, l'un aussi saillant que l'autre, sont posés et repris en parallèle.

- (11) Pour comprendre, il faut d'abord comparer. **DSK** est un prof d'université. **Fabius** est un énarque normalien. **L'un** est rond et séducteur. **L'autre** est sec et élégant. **Strauss-Kahn** est brouillon. **Fabius** est précis. **Le premier** a de l'humour. **Le second**, de l'ironie. **DSK** ne se méfie de personne. **Fabius** se méfie de tout le monde. **DSK** pense – à tort – n'avoir que des amis. **Fabius** croit – à tort – n'avoir que des ennemis.

On peut remarquer que dans le déroulement en parallèle la fonction syntaxique des deux référents est la fonction sujet, ce qui pourrait confirmer le rôle de la fonction syntaxique comme indicateur de la saillance des référents.

2.3.5 Autres facteurs de variation

D'autres facteurs, externes et internes, peuvent avoir un impact sur la dimension et sur la composition des chaînes de référence. Nous en proposons ci-dessous un aperçu succinct. Ces aspects n'étant pas pris en compte dans le cadre de cette thèse, nous renvoyons donc aux travaux existants pour une présentation détaillée des résultats obtenus³.

Époque En premier lieu, nous pouvons mentionner l'époque, en tant que facteur externe ayant un impact sur la dimension et la composition des CR. Différentes études se sont intéressées aux chaînes de référence en analysant leurs caractéristiques aussi bien en synchronie ([Glikman et al., 2014](#)) qu'en diachronie ([Capin et al., 2021](#) ; [Landragin et al., 2024](#) ; [Obry et al., 2017](#)).

[Glikman et al. \(2014\)](#) et [Obry et al. \(2017\)](#) proposent une analyse diachronique des CR à référent humain dans les récits brefs. Les résultats de ces deux études montrent

3. Pour aller plus loin, nous renvoyons également à l'ouvrage de [Schneidecker \(2021\)](#) qui recense plusieurs études portant sur ces facteurs et en présente les principaux résultats obtenus.

que la composition des CR varie selon les évolutions de la langue : les auteurs mettent notamment en évidence la baisse de fréquence du sujet zéro, une évolution de l'article indéfini au profit du nom propre, et l'apparition de nouvelles formes référentielles (ex. : *ledit, lequel*). Les auteurs soulignent également l'individualisation croissante des personnages principaux, ce qui influence la composition des chaînes (Obry et al., 2017). Récemment, une étude à plus large échelle en diachronie longue a été proposée par Landragin et al. (2024), à partir du corpus Democrat (Landragin, 2016) (cf. chapitre 3), prenant en considération des textes narratifs et non narratifs, issus de plusieurs genres textuels. Les auteurs proposent de prendre en considération la nature des maillons, et comme dans l'étude de Obry et al. (2017) d'analyser la nature du premier maillon, la distance entre les maillons (calculée en nombre de mots), la longueur des chaînes (en nombre de maillons et nombre de mots), et la stabilité référentielle. Les résultats obtenus confirment les observations des deux études précédemment présentées, notamment pour les sujets zéro. Les auteurs mettent également en évidence une augmentation de la distance entre les maillons pour les textes de certaines périodes (14^e, 17^e, 20^e, 21^e siècles) et l'augmentation progressive des SN définis et indéfinis au fil du temps. La stabilité référentielle reste globalement constante, à l'exception des CR dans les textes du 21^e siècle, où le coefficient de stabilité est plus élevé, en particulier pour les textes non-narratifs. Encore une fois, d'autres facteurs, tels le type et le genre textuel, semblent entrer en jeu, en plus de l'époque, comme déjà mis en évidence par Aktaş et Stede (2022).

Langue Un autre facteur externe ayant un impact sur la variation des CR est la langue. Les études contrastives se sont intéressées aussi bien aux phénomènes d'anaphore et coréférence qu'aux caractéristiques des chaînes (Kunz et al., 2016 ; Lundquist, 2005). Baumer (2012) par exemple, analyse l'emploi des anaphores nominales en français et en anglais. Ses résultats mettent en évidence que ce type d'anaphores est plus utilisé en français qu'en anglais. D'autres études, telles que celles de Kunz (2010) ; Lapshinova-Koltunski et Kunz (2020) analysent les chaînes de référence en anglais et en allemand, en prenant en considération aussi d'autres facteurs, tels le mode de communication (oral vs écrit) et le genre textuel. Ces travaux ont pour objectif de comprendre quel est l'impact de ces facteurs sur les caractéristiques des chaînes de référence, et d'analyser dans quelle mesure les variations rencontrées sont liées à la langue elle-même, au mode ou au genre textuel. Les deux études convergent vers une même conclusion : le genre ou la structuration textuelle ont un impact plus significatif que la langue ou le mode de communication sur les caractéristiques des chaînes de référence. Guo (2021) propose une analyse contrastive des CR dans un corpus comparable de textes encyclopédiques en français et en chinois. Il observe que les chaînes principales sont en moyenne plus longues en chinois, avec une prédominance de sujets zéro, tandis que le français privilégie les pronoms. Les syntagmes nominaux dominent dans les deux langues, mais le chinois présente une plus grande diversité lexicale. La fonction sujet est majoritaire dans toutes les CR, particulièrement dans les chaînes principales, surtout en chinois. Enfin, les modalités de cohabitation des chaînes sont globalement similaires dans les deux langues, avec des différences qui semblent davantage liées à la nature des paragraphes qu'à la langue elle-même.

Pour résumer, les études contrastives ont mis en évidence qu'il existe des variations

dans les caractéristiques des CR liées à la langue. Cependant, les résultats suggèrent que d'autres facteurs, tels le genre discursif ou la structure textuelle, ont un impact plus important que la langue. Cela offre des implications importantes pour la linguistique contrastive et le traitement automatique des langues.

Canal ou support de transmission Aktaş et Stede (2022) analysent la façon dont le genre et le support de transmission influencent les stratégies référentielles. Pour ce faire, les auteurs examinent la variation de la distance entre les expressions référentielles à l'oral et à l'écrit, en combinant des analyses de corpus et des données expérimentales. Les résultats indiquent que lorsque la distance entre les expressions référentielles est importante, les participants utilisent davantage de pronoms à l'oral qu'à l'écrit. Cette tendance suggère que la distance entre les expressions référentielles à l'écrit a un impact sur le choix des expressions référentielles, corroborant ainsi les conclusions des études de corpus antérieures. Amoia et al. (2012) proposent une analyse empirique des chaînes de référence dans des textes oraux et écrits, en comparant des interviews (oral) et des textes de vulgarisation scientifique (écrit). Les résultats mettent en évidence que les interviews présentent des chaînes de référence plus courtes et une utilisation plus fréquente de pronoms personnels, tandis que les textes écrits favorisent des chaînes plus longues avec une prédominance de syntagmes nominaux.

Les résultats de ces études offrent une meilleure compréhension des stratégies référentielles mobilisées en fonction du canal de transmission. Par ailleurs, les auteurs suggèrent des pistes d'amélioration pour la résolution automatique de la coréférence, en mettant en évidence les paramètres à adapter aux spécificités de l'oral, de l'écrit, ainsi qu'à la diversité des genres discursifs.

Structure textuelle En plus des facteurs externes que nous venons de présenter, des facteurs internes peuvent avoir un impact sur la composition des chaînes de référence. Les travaux de (Guillot-Barbance & Quignard, 2019), par exemple, mettent en lumière les relations entre les chaînes de référence et la structure textuelle. Leur analyse révèle que les chaînes longues et continues sont souvent associées aux thèmes développés à l'échelle du chapitre, tandis que les chaînes plus courtes et discontinues correspondent aux thèmes spécifiques des paragraphes. Cette distinction suggère que la structuration textuelle influence la configuration des chaînes de référence.

Capin et al. (2021), s'inscrivant dans le cadre théorique de Fox (1987), distinguent deux modes d'interaction entre contexte et usage des expressions référentielles : d'une part, le contexte structurel (par exemple, le paragraphe) conditionne l'utilisation des expressions référentielles ; d'autre part, l'usage des expressions contribue à construire le contexte. Cette dualité offre un éclairage pertinent sur la manière dont les chaînes de référence s'articulent avec la structure textuelle.

Par ailleurs, Guo (2021) propose une typologie des paragraphes en fonction de leur configuration référentielle, distinguant les paragraphes mono-référentiels, centrés sur un seul référent, et les paragraphes pluri-référentiels, impliquant plusieurs référents. Il observe que, dans les environnements pluri-référentiels, les reprises sont plus variables, ce qui complexifie la résolution de la coréférence.

Ces études convergent vers l'idée que la structure textuelle, au niveau du titre, du paragraphe ou de la section, joue un rôle déterminant dans la configuration des chaînes de référence. Ainsi, pour améliorer la résolution automatique de la coréférence, il est essentiel de prendre en compte ces éléments structurels, en adaptant les modèles aux spécificités des différents genres et supports textuels.

2.4 Bilan

Ce chapitre avait pour objectif de dresser un état des lieux des différents critères de description et des facteurs de variation identifiés dans les travaux s'intéressant aux chaînes de référence (Kibrik, 2011 ; Kunz & Steiner, 2012 ; Recasens & Hovy, 2010 ; Rousier-Vercruyssen & Landragin, 2019 ; Schnedecker, 2021). Nous avons fait une distinction entre les critères qui permettent de décrire la composition des chaînes, par la caractérisation des maillons (nature, fonction syntaxique, etc.), et ceux qui permettent de caractériser la dimension des CR (longueur, distance entre les maillons, etc.).

En présentant les aspects des CR que ces critères permettent de mettre en évidence, nous avons identifié des types de chaînes, que nous résumons dans ce qui suit :

- chaînes homogènes vs chaînes hétérogènes : relevé selon la nature des maillons. Nous pouvons faire un lien entre ce type de chaîne et la configuration des chaînes des référents saillants et non saillants (cf. 1.3, Givón (1983)) ;
- chaînes autonomes vs chaînes dépendantes d'autres CR : relevé principalement pas la nature et la fonction syntaxique des maillons ;
- chaînes principales vs chaînes secondaires : relevé selon la longueur de la chaîne, leur portée ou couverture, la distance entre les maillons et la fonction syntaxique des maillons ; ce type de CR peut être lié aux suivants :
 - chaînes locales vs chaînes globales : relevé selon la longueur de la chaîne, leur portée ou couverture ;
 - chaînes longues vs chaînes éphémères : relevé selon la longueur de la chaîne et leur portée. Les chaînes longues pourraient correspondre à des référents principaux avec une couverture globale, et les chaînes éphémères à des référents secondaires, avec une couverture locale.

Il est important de souligner que ces types émergent rarement d'un seul critère isolé : d'après nous, c'est bien la combinaison de plusieurs critères qui permet d'aboutir à une typologie des chaînes. Une approche fondée sur des configurations de critères croisés est donc indispensable pour saisir la complexité de leur fonctionnement discursif.

Par ailleurs, nous avons présenté différents facteurs de variation susceptibles d'influencer la composition et la dimension des chaînes. Ces facteurs sont souvent considérés comme des variables externes, mais ils pourraient également être interprétés comme des indices de types de chaînes. Par exemple, certaines configurations semblent récurrentes dans des genres narratifs, juridiques ou scientifiques, ou encore pour certains types de référents (humains, concrets, abstraits). On pourrait alors parler de chaînes typiques de certains types ou genres textuels, comme des « chaînes narratives » ou des « chaînes juridiques » ; de certains référents, par exemple des CR pour les référents humains ou celles pour les référents abstraits ; de situations discursives, par exemple,

des chaînes en danger (en compétition avec d'autres référents). Néanmoins, à ce stade de notre travail, il nous paraît prématuré de considérer ces facteurs comme les fondements d'une typologie préétablie fondée uniquement sur ces facteurs de variation. Notre hypothèse est inverse : ce sont les types de chaînes que nous allons identifier empiriquement – à partir de leur composition et de leur dimension – qui permettront d'interroger leur distribution selon les genres, les types de textes ou les natures des référents. Autrement dit, notre objectif est d'évaluer dans quelle mesure les types de chaînes que nous identifions empiriquement – via des méthodes de linguistique outillée et dans une approche bottom-up – sont associés à ces facteurs de variation.

Afin de pouvoir décrire les chaînes de référence selon les critères présentés et d'analyser l'impact des facteurs de variation, il est essentiel de disposer de données annotées. Dans le chapitre suivant, nous présentons donc les corpus annotés en chaînes de référence existants.

Chapitre 3

Des corpus annotés en référence

Contents

3.1	En quoi consiste l'annotation de la référence ?	54
3.2	Guides et modèles d'annotation	58
3.3	Deux corpus de référence : AnnoDis et DEMOCRAT	60
3.3.1	Le corpus AnnoDis	61
3.3.2	Le corpus DEMOCRAT	63
3.3.3	Comparaison des modèles d'annotation	64
3.4	Impact du modèle d'annotation sur la description des chaînes de référence	66

Pour décrire les chaînes de référence selon les critères que nous avons présentés dans le chapitre précédent et étudier l'impact des différents facteurs de variation, il est nécessaire d'avoir à sa disposition des données. En fonction de l'approche que l'on suit pour étudier le fonctionnement des expressions référentielles et des chaînes de référence, il est possible de mettre en place des protocoles expérimentaux, ou bien d'utiliser des données attestées, en menant une analyse de corpus annotés en référence (cf. 1.3). Dans le cadre de notre thèse, nous avons opté pour l'exploitation de corpus annotés en chaînes de référence. Actuellement, il existe une quantité considérable de corpus annotés en CR. Les campagnes MUC-6 (Grishman & Sundheim, 1996), MUC-7 (Hirschman & Chinchor, 1998) et SEMEVAL (Recasens, Màrquez, et al., 2010) ont favorisé la création de larges corpus avec des annotations de la référence. La majorité de ces corpus est en anglais (Ghaddar & Langlais, 2016 ; Pradhan, Moschitti, Xue, Uryupina, & Zhang, 2012), mais ces campagnes ont également contribué à la diffusion de corpus annotés dans d'autres langues. La mise à disposition de ces corpus a permis non seulement de développer des systèmes de résolution automatique de la référence, mais aussi de systématiser les études linguistiques et d'accompagner la description des différents phénomènes de la référence et de proposer des études à plus grande échelle. Cependant, ce type d'études reste minoritaire, l'accent étant mis sur le développement des systèmes de résolution.

À partir des recherches sur les données existantes, nous avons constaté que les corpus annotés en chaînes de référence présentent une hétérogénéité, tant en termes de

choix d’annotation que de formats de diffusion. Cette hétérogénéité reflète la diversité des bases théoriques adoptées lors de la constitution des ressources et la diversité des objectifs de recherche pour lesquels chaque ressource est créée (Poesio, Pradhan, Recasens, Rodriguez, & Versley, 2016 ; Recasens & Martí, 2010). Si les ressources annotées en CR ont été constituées pour pouvoir effectuer des descriptions linguistiques fines, les annotations présentent un niveau de granularité assez fin, permettant d’ajouter le plus d’informations possible, comme par exemple dans le cas de la constitution du corpus MC4 (Mélanie-Becquet & Landragin, 2014). Si les ressources sont annotées pour pouvoir développer et entraîner des systèmes de résolution automatique de la référence, les modèles sont moins complexes pour favoriser l’annotation d’une grande quantité de données, comme par exemple dans le cadre du projet Democrat (Landragin, 2016) et des campagnes MUC (van Deemter & Kibble, 2000) et CRAC (Yu et al., 2022).

Étant donné la complexité de la tâche, les annotations se limitent parfois (LDC, 2008 ; Zhu, Pradhan, & Zeldes, 2021) à certains types de référents (par exemple, on n’annote pas les événements ou on n’annote que les référents humains). Une autre limite concerne le type d’anaphore, par exemple les anaphores associatives sont le plus souvent exclues des modèles d’annotation. Cette hétérogénéité est souvent perçue comme une limite à la comparabilité des résultats. Plutôt que de la voir comme une limite, nous interrogeons la possibilité d’en tirer parti. Dans notre thèse, nous exploitons deux ressources diversifiées. Afin de comparer les résultats obtenus à partir de ces deux ressources hétérogènes, il est essentiel de bien connaître leurs spécificités et de prendre les précautions nécessaires lors de la comparaison des résultats. Ce chapitre vise à comprendre les différences entre les modèles d’annotation des corpus, avec un focus sur les corpus exploités dans la thèse. Comprendre la diversité des modèles d’annotation nous permet de mieux appréhender leur impact potentiel sur les résultats obtenus.

Tout d’abord, nous présentons en quoi consiste la tâche d’annotation de la référence. Ensuite, nous présentons les différents corpus annotés en référence et la diversité des modèles d’annotation. Nous mettons en avant les divergences et les similitudes entre les corpus. Cela nous amènera aussi à nous intéresser à l’hétérogénéité des formats de diffusion. Par conséquent, nous discutons de la question de la compatibilité entre les annotations les corpus et les outils d’analyse de corpus.

3.1 En quoi consiste l’annotation de la référence ?

La tâche d’annotation de la référence est chronophage et plutôt complexe, comme le soulignent Mitkov et al. (2000) et Landragin (2021a) :

« Annotating anaphora is a notoriously difficult, timeconsuming and labour-intensive task even when focusing on one single variety of the phenomenon. » (Mitkov et al., 2000)

« Les expressions référentielles et les chaînes de référence parsèment les textes, contribuant à leur ancrage contextuel et à leur cohérence textuelle.

Les annoter pose des difficultés (Landragin & Schnedecker, 2014 ; Schnedecker, Glikman, & Landragin, 2017), qui relèvent d'une part de leur détection — on peut « rater » non seulement un sujet zéro, mais aussi un pronom comme « l' » —, d'autre part de leur renvoi plus ou moins clair vers un référent du monde extralinguistique : l'ambiguïté comme la sous-détermination font partie intégrante de la tâche (Landragin, 2011) » (Landragin, 2021a, 40).

Qu'il s'agisse d'annoter les relations anaphoriques ou les chaînes de référence, nous pouvons distinguer quatre étapes théoriques :

1. identifier les référents et les expressions référentielles (ER) ;
2. délimiter les ER ;
3. attribuer des propriétés aux expressions référentielles annotées ;
4. relier les ER.

Les deux premières étapes sont strictement liées l'une à l'autre, nous les avons distinguées pour mieux mettre en évidence les différents choix qu'elles impliquent.

Identifier les référents et les expressions référentielles

La première étape de l'annotation consiste à identifier dans le texte les référents. Il s'agit notamment de savoir quel type d'entité il faut annoter dans les textes. Par exemple, si l'on s'intéresse uniquement aux référents humains, si on prend en considération les événements, les références temporelles. Certains projets se sont focalisés par exemple sur les entités nommées, telles les organisations, les entités géopolitiques, les personnes, etc. (Poesio et al., 2016).

Concernant les expressions référentielles, il s'agit de savoir quel type d'expressions référentielles il faut annoter en tant que mentions (ou maillon dans le cas des chaînes). Cela va dépendre du type de référent et de la définition de maillon que l'on adopte. Nous avons pu relever trois conceptions plus ou moins strictes. La figure 3.1 résume les types d'expressions référentielles pris en compte par chacune des trois conceptions.

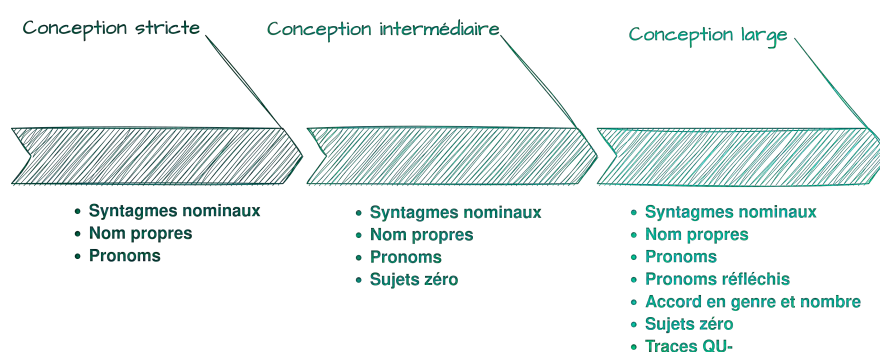


FIGURE 3.1 – Trois conceptions de maillons

- **Une conception stricte** : si l'on se tient aux théories classiques de la référence (Charolles, 2002), la définition du maillon est plutôt stricte : seules les

expressions référentielles ayant une forme linguistique (Schneidecker & Landragin, 2014), telles que les syntagmes nominaux, les noms propres, les pronoms (personnels, démonstratifs, adverbiaux, possessifs, indéfinis, relatifs) sont aptes à constituer des maillons.

- **Une conception large :** elle se base sur l'échelle d'accessibilité d'Ariel (2001) (figure 1.1). Suivant cette échelle les pronoms réfléchis, les accords, les traces QU- (relatives) et les sujets zéro peuvent être considérés comme des expressions référentielles aptes à indiquer une continuité référentielle ou, en termes d'accessibilité, à maintenir accessible un référent. À ces éléments Schneidecker et Landragin (2014) ajoutent les syntagmes nominaux en position initiale, comme par exemple les appositions.
- **Une conception intermédiaire :** Landragin (2011) décide d'ignorer les phénomènes d'accords verbaux et participiaux en raison de leur aspect grammaticalisé. Il estime aussi qu'une parenthèse, une construction attributive ou des syntagmes apportant une information à l'expression référentielle peuvent être considérés comme des maillons (Landragin, 2011, 8).

Une autre différence entre les ressources concerne l'annotation des singletons (référents qui ne font pas l'objet d'une reprise référentielle). Certains considèrent qu'il est indispensable de les annoter pour pouvoir entraîner les systèmes de détection des expressions référentielles ou pour pouvoir mesurer le rapport entre les ER qui constituent des CR et le nombre total d'ER dans le texte. Cela permet de mesurer la densité référentielle des textes (cf. 2.2.5). D'autres, ayant des objectifs différents, ne les annotent pas. De Marneffe, Recasens, et Potts (2015) ont démontré que l'annotation de singletons améliore les performances des systèmes de résolution.

Délimiter les expressions référentielles

La notion de *mention* ou *markable* concerne la partie de texte qui doit être annoté en tant que mention d'une entité. Les décisions en termes de délimitation des maillons, peuvent varier d'une annotation à l'autre. La question qui se pose est de savoir s'il faut considérer le syntagme nominal complet, si l'on doit annoter les compléments comme des maillons à part, si l'on considère que les compléments donnent naissance à plusieurs chaînes ou non. Les décisions peuvent varier aussi d'une langue à l'autre. Dans certains modèles d'annotation pour l'anglais, les SN sont annotés intégralement, comme c'est le cas dans le corpus WikiCoref (Ghaddar & Langlais, 2016). De ce fait, des syntagmes tels que *[Zsa Zsa, who slap a security guard]_{zsazsa}* ou *[a colorful array of magazines]_{array}* sont annotés comme un seul et unique maillon. À l'inverse, dans le corpus OntoNotes (Hovy, Marcus, Palmer, Ramshaw, & Weischedel, 2006) les chercheurs ont annoté les SN en distinguant le syntagme complet (en gras) et la mention du référent (entre crochets), obtenant des annotations telles que *[Zsa Zsa]_{zsazsa}, who slap a security guard* ou *[a colorful array]_{array} of magazines*. Suivant à peu près la même approche, Landragin (2016), dans le modèle d'annotation pour le projet français DEMOCRAT, indique que ce type de SN est susceptible d'ouvrir deux CR différentes, et qu'il faut annoter aussi bien le SN complet que tous les compléments qui pourraient ouvrir une nouvelle CR, par exemple : *J'ai bu [une bouteille d'_{eau}]_b ou [L'intérêt [des écrivains]_{ecrivains}]_{interetE}*.

Attribuer des propriétés aux expressions référentielles

Cette étape consiste à renseigner des propriétés qui donnent des informations sur la mention annotée. Le nombre de propriétés à annoter peut varier d'un projet à l'autre. Par exemple, dans les corpus à visée linguistique, on privilégie des modèles assez complexes quitte à avoir peu de données à disposition. Inversement, les projets visant le développement de systèmes automatiques vont privilégier la quantité sur la richesse des descriptions.

Les propriétés peuvent correspondre à des informations de surface, telles l'identifiant attribué au référent, la catégorie grammaticale de l'expression référentielle, sa fonction syntaxique et les autres propriétés morphosyntaxiques. Néanmoins, ces informations sont souvent confiées à l'annotation automatique. D'autres propriétés correspondent à des informations plus complexes, comme le type d'entité annoté (nature ontologique du référent), le type d'emploi de l'expression référentielle (générique ou spécifique). Cela amène à avoir des modèles d'annotation plus ou moins complexes. Par exemple, le modèle d'annotation du corpus MC4 ([Mélanie-Becquet & Landragin, 2014](#)) prévoit l'attribution de 11 caractéristiques pour un ensemble de 78 valeurs, alors que dans le corpus Democrat ([Landragin, 2016](#)) aucune information n'est attribuée aux maillons (seul l'identifiant du maillon est signalé).

Relier les expressions référentielles

Cette étape consiste à relier entre elles les différentes expressions référentielles. Elle implique un double choix : un premier, d'ordre sémantique, concerne les types de relation entre les maillons qui doivent être pris en compte ; et un deuxième, d'ordre technique, concerne la mise en œuvre pratique de la relation entre les ER.

Concernant les types de relations entre les expressions référentielles, il y a des décisions difficiles à prendre dans plusieurs cas de figure. En effet, la continuité référentielle n'est pas toujours facile à évaluer.

Un premier cas de figure, porte sur le type de relation entre les ER. On peut annoter uniquement les ER qui entretiennent une relation de coréférence stricte (cf. [chapitre 1](#)). Ce type d'annotation est la plus fréquente lorsque l'on étudie les chaînes de référence. Un choix délicat concerne l'annotation des référents évolutifs : le statut du référent change et le référent n'est plus exactement le même après l'évolution, comme illustré dans [\(12\)](#) avec un exemple bien connu (poulet vivant vs poulet rôti). Il s'agit alors de décider si considérer les différentes mentions du référent comme coréférentes (comme dans Democrat) ou pas.

- (12) Tuez un **poulet bien actif et bien gras**. Préparez-**le** pour le four. Coupez-**le** en quatre morceaux et faites-**le** rôti avec du thym pendant une heure. ([Kleiber & Riegel, 2003](#), 126)

Lorsque l'annotation porte sur les CR, les cas de fusion (cf. [2.3.4](#)) peuvent poser problème quant à l'annotation de l'antécédent. Le plus souvent, puisque la relation entre l'antécédent et sa reprise n'est pas de coréférence stricte, l'antécédent n'est pas relié

à la CR (comme dans Democrat). Parfois, les antécédents sont annotés avec une indication spécifique pour signaler la fusion (c’est le cas dans le corpus RésolCo (Garcia-Debanc et al., 2017)). Enfin, on peut annoter des phénomènes anaphoriques sans co-référence stricte, comme les anaphores associatives ou les relations de quasi-identité (Recasens, Hovy, & Martí, 2010) (e.g. *Le village* → *L’Église*).

Sur un plan plus technique, Landragin (2014) souligne que deux approches sont utilisées pour l’annotation des CR. La première consiste à annoter les expressions référentielles comme des unités. Une relation entre deux expressions référentielles peut alors être indiquée. On dispose d’une liste d’unités et de relations que ces unités entretiennent entre elles. Cette solution est utilisée dans les campagnes MUC *Message Understanding Conferences* (van Deemter & Kibble, 2000) ou via des outils tels que MMAX et MMAX2 (Landragin, 2014 ; Müller & Strube, 2006). Une autre solution pour l’annotation des relations est le codage en première mention, utilisé pour l’annotation du corpus AnCor (Muzerelle et al., 2013). Ce type de codage consiste à tisser des relations entre le premier maillon et un maillon-x. Dans les deux cas, il est possible de renseigner des informations sur le type de relation qui existe entre les ER. L’autre façon de faire consiste à considérer les CR comme des schémas qui regroupent les différentes unités (les expressions référentielles). Un schéma regroupe alors toutes les mentions du même référent sans avoir à spécifier le type de relation entre deux unités. Deux approches sont possibles : soit on relie manuellement toutes les mentions à une même chaîne, soit on crée les *schémas* à posteriori, à partir des identifiants attribués aux référents, comme dans le corpus Democrat (Landragin, 2016).

3.2 Guides et modèles d’annotation

Les guides d’annotation et les modèles utilisés vont donc expliquer et refléter les différents choix qui ont été faits pour mener à bien la tâche d’annotation. Tout d’abord, les guides présentent les indications sur la manière dont les annotateurs doivent procéder pour effectuer la tâche d’annotation. Ensuite, le modèle d’annotation est présenté en détail.

Concernant la mise en œuvre de la tâche d’annotation, plusieurs solutions sont envisageables. Pour l’identification des référents, nous pouvons distinguer deux approches principales : une annotation entièrement manuelle, qui consiste à annoter manuellement toutes les expressions référentielles correspondant aux référents ; une annotation manuelle précédée d’un pré-marquage automatique des expressions référentielles qui seraient susceptibles de désigner les référents. Cette étape, étant étroitement liée à la délimitation des expressions référentielles : il faut prévoir comment délimiter les ER, laissant ou pas la possibilité à l’annotateur de modifier les délimitations. L’objectif du pré-marquage est de faciliter et d’accélérer la tâche. Mais le gain de temps est à relativiser : les expérimentations menées par Landragin (2021b), avant le lancement du projet Democrat, ont montré que le pré-marquage nécessite des corrections, ce qui le rend finalement plus coûteux pour les annotateurs.

Parmi les pratiques d’annotation des corpus, nous pouvons distinguer une approche ascendante et une approche descendante. La première, la plus utilisée, consiste à d’abord

annoter toutes les expressions référentielles, en effectuant une lecture linéaire de l'intégralité du texte. Une fois toutes les expressions annotées, elles sont reliées avec des relations ou des schémas. Cette approche ascendante est la plus utilisée (Landragin, 2016 ; Recasens & Martí, 2010 ; Zhu et al., 2021). Une deuxième approche consiste à zoomer sur des parties de texte, où les expressions référentielles ont été pré-marquées automatiquement. L'annotateur se focalise alors sur la portion de texte où les expressions pré-marquées apparaissent, et il complète l'annotation manuellement. Cette approche est utilisée en particulier dans l'annotation du corpus AnnoDis (Colléter et al., 2012).

Une autre différence que nous pouvons mettre en lumière concerne le profil des annotateurs. Certains projets font appel à des annotateurs experts, d'autres projets, en revanche, s'appuient sur des annotateurs non experts ou « naïfs », tels que les étudiants. D'autres font appel à des plateformes de *crowdsourcing*, c'est-à-dire dire que l'on propose la tâche d'annotation à la foule, comme par exemple dans *Phrase Detective* (Chamberlain, Poesio, & Kruschwitz, 2008). La distinction entre annotateurs experts et naïfs n'est pas anodine et peut avoir un impact sur la qualité et sur la nature des annotations des CR. La compréhension de l'annotation fine de ces phénomènes requiert une expertise linguistique approfondie. Ainsi, dans les projets impliquant des annotateurs non experts, les indications d'annotation sont généralement simplifiées pour faciliter la tâche. Par exemple, l'outil CoRefi (Bornstein, Cattan, & Dagan, 2020), conçu pour le crowdsourcing, propose une interface intuitive et un processus de formation guidée pour aider les annotateurs à comprendre les décisions d'annotation. D'autres projets, tels que AnnoDis, ont choisi d'impliquer des annotateurs non experts pour sonder les intuitions des locuteurs sur les indices de continuité référentielle. La définition de chaînes qui a été donnée aux annotateurs était très large, et une grande liberté a été laissée concernant la délimitation des expressions référentielles.

Passons maintenant aux modèles d'annotations. Une différence que nous pouvons identifier concerne la délimitation des ER. Les schémas d'annotation pour la résolution des coréférences, tels que MUC (Grishman & Sundheim, 1996 ; Hirschman & Chinchor, 1998), MATE (Poesio, Bruneseaux, & Romary, 1999), ACE (Doddington et al., 2004), GNOME (Poesio, 2000), et ONTONOTES (Hovy et al., 2006), font généralement annoter les syntagmes nominaux (SN) dans leur intégralité, y compris tous les post-modificateurs. Cette pratique vise à améliorer l'accord entre annotateurs, mais peut poser des défis pour les systèmes automatiques en raison d'éventuelles erreurs de prétraitement, comme des inexactitudes de parsing. D'autres, font annoter uniquement la tête lexicale de l'ER (Doddington et al., 2004 ; Poesio et al., 2016).

D'autres différences concernent le type d'entités annotées, par exemple, le corpus français RésolCo (Garcia-Debanc et al., 2017), ne considère que les référents humains ; le corpus ACE (Doddington et al., 2004), se limite à certains types, telles les personnes, les organisations, les lieux, les entités géopolitiques, les armes et les véhicules (Poesio et al., 2016).

Les modèles diffèrent également sur les formes d'anaphores annotées. Les modèles MUC (Grishman & Sundheim, 1996) n'annotent pas les anaphores sans antécédent nominal (ex. : événements, deixis). En revanche, des corpus comme ARRAU

(Uryupina et al., 2020), GNOME (Poesio, 2000) ou ONTONOTES (Hovy et al., 2006) intègrent des cas plus complexes : références génériques, anaphores incorporées (ex. clitiques en italien), ou encore références à des événements ou au discours précédent.

Enfin, nous soulignons qu’en fonction des phénomènes que l’on décide d’annoter, les formats d’annotation varient, car ils doivent être adaptés à ces choix.

Les formats d’annotation et de diffusion Plusieurs formats d’annotation sont possibles, souvent déterminés par les outils utilisés pour mener à bien la tâche. De manière générale, l’annotation de la référence suit un format avec des annotations déportées (*standoff*), permettant de conserver le texte original d’un côté et les annotations dans un fichier séparé. Il est alors crucial de se référer à un *offset* pour relier chaque annotation au texte source. Ce lien peut être réalisé de différentes manières, selon les outils et les choix opérés : soit par *offset* sur les caractères, soit par *offset* sur les tokens.

Avec un *offset* caractère, on spécifie les expressions référentielles par leurs positions de début et de fin dans le texte brut (mesurées en caractères). Ce mode offre une grande flexibilité — par exemple, il permet d’annoter des anaphores internes à un mot, comme les clitiques en italien — et a l’avantage de ne pas dépendre de la tokenisation. À l’inverse, l’*offset* token repose sur des unités pré-segmentées ; cela rend l’annotation dépendante d’un découpage préalable, et empêche d’isoler des morphèmes ou des affixes, obligeant à annoter le token entier.

Le format de diffusion des annotations dépend ensuite de l’outil. Nous pouvons distinguer le format CoNLL (notamment CoNLL-2012 (Pradhan et al., 2012)), qui est un format tabulaire où chaque token correspond à une ligne avec colonnes (POS, etc.) ; l’anaphore est indiquée par une colonne spécifique. Ce format est très utilisé pour les systèmes de résolution automatiques (Grobo, 2020 ; Wilkens et al., 2020) ; et les formats XML avec une annotation déportée :

- format MMAX2 (Müller & Strube, 2006) : format XML avec une annotation déportée qui repose sur l’*offset* de token ;
- le format XML de l’outil Glozz (Widlöcher & Mathet, 2009) avec le modèle URS (Unit-Relation-Schema). L’outil stocke le texte (.ac) et avec annotation déportée (.aa) en XML et utilise l’*offset* de caractère.
- XML-TEI-URS, développé dans le projet Democrat, s’appuie sur TEI/XML et URS : les annotations sont déportées en XML (standard TEI) avec *offset* sur le token.

3.3 Deux corpus de référence : AnnoDis et DEMOCRAT

Dans cette section, nous évoquons les corpus disponibles pour le français. Des corpus de petite taille ou qui se focalisent uniquement sur un certain type d’anaphore étaient disponibles déjà avant la mise à disposition de corpus de plus grande taille. Nous pouvons mentionner notamment le corpus DéDé (Gardent & Manuélian, 2005), qui cible l’annotation des descriptions définies, et ARCADE (Tutin et al., 2000), qui

constitue un premier corpus de grande taille, se focalisant uniquement sur l’annotation des expressions anaphoriques et cataphoriques. Le corpus RésolCo, constitué dans le cadre du projet E-Calm ([Garcia-Debanc et al., 2017](#)) est un corpus d’écrits d’élèves et d’étudiants à différents niveaux de scolarité qui présente plusieurs couches d’annotation, parmi lesquelles l’annotation de CR. Seule les référents humains y sont annotés ¹.

À plus large échelle, le corpus AnCor ([Muzerelle et al., 2013](#)) a été annoté pour le français oral. Ce corpus a été annoté par des annotateurs experts. Il propose une annotation en première mention. Le modèle d’annotation inclut des informations sur le caractère générique ou spécifique des ER et sur la nature des relations entre les mentions.

Dans ce qui suit, nous nous focalisons sur les corpus pour le français écrit que nous exploitons.

3.3.1 Le corpus AnnoDis

Le corpus AnnoDis (ANNOtation DIScursive) ([Péry-Woodley et al., 2011](#)) est un corpus du français écrit d’assez grande taille constitué dans l’objectif de sonder l’intuition des locuteurs naïfs en termes de continuité référentielle. Pour cette raison, l’annotation a été effectuée par des annotateurs non-experts. Le corpus est enrichi de deux types d’annotation manuelle ayant abouti à la composition de deux ressources, une ressource annotée en relations rhétoriques et une ressource enrichie de l’annotation de structures multi-échelles. Ces dernières correspondent aux chaînes topicales (CT) — annotation qui nous intéresse ici — et aux structures énumératives (SE). Les CT sont décrites comme des structures ayant une continuité référentielle. :

« Quant aux chaînes topicales, nous les définissons pour cette annotation comme un type particulier de chaîne de cohésion dont les éléments contiennent des unités coréférentielles (potentiellement) topicales [...] Les segments des textes [constitués par les CT] regroup[ent] des phrases reliées par un référent commun. Ce référent doit être exprimé en position préverbale (e.g. potentiellement topical) dans plusieurs phrases du segment » ([Péry-Woodley et al., 2011](#), 90).

La ressource AnnoDis annotée en structures multi-échelles est composée principalement de textes longs non narratifs relevant de trois domaines différents et divisés en trois sous-corpus :

- le sous-corpus appelé GEOP, constitué de rapports et d’articles scientifiques dans le domaine de la géopolitique publiés par l’Institut Français de Relations Internationales (IFRI) entre 2001 et 2003 ;
- le sous-corpus appelé LING, constitué d’articles scientifiques dans le domaine de la linguistique, issus du Congrès mondial de Linguistique Française de 2008 (CMLF08) ;
- le sous-corpus appelé WIK2, constitué d’articles longs tirés de l’encyclopédie Wikipédia (version de 2009) ;

1. Le modèle d’annotation, auquel nous avons contribué, s’appuie sur les modèles AnnoDis et Democrat (cf. 3.2).

Dans le corpus AnnoDis seules les chaînes potentiellement topicales ont été annotées. Pour annoter ces chaînes, les annotateurs ont suivi une approche descendante qui consiste à repérer les structures en premier, et ensuite les indices qui « participent au signalement de la structure » (Colléter et al., 2012). Les annotateurs étaient guidés par un pré-marquage de certaines expressions, toutes en position de sujet. Cela permettait aux annotateurs de faire un zoom sur les segments de textes riches en pré-marquage, sans besoin de lire linéairement l'intégralité du texte. Il est donc probable que certains maillons n'ont pas été repérés.

Concernant l'identification des indices, le projet a adopté une conception stricte du maillon (cf. 3.1). En revanche, les annotateurs n'ont pas reçu d'indications sur la façon de délimiter les maillons. En ce qui concerne les chaînes, elles ne peuvent pas dépasser l'échelle d'une section. Toutes les mentions sont reliées en suivant l'approche par *schéma* (URS).

Le [tableau 3.1](#) indique la composition de la ressource annotée en chaînes topicales telle qu'elle a été mise à disposition à la fin du projet ANR en 2012². Le tableau indique pour chaque sous-corpus son nom, la source, le genre textuel, le nombre d'articles annotés, le nombre total de mots, le nombre de chaînes annotées, ainsi que le nombre d'*indices* qui, selon la terminologie AnnoDis, recouvrent en partie les maillons des chaînes. Pour la mise à disposition de la ressource, les responsables ont réalisé

<u>Sous-corpus</u>	<u>Source</u>	<u>Genre</u>	<u>Textes</u>	<u>Mots</u>	<u>CT</u>	<u>Indices</u>
GEOP	IFRI	rapports et articles	32	266 000	234	1 125
LING	CMLF08	articles de recherche	25	169 000	88	478
WIK2	Wikipédia	articles encyclopédiques	30	231 000	266	1 853
AnnoDis			87	666 000	588	3 456

TABLEAU 3.1 – Composition de la ressource AnnoDis annotée en chaînes topicales

des post-traitements consistant en des « opérations de nettoyage et d'harmonisation : élimination de doublons, mise en cohérence des indices ajoutés par les annotateurs » (Péry-Woodley et al., 2011). Ces traitements avaient pour objectif de corriger des incohérences éventuelles et favoriser la manipulation des annotations hors Glozz³ (Widlöcher & Mathet, 2009), interface utilisée pour l'annotation de la ressource⁴.

Les annotations résultant de ces post-traitements ont été mises à disposition sous différents formats :

- au format Glozz (fichiers .ac et .aa) avec annotation déportée (cf. [chapitre 3](#));
- au format xml avec annotation intégrée, celui que nous avons exploité dans notre travail (cf. [4.1.1](#)).

2. Les informations présentées dans ce tableau sont reprises directement du site de la ressource : <http://redac.univ-tlse2.fr/corpus/annodis/>.

3. <http://www.glozz.org>

4. Pour plus de détails sur les post-traitements réalisés avant la mise à disposition de la ressource, voir [Ho-Dac et Péry-Woodley \(2014\)](#) et [Péry-Woodley et al. \(2011\)](#).

3.3.2 Le corpus DEMOCRAT

Le corpus DEMOCRAT⁵ (Landragin, 2016), mis à disposition en juin 2019, a été constitué dans le cadre du projet ANR Democrat (DEscription et MOdélisation des Chaînes de Référence : outils pour l'Annotation de corpus (en diachronie et en langues comparées) et le Traitement automatique). Democrat est un corpus annoté en chaînes de référence pour le français écrit. Il a été constitué dans l'objectif de créer un corpus à grande échelle pour l'étude des chaînes de coréférence en linguistique descriptive et assez grand pour permettre des applications de traitement automatique des langues, et notamment le développement des systèmes de résolution automatique de la référence, impliquant aussi la détection automatique des expressions référentielles (Delaborde, 2020 ; Landragin, 2021b).

Le corpus Democrat est composé de textes narratifs et non narratifs, relevant de 18 genres textuels différents, allant du 11^{ème} au 21^{ème} siècle. L'annotation a été faite sur des blocs de texte d'environ 10 000 mots, ce qui signifie que certains textes ont été annotés intégralement, comme les nouvelles, alors que les textes plus longs, par exemple les romans, ont été annotés partiellement. Dans ce cas, ce sont les 10 000 premiers mots qui sont pris en considération.

Les annotations ont été effectuées par des annotateurs experts. Le corpus a été annoté en suivant une approche ascendante, qui consiste à repérer toutes les expressions référentielles en premier, et ensuite à construire les CR. Dans Democrat aucun pré-marquage n'a été effectué. La définition du maillon est de niveau intermédiaire (cf. 3.1). Dans DEMOCRAT, on retrouve les expressions considérées comme des maillons dans la conception intermédiaire (cf. 3.1), avec des petites modifications : les pronoms réfléchis ne sont pas annotés. Toutes les expressions référentielles ont été annotées, y compris les singletons. Ce qui explique le grand nombre de CR et de maillons présents dans cette ressource (tableau 3.3). Dans le guide d'annotation Democrat une grande attention est portée à la délimitation des maillons : les SN complets sont annotés ; les bouts des syntagmes pouvant donner lieu à d'autres CR sont également annotés. Enfin, aucune délimitation des CR n'est prévue, c'est-à-dire que les CR peuvent couvrir l'intégralité du texte sans jamais se fermer. Le tableau 3.2 illustre la composition du corpus DEMOCRAT. Dans ce tableau les textes sont regroupés par siècle. Celui-ci est indiqué dans la première colonne. La deuxième et la troisième colonne donnent respectivement le nombre (#) et le pourcentage (%) des textes appartenant à chaque siècle. Ensuite, les informations portent sur le nombre total de mots, de chaînes et d'expressions référentielles, y compris les singletons, annotées dans le corpus.

Dans le corpus 54,21% des CR (11 062 CR) ne sont constituées que de deux maillons. 45,79% des CR (9 345 CR) se composent d'au moins trois maillons. Democrat est diffusé sous deux formats : au format TXM, format lié à la plateforme TXM (Heiden, Magué, & Pincemin, 2010), et au format XML-TEI-URS avec annotation déportée, où URS correspond au format *Unité, Relation, Schéma*.

5. <https://www.ortolang.fr/market/corpora/democrat>

	Siècle		Nb de mots	Nb de CR	Nb d'ER
	#	%			
21 ^e	2	3,45	21801	711	6298
20 ^e	15	25,86	175434	5302	49964
19 ^e	13	22,41	161897	4356	45589
18 ^e	6	10,34	66889	2301	18956
17 ^e	7	12,07	86147	3167	23406
16 ^e	5	8,62	61216	1630	15716
15 ^e	3	5,17	34416	728	10723
14 ^e	3	5,17	34616	1221	9678
13 ^e	2	3,45	21729	700	7412
12 ^e	1	1,72	11647	124	4138
11 ^e	1	1,72	13059	170	5043
Total	58	100	688851	20410	196923

TABLEAU 3.2 – Composition du corpus DEMOCRAT

3.3.3 Comparaison des modèles d'annotation

Dans cette section nous présentons les principales différences entre ces deux corpus et leurs modèles d'annotation. Dans le [tableau 3.3](#) nous donnons le nombre de textes, de mots, de CR et de maillons qui sont annotés dans chaque ressource. Nous donnons également la longueur moyenne des CR en nombre de maillons et l'étendue.

Corpus	Nb textes	Nb mots	Nb CR	Nb maillons	Long. moyenne	Étendue
AnnoDis	87	666 000	581	3456	5,95	38
DEMOCRAT	58	688 851	20 410	124 529 ⁶	6,33	1 439

TABLEAU 3.3 – Constitution des corpus AnnoDis et DEMOCRAT

Il existe une différence remarquable en termes de nombre de CR et de maillons annotés dans le corpus AnnoDis et DEMOCRAT. Cette différence s'explique par les différents choix d'annotation que nous allons présenter dans ce qui suit. Pour mieux illustrer cette différence, prenons l'exemple d'une CR tirée du corpus AnnoDis. Dans l'exemple (13), nous illustrons une chaîne autour de Léonard de Vinci, telle qu'annotée dans le corpus AnnoDis. Les expressions en rouge correspondent aux *indices* qui, d'après les annotateurs, signalent une continuité référentielle. Le résultat de l'annotation est une chaîne composée de 8 maillons.

- (13) **[Le jeune Léonard]**_{Léo} est proche de la nature, qu'**[il]**_{Léo} observe avec une vive curiosité et s'intéresse à tout. Il dessine déjà des caricatures et pratique l'écriture spéculaire en dialecte toscan. Giorgio Vasari, dans sa **[biographie de Léonard]**_{Léo}, raconte une anecdote sur les premiers pas dans la carrière artistique de celui qui allait devenir un des plus grands peintres de la Renaissance. Un jour, **[le père de Léonard]**_{Léo}, ser Piero, « prit plusieurs de ses

6. Les expressions référentielles annotées dans la ressource sont 196 923 au total. Nous avons calculé que 36,76% des expressions annotées (soit 72 394) sont des singletons.

dessins et les soumit à son ami Andrea del Verrocchio qu'il pria instamment de lui dire **[si Léonard]**_{Léo} devait se consacrer à l'art du dessin et s'il pourrait parvenir à quelque chose en cette matière. Andrea s'étonna fort des débuts extraordinaires **[de Léonard]**_{Léo} et exhorta ser Piero à **[lui permettre]**_{Léo} de choisir ce métier, sur quoi, ser Piero résolut que **[Léonard]**_{Léo} entrerait à l'atelier d'Andrea [...]

[AE_{w} ; 

Cet exemple montre bien la particularité du type d'annotation effectué pour la constitution de la ressource : certaines expressions, généralement considérées comme référentielles n'ont pas été annotées (e.g. *[ses] dessins*, *s'[/il] pourrait*). Un autre aspect particulier de cette ressource concerne la délimitation des maillons. Les maillons *biographie de Léonard*, *si Léonard*, *lui permettre* reflètent le manque d'indication à ce sujet. De plus, la délimitation des maillons varie en fonction de l'annotateur.

La même chaîne selon le modèle Democrat serait constituée de 16 maillons, soit deux fois plus de maillons, comme le montre l'exemple (14) dans lequel nous avons laissé en rouge les maillons annotés dans AnnoDis, nous mettons en orange les maillons pour lesquels la délimitation est différente suivant le modèle Democrat et en bleu les maillons qui ne sont pas annotés dans AnnoDis.

- (14) **[Le jeune Léonard]**_{Léo} est proche de la nature, qu'**[il]**_{Léo} observe avec une vive curiosité et **[s'intéresse]**_{Léo} à tout. **[Il]**_{Léo} dessine déjà des caricatures et **[pratique]**_{Léo} l'écriture spéculaire en dialecte toscan. Giorgio Vasari, dans sa biographie **[de Léonard]**_{Léo}, raconte une anecdote sur les premiers pas dans la carrière artistique de **[celui]**_{Léo} **[qui]**_{Léo} allait devenir **[un des plus grands peintres de la Renaissance]**_{Léo}. Un jour, le père **[de Léonard]**_{Léo}, ser Piero, « prit plusieurs de **[ses]**_{Léo} dessins et les soumit à son ami Andrea del Verrocchio qu'il pria instamment de lui dire si **[Léonard]**_{Léo} devait se consacrer à l'art du dessin et s'**[il]**_{Léo} pourrait parvenir à quelque chose en cette matière. Andrea s'étonna fort des débuts extraordinaires **[de Léonard]**_{Léo} et exhorta ser Piero à **[lui]**_{Léo} permettre de choisir ce métier, sur quoi, ser Piero résolut que **[Léonard]**_{Léo} entrerait à l'atelier d'Andrea [...].

De plus, le même segment de texte, annoté selon le modèle Democrat donnerait lieu à plusieurs CR et pas seulement à la chaîne autour de Léonard (considérée dans AnnoDis comme topicale), comme illustré dans l'exemple (15). Dans le corpus DEMOCRAT, les singletons sont également annotés (en noir dans l'exemple suivant) ainsi que les chaînes composées de deux maillons. Les CR sont signalées par des couleurs différentes.

- (15) **[Le jeune Léonard]**_{Léo} est proche de **[la nature]**_{Nat}, **[qu']**_{Nat}**[il]**_{Léo} observe avec **[une vive curiosité]**_C et **[s'intéresse]**_{Léo} à tout. **[Il]**_{Léo} dessine déjà **[des caricatures]**_{Car} et **[pratique]**_{Léo} l'écriture spéculaire en **[dialecte toscan]**_{DT}**[écrit]**. **[Giorgio Vasari]**_{Gio}, dans **[sa]**_{Gio} biographie **[de Léonard]**_{Léo}**[Bio]**_{Léo}, raconte une anecdote sur les premiers pas dans la carrière artistique de **[celui]**_{Léo} **[qui]**_{Léo} allait devenir **[un des plus grands peintres de la Renaissance]**_{Léo}. Un jour, **[le père [de Léonard]]**_{Léo}**[père]**_L, ser Piero, « prit plusieurs de **[ses]**_{Léo} **[dessins]**_{dessins} et **[les]**_{dessins} soumit à **[son]**_{père}**[ami Andrea del Verrocchio]**_{Andrea} qu'**[il]**_{père}**[père]**_L pria instamment de **[lui]**_{père}**[père]**_L dire si **[Léonard]**_{Léo} devait se consacrer à **[l'art]**

du dessin]_{Ad} et s’[il]_{Léo} pourrait parvenir à quelque chose en [cette matière]_{Ad}. [Andrea]_{Andrea} s’étonna fort [des débuts extraordinaires [de Léonard]_{Léo} débuts et [exhorta]_{Andrea} [ser Piero]_{pèreL} à [lui]_{Léo} permettre de choisir [ce métier]_{Ad}, sur quoi, [ser Piero]_{pèreL} résolut que [Léonard]_{Léo} entrerait à [l’atelier d’[Andrea]_{Andrea} l’atelierA [...].

Dans le guide d’annotation Democrat une grande attention est portée à la délimitation des maillons. Comme le montre l’exemple que nous avons montré, les SN complets sont annotés ; les bouts des syntagmes pouvant donner lieu à d’autres CR sont également annotés. Nous ne sommes pas en mesure de faire la même comparaison en partant d’une CR annotée dans Democrat et voir ce que cela donnerait en suivant le modèle d’annotation AnnoDis, car comme nous l’avons dit, dans ce dernier les annotateurs suivaient une approche descendante et l’objectif était de tester l’intuition de locuteurs naïfs.

3.4 Impact du modèle d’annotation sur la description des chaînes de référence

Les différences entre les différents modèles d’annotation peuvent avoir un impact sur les résultats obtenus à partir de l’analyse des corpus annotés en chaînes de référence. Un même critère de description peut produire des résultats différents selon le modèle suivi, tant sur le plan de la dimension que de la composition des chaînes. Il est donc essentiel de prendre en compte ces effets lors de l’interprétation des résultats, en particulier dans le cadre d’analyses comparatives des corpus hétérogènes.

Dans cette section, nous présentons l’impact potentiel du modèle d’annotation sur les critères de description présentés dans le [chapitre 2](#), en nous appuyant sur les deux ressources principales mobilisées dans cette thèse : AnnoDis et DEMOCRAT. Certains choix d’annotation influencent directement les indicateurs classiques de la dimension des chaînes : longueur, couverture et distance entre les maillons. Dans AnnoDis, le choix méthodologique a été de fermer et de redémarrer les chaînes à chaque nouvelle section, ce qui conduit logiquement à des chaînes plus courtes, notamment pour les référents principaux. À l’inverse, dans DEMOCRAT, les chaînes ne sont jamais interrompues, ce qui permet leur maintien tout au long du texte. Ce choix favorise une représentation plus complète de la continuité référentielle, et permet, le cas échéant, d’opérer une segmentation à posteriori ([Landragin, 2016](#) ; [Schnedecker, 2021](#)). Dans DEMOCRAT, la non-interruption des chaînes implique que certains maillons peuvent être très éloignés les uns des autres, même dans des chaînes courtes. Il n’est pas rare d’observer des chaînes secondaires ou d’arrière-plan qui ne comptent que trois maillons, mais s’étendent sur une grande portion du texte, comme dans (16). À l’inverse, dans AnnoDis, la distance entre les maillons est globalement plus réduite, notamment dans les chaînes les plus courtes. Cela est lié à l’approche descendante focalisée sur des segments de texte localisés et encore une fois, au choix de fermer les chaînes à chaque section.

- (16) Tout ressemblait **au désordre qui** conclut les maladies : avant la mort qui se charge de rendre tous les corps invisibles, l'unité de la chair se dissipe, chaque partie dans cette multiplication tire dans son sens.

[85 phrases]

La mort de nos cousins et de nos frères, la licence donnée par l'absence de nos pères, les objets meurtriers de nos aînés ont fourni **au désordre** de mystérieux aliments : c'était celui de l'enfance miraculeusement soustrait aux complots pacifiques de l'ordre : la guerre nous a permis de vivre.

[255 phrases]

Je fus ce voyageur : circuler sur de petits vapeurs écaillés, sur des dhows indigènes de l'un à l'autre bord de ce profond canal des enfers, rebondir sur les remparts de l'Afrique et de l'Arabie, ces mouvements **du désordre** n'imitent pas longtemps les allures de la liberté. [P_{A} ; ⚡]

Au niveau de la composition, l'influence des modèles d'annotation se manifeste notamment à travers la nature des maillons, leur fonction syntaxique et leur complexité morphosyntaxique. Dans une approche descendante comme celle d'AnnoDis, on peut s'attendre à une sous-représentation des formes peu saillantes (pronoms, sujets zéro, etc.), qui sont plus difficiles à identifier sans une lecture linéaire complète. À l'inverse, dans DEMOCRAT, les maillons sont identifiés de façon systématique (approche ascendante), ce qui garantit une meilleure représentativité des formes pronominales et implicites. Par ailleurs, concernant la nature des maillons, la différence entre annotateurs experts et non experts joue également un rôle important. Les annotateurs non experts ne vont pas annoter systématiquement toutes les formes, et encore moins les sujets zéro qui n'ont pas de forme linguistique. Cette différence d'approche se traduit également dans la distribution des fonctions syntaxiques. Dans AnnoDis, on observe une prépondérance des sujets, tandis que dans DEMOCRAT, la diversité des fonctions est plus importante (objets, compléments, etc.), reflétant une annotation plus exhaustive, comprenant également tous les référents secondaires, comme illustré dans (16). La délimitation des maillons est également affectée par le modèle. Dans DEMOCRAT, les annotateurs ont reçu la consigne d'inclure tous les modificateurs dans les maillons, ce qui produit des unités syntaxiquement plus longues et complexes. À l'inverse, dans AnnoDis, où la délimitation est plus libre et soumise à l'intuition des annotateurs, les maillons pourraient être généralement plus courts.

Ces observations montrent que le modèle d'annotation peut être considéré comme un facteur de variation à part entière, aux côtés d'autres variables externes ou internes identifiées dans le chapitre 2. Il agit directement sur les résultats obtenus, au même titre que le type ou le genre textuel. AnnoDis et DEMOCRAT permettent d'interroger la variation des CR selon les facteurs de variation suivants : le type et genre textuel (narratif vs non narratif, scientifique vs encyclopédique, etc.) ; le modèle d'annotation (guidé vs souple, approche ascendante vs descendante, experts vs naïfs, restriction aux référents topicaux) ; En revanche, certains facteurs ne peuvent pas être investigués en l'état : la nature des référents n'est pas annotée systématiquement dans les corpus existants ; des enrichissements sont donc nécessaires (cf. 6.1). Les facteurs internes liés à la structure textuelle sont partiellement disponibles : AnnoDis propose une annotation

des paragraphes, mais ce n'est pas le cas dans DEMOCRAT. L'absence de structuration en XML-TEI, notamment, limite les possibilités d'analyse croisée.

Afin de pouvoir décrire les chaînes issues de ces deux corpus hétérogènes et analyser leur fonctionnement, il est nécessaire d'opérer des traitements pour préparer les données à l'analyse et de mettre en place une méthode d'analyse adaptée. La seconde partie de cette thèse est ainsi consacrée à la présentation des étapes de préparation des données, ainsi qu'à la méthode que nous proposons pour analyser les chaînes de référence, avec une attention particulière aux modalités d'enchaînement des maillons.

Deuxième partie

Caractériser les chaînes de référence par une approche linéaire

Analyser les chaînes de référence et identifier des types de chaînes n'est pas chose aisée : à l'hétérogénéité des ressources annotées, qui traduit un manque de consensus autour de la notion même de chaîne et complique la comparabilité des résultats, s'ajoute l'absence d'une méthodologie d'analyse unifiée. Comme le souligne [Delaborde \(2020, 63\)](#) « il n'existe pas réellement de méthode scientifique universelle d'analyse [des chaînes de référence] » bien que certaines propositions aient émergé, notamment en ce qui concerne les critères permettant de décrire leurs caractéristiques ([Schneidecker, 2019, 2021](#)) :

Pour décrire les [...] chaînes et mettre au jour leurs différences, il faut adopter un système de mesures permettant de rendre compte de la manière la plus objective possible d'un certain nombre de facteurs relatifs aux « maillons » de ces CR ([Schneidecker, 2019](#)).

Pour appréhender les caractéristiques des chaînes, il est essentiel de s'appuyer sur un système de mesures objectif, capable de rendre compte des propriétés des maillons qui les composent ([Schneidecker, 2019](#)). Ainsi, comme nous l'avons vu (cf. [chapitre 2](#)), plusieurs critères de description ont été mis au jour dans les travaux s'intéressant aux chaînes de référence ([Kibrik, 2011](#) ; [Kunz, 2010](#) ; [Recasens & Martí, 2010](#) ; [Rousier-Vercruyssen & Landragin, 2019](#)). Bien que ces études représentent un apport précieux, la plupart d'entre elles, en linguistique comme en TAL, considèrent les chaînes comme des classes d'expressions équivalentes ([Potau, 2010](#) ; [Uryupina, Kabadjov, & Poesio, 2016](#)). Suivant cette conception, définie comme *paradigmatisante* par [Schneidecker \(2021\)](#), « les chaînes sont assimilées à un paradigme ou une famille d'expressions renvoyant à une même entité, abstraction faite de leur environnement discursif » ([Schneidecker, 2021, 86](#)). Les travaux qui adoptent cette conception utilisent des critères formels tels que la longueur ou l'interdistance moyenne des chaînes. Au niveau des maillons, les critères utilisés sont le plus souvent morpho-syntaxiques et syntaxiques, telles les catégories grammaticales des maillons ou leurs fonctions syntaxiques (cf. [chapitre 2](#)). Ces critères sont utilisés pour caractériser les chaînes suivant une approche *globale* : il s'agit alors d'identifier et de décrire les caractéristiques globales des CR en quantifiant, par exemple, les catégories grammaticales du premier maillon et/ou de l'ensemble des maillons. Ensuite, il s'agit de comparer les fréquences obtenues dans différents genres textuels pour rendre compte d'éventuelles variations entre les genres analysés.

Ces critères ne sont pas en mesure de rendre compte de « l'enchaînement, au sens littéral du terme, des expressions référentielles » ([Schneidecker, 2017](#)).

[S]i une chaîne de référence se constitue d'un ensemble d'expressions coréférentielles « équivalentes », cet ensemble est aussi contraint d'une double manière : par la linéarité du discours et par la cohabitation avec les expressions coréférentielles d'autres chaînes ([Schneidecker, 2021, 54](#)).

C'est pour cette raison qu'une approche *globale* n'est pas complètement satisfaisante dès lors que l'on s'intéresse à leur fonctionnement textuel, à leur rôle dans la création de liens de cohésion et que l'on considère qu'une chaîne n'est pas seulement une classe d'expressions équivalentes, mais une unité discursive, contrainte par la linéarité du discours.

Cela ne signifie pas qu'il faut s'intéresser exclusivement aux enchaînements au détriment d'une description globale. Au contraire, nous pensons qu'il est important de considérer les deux niveaux d'analyse, ceux-ci étant, de notre point de vue, complémentaires. Notre proposition est donc d'utiliser une approche *linéaire*, consistant à exploiter des critères pour décrire les modalités d'enchaînement des maillons, afin de compléter une description des caractéristiques globales des CR et décrire les stratégies d'introduction et maintien des référents.

Pour mieux comprendre l'intérêt de prendre en compte la linéarité du discours, nous proposons une représentation schématique de deux chaînes CR1 et CR2 dans la [figure 3.2](#) qui montre comment, malgré une composition globale similaire (nombre de maillons, proportion de noms propres et de pronoms), ces deux chaînes peuvent différer si on les considère dans la linéarité du discours. Par exemple, la distance entre les maillons, mesure qui permet de rendre compte de la linéarité du discours, nous pouvons remarquer une différence entre les deux chaînes. Cette approche, consistant à rester dans la linéarité, est appelée *syntagmatisante* par [Schneidecker \(2021\)](#).

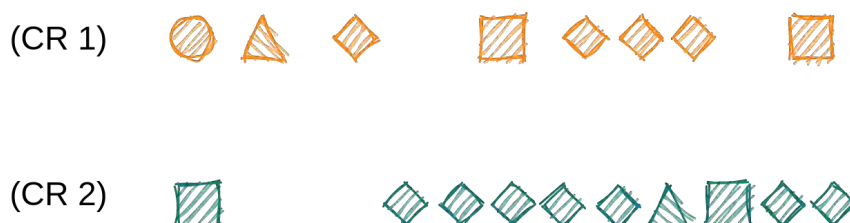


FIGURE 3.2 – Représentation schématique de deux chaînes de référence, adaptée de [Schneidecker \(2021, 55\)](#)

Les chaînes de la [figure 3.2](#) sont codées par des formes géométriques qui représentent les types d'expressions référentielles constituant les chaînes (carré pour les noms propres, losange pour les pronoms, cercle pour les syntagmes nominaux définis, triangle pour les possessifs). Elles sont disposées linéairement en fonction de la distance relative entre les maillons. Cette visualisation met en évidence les différences linéaires entre CR1 et CR2 : les principales différences entre les deux chaînes concernent la distance entre les maillons, la manière dont le référent est introduit (catégorie de la première mention) et la manière dont il est maintenu, c'est-à-dire la distribution linéaire des catégories tout au long des chaînes (i.e. plus grande succession de pronoms dans la CR2 que dans la CR1). Si l'on veut capter ces différences dans la composition des CR, il est important de compléter la description des caractéristiques globales et de la dimension des chaînes avec une description des caractéristiques liées à l'enchaînement des maillons.

La description des enchaînements favoriserait la prise en compte de toute la complexité de composition des chaînes. L'analyse des enchaînements permettrait de distinguer des CR comme celles de la [figure 3.2](#) qui présentent des stratégies relativement différentes malgré une ressemblance au niveau de leur composition globale. Une telle

description compléterait ainsi les analyses plus traditionnelles basées sur les caractéristiques globales des CR, sur la nature du premier maillon ou la répartition des catégories grammaticales des maillons (Longo, 2013 ; Longo & Todirascu, 2014 ; Schnedecker, 2005, 2014). Décrire les chaînes sous l'angle des enchaînements permettrait aussi d'appréhender les stratégies de continuité référentielle que réalisent les chaînes et l'influence de l'organisation globale du discours sur ces stratégies (influence de la structure textuelle ou de la présence simultanée de plusieurs chaînes).

Ainsi, grâce à la description systématique des enchaînements, il serait possible de rendre compte des régularités et des différences qui caractérisent les CR, tout en prenant en compte leur complexité, pour « dégager des « patrons » ou, en d'autres termes, des modes de composition similaires » (Schnedecker, 2021, 87). C'est cet objectif qui guide notre analyse. L'objectif de cette deuxième partie est de proposer une méthodologie inédite permettant d'analyser de manière systématique les enchaînements des maillons dans l'objectif d'identifier des modes de composition différents, formalisés par la proposition de patrons des CR.

Dans un premier temps, nous présentons les traitements effectués sur AnnoDis et DEMOCRAT pour les préparer à l'analyse. Nous proposons également une description de l'anatomie des CR des deux ressources, suivant l'approche globale, comme dans les études existantes (chapitre 4). Dans le chapitre 5, nous expliquons en détail la méthode que nous proposons afin de décrire les enchaînements des maillons suivant une approche linéaire. Enfin, dans le chapitre 6, nous présentons l'enrichissement des données et le paramétrage utilisé pour appliquer la méthode proposée à l'étude des chaînes de référence.

Chapitre 4

Au cœur des données : de la sélection à la description

Contents

4.1	Le corpus AnnoDis : harmonisation des annotations	77
4.1.1	Format xml avec annotation intégrée	78
4.2	Du corpus DEMOCRAT à la ressource DEMRC	80
4.2.1	Nettoyage des annotations et constitution de la ressource DEMRC	80
4.2.2	Format d'analyse proposé	81
4.2.3	Pré-traitement de la ressource DEMRC	84
4.3	Anatomie des chaînes dans les deux ressources exploitées	86
4.3.1	Longueur des chaînes	87
4.3.2	Distance entre les maillons	90
4.3.3	Nature des maillons	92
4.4	Bilan	95

Dans notre travail, nous adoptons une approche bottom-up (Biber, Connor, & Upton, 2007), consistant à faire ressortir des régularités à partir de l'observation des données. Pour ce faire, nous exploitons des corpus annotés en CR, et la première étape de notre méthodologie (figure 4.1) consiste en la sélection de corpus à exploiter. Dans le chapitre 3, nous avons vu que les corpus annotés en référence sont hétérogènes. Cette hétérogénéité engendre des difficultés pour la comparabilité des résultats. Toutefois, cela peut également représenter un atout pour l'étude des CR : la prise en compte de corpus hétérogènes pourrait permettre d'étudier différentes facettes des phénomènes référentiels. Cela nous amène à nous poser plusieurs questions :

- les résultats obtenus à partir de ressources hétérogènes sont-ils vraiment incomparables ? Peut-on identifier des régularités malgré les différences dans les choix d'annotation ?
- quel est l'impact réel des modèles d'annotation sur les résultats obtenus ?

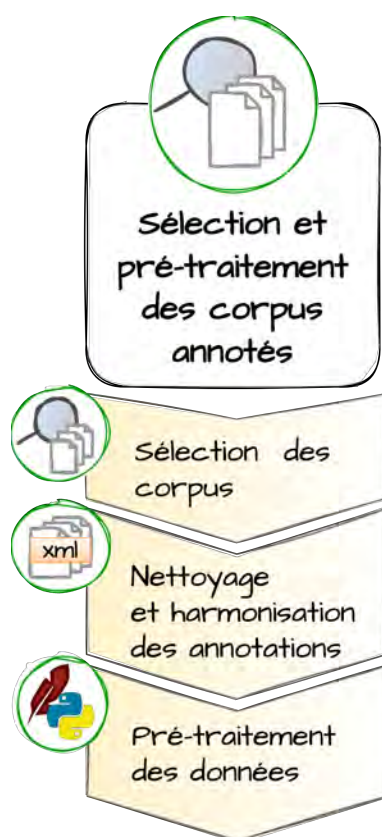


FIGURE 4.1 – Première étape : sélection et pré-traitement des données

- peut-on tirer parti de cette hétérogénéité pour mieux comprendre le fonctionnement des CR ? Le cas échéant, comment faire ?

Afin d'apporter des éléments de réponse à ces questions, nous avons décidé d'utiliser deux corpus du français écrit aux caractéristiques très différentes, AnnoDis et DEMOCRAT (cf. 3.3.1 et 3.3.2). Les modèles et procédures d'annotation, les formats de diffusion ainsi que la composition de chaque corpus, ont été abordés en détail dans le chapitre 3. Dans le présent chapitre, nous nous concentrons sur les traitements effectués afin de les préparer à l'analyse. Les deux corpus présentent une double difficulté : l'hétérogénéité des formats de diffusion et l'hétérogénéité des choix d'annotation. Pour pallier ces difficultés, nous proposons d'effectuer des traitements adaptés aux caractéristiques de chaque corpus. Ceux-ci visent à préparer les ressources de sorte à pouvoir en extraire les mêmes critères de description et leur appliquer la même méthode d'analyse. Certains choix de traitement sont donc liés à la nature des annotations, à savoir le nettoyage et l'harmonisation des annotations (cf. 4.1 et 4.2). D'autres choix sont, quant à eux, liés au format de diffusion. Nous proposons un *format d'analyse* (cf. 4.2.2) qui facilite :

- les traitements automatiques via des analyseurs morpho-syntaxiques comme Talismane (Urieli, 2013) ou Stanza (Qi, Zhang, Zhang, Bolton, & Manning, 2020) ;
- l'enrichissement des informations dans le corpus d'origine (nature des maillons, entités nommées, par exemple) grâce aux traitements mentionnés ci-dessus. Ces

informations, utiles pour la description des CR, sont rarement annotées dans les corpus ;

- la visualisation et l’analyse qualitative des données. Ce dernier point est particulièrement important à notre avis : d’après nous, les analyses quantitatives devraient systématiquement être accompagnées d’un retour au contexte, permettant une interprétation qualitative.

Le chapitre se conclut avec un bilan des ressources résultant des différents traitements et que nous allons utiliser pour les analyses des CR.

4.1 Le corpus AnnoDis : harmonisation des annotations

Nous rapportons ici le travail réalisé dans le cadre de notre mémoire de master 2 qui a donné lieu à une publication (Federzoni et al., 2020). Ce travail avait pour objectif de comparer les chaînes topicales (CT) annotées dans la ressource aux CR décrites dans les études existantes (Salles (2015a, 2015b), Longo et Todirascu (2014), Schnedecker (2014), Schnedecker (2005), entre autres). En nous inspirant d’un travail de Ho-Dac, Péry-Woodley, et Tanguy (2010) sur les structures énumératives, nous avons tout d’abord mené une analyse exploratoire des annotations visant à obtenir une vision globale de l’ensemble des CT et des étiquettes associées aux indices. Cette étude nous a amenée à réaliser des opérations de nettoyage et d’harmonisation des annotations, nécessaires à l’exploitation de la ressource pour l’étude des chaînes de référence. Ces post-traitements ont été effectués à partir de la version de la ressource au format `xml` parsée par Talismane (Urieli, 2013). Ils ont principalement porté sur la (re-)délimitation des maillons¹ (par exemple, lorsqu’un déterminant n’avait pas été inclus dans le maillon) et sur l’attribution de la catégorie grammaticale aux maillons (lorsqu’elle n’avait pas été indiquée par l’annotateur). Cela nous a notamment permis d’attribuer une catégorie grammaticale à la majorité des indices (Federzoni, 2019 ; Federzoni et al., 2020). Le tableau 4.1 montre la répartition des catégories grammaticales

Nature des maillons	AnnoDis	
	#	%
SN _{def}	1198	34,66
Pro	1026	29,69
Np	442	12,79
SN _{dem}	272	7,87
Poss	182	5,27
Autre	146	4,22
SN _{ind}	126	3,65
SN _∅	64	1,85
Total	3456	100

TABLEAU 4.1 – Distribution de la nature des maillons dans AnnoDis après harmonisation des annotations

1. Rappelons que nous considérons les indices annotés dans AnnoDis comme des maillons.

des maillons, triées par fréquence, dans la version harmonisée de AnnoDis.

Les post-traitements que nous avons brièvement présentés ci-dessus ont permis de convertir les CT en chaînes de référence et d'éliminer certains biais qui auraient pu être causés par le protocole d'annotation (Federzoni et al., 2020). Nous exploitons donc cette dernière version de AnnoDis, dont nous donnons un aperçu quantitatif ci-dessous :

- 581 chaînes (588 CR avaient été annotées dont 7 vides que nous avons éliminées)
- 3 456 maillons dont :
 - 1 747 prémarqués automatiquement et validés (50,55%)
 - 1 709 annotés manuellement (49,45%)

Enfin, nous proposons de renommer les trois sous-corpus :

- GEOP devient *articles_scientifiques*_{Geop}
- LING devient *articles_scientifiques*_{Ling}
- WIK2 devient *articles_encyclopédiques*_{Wiki}

Nous avons fait le choix de mettre en avant le genre textuel tout en gardant la trace du domaine duquel les textes ont été tirés pour deux raisons. D'une part, le genre textuel est l'un des facteurs de variation que nous analyserons ([chapitre 4](#)), et d'autre part, cela nous permet de nous aligner sur le corpus DEMOCRAT. Nous soulignons toutefois que la notion de genre textuel est ici à nuancer : les sous-corpus GEOP et LING correspondent à un même genre, mais appartiennent à des domaines diversifiés.

4.1.1 Format **xml** avec annotation intégrée

Le format **xml** avec annotation intégrée consiste en la reproduction de toutes les CT et de tous les maillons annotés dans la ressource AnnoDis. Les chaînes sont présentées comme des *segments de texte*, chaque segment correspondant à la portion de texte couverte par la chaîne, comme illustré dans la [figure 4.2](#). Si deux chaînes couvrent la même partie du texte, le segment est dupliqué. Le fichier au format **xml** contient exclusivement les CT et les maillons annotés. Les chaînes sont délimitées via la balise `<CT>` et les indices via la balise `<cue>`.

Nous avons trouvé ce format particulièrement intéressant pour plusieurs raisons. Tout d'abord, il favorise le retour aux données et la visualisation des chaînes, facilitant ainsi une analyse qualitative des CR. En effet, il est possible de revenir aux données pour visualiser et analyser aisément les CR et les maillons dans leur contexte. Il permet de considérer une chaîne à la fois, indépendamment des autres. Par ailleurs, ce format rend les annotations accessibles à tous, sans devoir passer par des outils ou des manipulations spécifiques avant de pouvoir visualiser les annotations. Ce qui est impossible avec des annotations déportées, sans l'utilisation d'un logiciel dédié. Ensuite, ce format permet d'appliquer des traitements automatiques, comme l'étiquetage morpho-syntaxique, tout en conservant les annotations. Les outils d'étiquetage automatique s'appuient sur l'*offset* au niveau des caractères, c'est-à-dire le début et la fin du mot en nombre de caractères. Nous avons vu dans le [chapitre 3](#) (cf. [3.2](#)) que ce type d'information n'est pas toujours disponible. Dans le cas de DEMOCRAT, par exemple,


```

<structure>
<context type="before">... si le président choisissait de ne pas se présenter au
Capitole</context>
<CT NbCar="705" startofs="37685" para="0.5" list="0" heading="-1" id=
"geop_19CT_coder1_1253610481687" file="geop_19">
<firstCOREF>George H.W.</firstCOREF>
<tags><tag nature="Ia_COREF_R" start="37687">George H.W.</tag><tag nature=
"Ia_COREFdef_R" start="37999">le président</tag><tag nature="Ia_COREFpro" start=
"38112">Il</tag><tag nature="Ia_COREFpro" start="38337">il</tag></tags>
<segment schema="CT_coder1_1253610481687" start="37685" end="38390"><fullVersion>.
<cue type="Ia_COREF_R">George H.W.</cue> Bush s'inquiéta de la volonté du régime
irakien de se doter d'armes de destruction massive, ce à quoi Al Gore, alors
sénateur (démocrate, Tennessee), répliqua qu'il s'agissait d'une manoeuvre
maladroite destinée à gagner le soutien du Congrès. Se sentant dans une situation de
plus en plus délicate, <cue type="Ia_COREFdef_R">le président</cue> comprit que la
guerre était inévitable, et que son issue déterminerait son propre avenir politique.
<cue type="Ia_COREFpro">Il</cue> estimait en effet que la mauvaise conduite des
opérations pouvait lui coûter cher, tandis qu'une victoire facile ne renforcerait
pas nécessairement sa position face au Congrès et à ses adversaires politiques.
Ainsi, comme <cue type="Ia_COREFpro">il</cue> l'écrivait dans son journal, le 20
décembre 1990 :<break type="paragraph"/></fullVersion>
<shortVersion>. George H.W. Bush s'inquiéta de la volonté du régime ...</shortVersion>
</segment>
</CT>
<context type="after"> " Si la guerre est rapide, avec le moins de pertes possible,
...</context>
</structure>

```

FIGURE 4.2 – Aperçu des CT dans la ressource AnnoDis : version xml avec annotation intégrée

les annotations déportées font référence au texte en exploitant les identifiants des tokens ; dans les sorties obtenues par l'analyseur morpho-syntaxique, cette référence est faite par les *offsets* sur le caractère. Cela rend difficile la mise en correspondance entre les annotations et les analyses morpho-syntaxiques obtenues ; sans oublier que la tokenisation peut varier d'un outil à l'autre. Par ailleurs, nous rappelons qu'exploiter le format étiqueté nous a permis de nettoyer et d'harmoniser les annotations de la ressource AnnoDis, et notamment d'attribuer une catégorie grammaticale à la majorité des indices annotés (Federzoni et al., 2020).

Cependant, le format proposé pour AnnoDis présente, à notre avis, deux limites. La première concerne l'absence d'identifiants attribués aux maillons. En effet, comme on peut le voir dans la figure 4.2, la balise *cue* n'est accompagnée d'aucun attribut. Cela engendre des difficultés lorsque l'on a besoin de revenir aux annotations d'origine. Par exemple, nous avons apporté des modifications sur la délimitation des indices et nous avons intégré des informations sur la nature des maillons. Ces modifications ont été effectuées sur le format xml tagué par Talisman. La deuxième limite concerne le fait que les balises utilisées pour délimiter les chaînes et les maillons ne correspondent pas aux standards proposés par la norme TEI. Toutefois, les balises sont explicites et ne posent aucun problème pour la manipulation des données².

En nous appuyant sur l'exemple de la ressource AnnoDis (Péry-Woodley et al., 2011) et sur les remarques que nous avons soulevées, nous proposons un format d'analyse similaire que nous appliquons au corpus DEMOCRAT afin d'obtenir un format

2. La mise à disposition d'une nouvelle version de la ressource avec une adaptation des balises aux standards TEI-P5 (comme nous le proposons pour la ressource DEMOCRAT_RC_19-21 (cf. 4.2.2)), avec l'ajout des modifications effectuées est envisagée.

proche à celui d'AnnoDis, nous permettant d'y appliquer les mêmes analyses par la suite. Avant de détailler les caractéristiques du format d'analyse proposé, nous allons décrire les traitements que nous avons effectués sur le corpus DEMOCRAT.

4.2 Du corpus DEMOCRAT à la ressource DEMRC

La deuxième ressource que nous utilisons est la ressource DEMOCRAT_RC_19-21. Celle-ci est le résultat d'une extraction automatique des CR du corpus DEMOCRAT (Landragin, 2016) (cf. 3.3.2).

4.2.1 Nettoyage des annotations et constitution de la ressource DEMRC

Les traitements que nous avons effectués sur les annotations d'origine ont donné lieu à la constitution de la ressource DEMOCRAT_RC_19-21³ (désormais DEMRC) que nous exploitons dans la suite de notre thèse. DEMRC est le résultat d'une extraction automatique de toutes les chaînes de référence (au moins 3 maillons) telles qu'annotées dans les textes du 19^{ème}, 20^{ème} et 21^{ème} siècles constituant le corpus DEMOCRAT.

Sélection des textes et des chaînes

Pour constituer cette ressource, nous avons tout d'abord effectué une sélection des textes à analyser. Dans le cadre de ce travail, nous ne nous intéressons pas aux CR d'un point de vue diachronique. Nous avons donc pris en compte les textes des 19^{ème}, 20^{ème} et 21^{ème} siècles (en gras dans le [tableau 3.2](#)).

Le deuxième traitement a consisté en une sélection des chaînes de référence. Nous avons fait le choix de ne considérer que les chaînes constituées d'au moins trois maillons. Les singletons (38 619 dans les 29 textes DEMOCRAT) et les chaînes constituées de deux maillons (5 652 dans les 29 textes DEMOCRAT) ne sont pas reproduits dans la ressource. Cette dernière est ainsi composée de 29 fichiers qui rassemblent toutes les chaînes d'au moins trois maillons annotés dans les textes retenus correspondants⁴. Les 29 textes pris en compte sont de type narratif (18 textes) et non narratif (11 textes) et couvrent 7 genres textuels différents, pour un total de 347 155 mots.

Nettoyage des annotations

Lors de la constitution de la ressource DEMRC, nous avons effectué deux autres traitements des annotations. Ceux-ci ont porté sur :

3. Mise à disposition sur Ortolang : https://hdl.handle.net/11403/democrat_rc_19-21

4. Pour des simples raisons techniques liées au traitement des fichiers xml lors de la transformation des annotations, le texte *Chateaubriandgenie* a été exclu de cette sélection en raison de sa structure XML-TEI qui était différente de celle des autres textes. En effet, c'est le seul texte où la structure textuelle (chapitres, titres, etc.) est explicitement annotée à l'aide des balises prévues à cet effet par la norme TEI-P5 (voir l'[annexe A](#)).

1. l'extraction et le nettoyage de la segmentation en phrases, effectuée automatiquement par TreeTagger (Schmid, 1994) dans le corpus DEMOCRAT. Nous avons corrigé automatiquement la segmentation en phrases lorsqu'un maillon chevauchait plusieurs segments identifiés comme des phrases. Nous avons observé que ces chevauchements étaient dus à la présence d'un point que TreeTagger considère comme un indice de fin de phrase, même lorsque cela n'est pas le cas. Ainsi, suite à l'erreur de segmentation observée dans l'exemple suivant où les balises <s> délimitent les phrases et les balises <rs> les maillons :

```
<s n="244"> [...] j'ai entendu mes parents parler de la visite
d'<rs type="mention"> un M. </s> <s n="245">
Grangier</rs> [...] </s>
```

les phrases ont été fusionnées pour aboutir au résultat suivant :

```
<s n="244,245"> [...] j'ai entendu mes parents parler de la vi-
site
d'<rs type="mention"> un M. Grangier</rs> [...] </s>.
```

Par ce traitement, 196 fusions ont été réalisées dont 81,6% (160) concernent la fusion entre deux phrases, 15,81% (31) la fusion de trois phrases et 2,55% (5) la fusion de plus de trois phrases.

2. La délimitation des maillons : dans la majorité des cas, les maillons ont été indiqués tels qu'ils l'étaient dans le corpus DEMOCRAT. Dans certains cas, plusieurs maillons imbriqués sont associés à une même chaîne, comme dans l'exemple suivant où « du CP » est un maillon imbriqué dans le maillon « Les enfants de la maternelle et du CP », tous deux appartenant à la même chaîne : <rs type="mention"> Les enfants de la maternelle<rs type="mention"> et du CP</rs></rs> . Dans des cas de ce type, le maillon imbriqué n'a pas été retenu et le maillon le plus large a été conservé (<rs type="mention"> Les enfants de la maternelle et du CP</rs>). Cette décision est justifiée par le fait que des maillons imbriqués devraient appartenir à deux chaînes de référence distinctes. Une fois ce traitement effectué, si la chaîne n'est plus constituée d'au moins trois maillons, celle-ci ne figurera pas dans la ressource DEMRC. Ainsi, 262 maillons imbriqués ont été supprimés et 14 chaînes ont été écartées.

Le [tableau 4.2](#) indique la composition de la ressource après les traitements de nettoyage décrits.

4.2.2 Format d'analyse proposé

En nous appuyant sur l'exemple de la ressource AnnoDis, nous avons transformé les annotations déportées en annotations intégrées dans DEMRC. Nous présentons donc chaque chaîne sous forme de *segment de texte*, celui-ci correspondant uniquement à la portion de texte couverte par la chaîne en question. De ce fait, une même partie de texte peut être répétée plusieurs fois, comme le montre la [figure 4.3](#), où deux chaînes couvrent à peu près la même portion de texte⁵.

5. Si les phrases ne contiennent pas de maillons, le texte correspondant à la phrase n'est pas présent.

Type de texte	Genre	Fichier	Nb. mots	Nb. CR	Nb. maillons
Non narratif	articles de presse	estrepublicain	12254	170	888
		estrepublicain-2	11605	151	811
	articles encyclopédiques	articleswiki	10628	118	854
		pamphlet	aden	13290	131
		codecivilsdesfrancais 1	12943	218	1853
		codecivilsdesfrancais 2	14431	202	1773
		codeprocedurepenale	12509	330	2086
	texte juridique	conventioninstitutiv	5384	73	586
		conventionmilieumarin	12860	294	2053
		convtransportaerien	9402	138	1233
		conventionthontropical	11173	200	1613
Total DemRC{Nar—}			126479	2025	14887
Narratif	biographie	madamedehautefort	11758	135	2083
		elisabethseton	12990	172	2167
		rosaliedeconstant	12688	143	1944
	nouvelle	mademoisellefifi 1	10893	79	1367
		mademoisellefifi 2	9812	79	1316
		mademoisellefifi 3	11624	87	1621
		morteamoureuse	14270	137	2122
		sarrasine	15346	119	1937
		bouvardetpecuchet	12241	103	1437
		capitainefracasse	11817	150	1235
	roman	pauline	12284	89	1620
		ventredeparis	12501	140	1572
		delavilleaumoulin	11945	133	2182
		diablaucorps	14596	154	2478
		doucelumiere	12268	110	1856
		jeanchristophe 1	10962	154	1951
		jeanchristophe 2	10303	96	1834
		nemoville	12378	130	2212
Total DemRC{Nar+}			220676	2210	32934
Total DemRC			347155	4235	47821

TABLEAU 4.2 – Composition de la ressource DEMOCRAT_19^e-21^e

Pour pallier les limites que nous avons rencontrées pour le format `xml` AnnoDis (cf. 4.1.1), nous proposons un encodage qui suit les recommandations du consortium TEI (version P5). Certains attributs non définis par la TEI ont été ajoutés. Une description complète de ces attributs est disponible dans la documentation mise à disposition avec la ressource DEMRC (annexe C). Afin d’assurer la compatibilité avec le corpus DEMOCRAT d’origine les `xml:id` des chaînes et des maillons correspondent aux identifiants attribués à ces éléments dans les fichiers DEMOCRAT; les `id` des tokens correspondent aux identifiants du corpus DEMOCRAT et permettent de récupérer les informations relatives aux tokens dans les fichiers source tagués par TreeTagger.

Dans les 29 fichiers DEMOCRAT pris en compte, la structure textuelle est signalée de façon minimale. Nous avons reproduit les marques de saut de ligne `<lb>` parce que cet élément pourrait aider à identifier des titres et/ou des paragraphes. Ci-dessous la liste des éléments que nous avons utilisés et leurs attributs :

- `<teiHeader>` : contient les métadonnées relatives au texte (titre, auteur, date de publication), aux contributeurs, à la licence et aux informations relatives au nombre d’éléments présents dans le fichier.
- `<div type="Reference_chain" xml:id=" " from=" " id_firstMention=" " to=" " id_lastMention=" " extent=" ">` : délimite les chaînes de référence. Ses attributs :
 - `xml:id` : indique l’identifiant DEMOCRAT de la CR

- `from` : correspond à l'identifiant DEMOCRAT du premier token du premier maillon
- `id_firstMention` : correspond à l'identifiant DEMOCRAT du premier maillon de la chaîne
- `to` : correspond à l'identifiant DEMOCRAT du dernier token du dernier maillon
- `id_lastMention` : correspond à l'identifiant DEMOCRAT du dernier maillon de la chaîne
- `extent` : correspond au nombre de maillons qui constituent la CR
- `<rs type="mention" xml:id=" " from=" " to=" ">` : délimite les maillons de la chaîne. Ses attributs :
 - `xml:id` : correspond à l'identifiant DEMOCRAT du maillon
 - `from` : indique l'identifiant DEMOCRAT du token de début du maillon
 - `to` : indique l'identifiant DEMOCRAT du token de fin du maillon
- `<s n=" " nbToks=" " nbWords=" " from=" " to=" ">` : délimite les phrases du texte. Rappelons qu'une phrase n'est reproduite que si elle contient au moins un maillon. Ses attributs :
 - `n` : correspond à l'identifiant DEMOCRAT de la phrase ; si deux phrases ont été fusionnées (cf. traitement 1), les deux identifiants sont présents, séparés par une virgule
 - `nbToks` : indique le nombre de tokens de la phrase, y compris la ponctuation
 - `nbWords` : indique le nombre de mots de la phrase sans prendre en compte la ponctuation
 - `from` : correspond à l'identifiant DEMOCRAT du premier token de la phrase
 - `to` : correspond à l'identifiant DEMOCRAT du dernier token de la phrase
 - `nbMentions` : si la phrase ne contient pas de maillons, cet attribut est ajouté à la balise avec la valeur 0
- `<lb n=" ">` : indique les sauts de ligne présents dans les fichiers d'origine. L'attribut `n` est l'identifiant DEMOCRAT de ces sauts de ligne.

En présentant les CR comme des segments de texte avec annotation intégrée, le format proposé facilite une analyse qualitative des CR. Il est en effet possible de revenir aux données pour visualiser et analyser aisément les CR et les maillons. Il permet de prendre en compte une chaîne à la fois, indépendamment des autres, tout en laissant la possibilité de revenir aux annotations originales grâce aux identifiants des chaînes et des maillons que nous avons reproduits tels quels dans les textes et les annotations d'origine.

Les annotations intégrées dans ce format pourront être complétées avec les informations attribuées aux maillons et aux chaînes. Par ailleurs, comme nous avons conservé les identifiants d'origine, une intégration de ces informations au format XML-TEI-URS avec annotation déportée est envisageable. Cela permettrait également d'intégrer les modifications effectuées lors du nettoyage ; sans oublier la possibilité d'ajouter d'autres couches d'annotation. Comme nous le verrons dans la [section 6.1](#), une


```

<div type="Reference_chain" xml:id="s-CHAINE-282" from="w_export_1397" id_firstMention="u-MENTION-400"
to="w_export_1462" id_lastMention="u-MENTION-416" extent="3 mentions">
  <s n="52" nbToks="25" nbWords="21" from="w_export_1397" to="w_export_1421">
    <rs type="mention" xml:id="u-MENTION-400" from="w_export_1397" to="w_export_1401">Sa langue bleue et
    préhensile</rs>
    est la plus puissante, la plus coriace et la plus longue ( 55 cm ) parmi les ongulés..
  </s>
  <s n="53" nbToks="19" nbWords="18" from="w_export_1422" to="w_export_1440">
    Elle peut
    <rs type="mention" xml:id="u-MENTION-405" from="w_export_1424" to="w_export_1424">l'</rs>
    allonger pour atteindre les pousses les plus tendres entre les barrières d'épines d'acacias..
  </s>
  <s n="54" nbToks="12" nbWords="11" from="w_export_1441" to="w_export_1452" nbMentions="0"/>
  <s n="55" nbToks="38" nbWords="35" from="w_export_1453" to="w_export_1490">
    Elle saisit donc les pousses d'acacias avec
    <rs type="mention" xml:id="u-MENTION-416" from="w_export_1461" to="w_export_1462">sa langue</rs>
    , puis les guide entre ses lèvres, referme la bouche et tire la tête en arrière pour racler les feuilles
    grâce à ses dents du bas.
    <lb n="21"/>
  </s>
</div>
<div type="Reference_chain" xml:id="s-CHAINE-285" from="w_export_1428" id_firstMention="u-MENTION-406"
to="w_export_1465" id_lastMention="u-MENTION-418" extent="3 mentions">
  <s n="53" nbToks="19" nbWords="18" from="w_export_1422" to="w_export_1440">
    Elle peut l'allonger pour atteindre
    <rs type="mention" xml:id="u-MENTION-406" from="w_export_1428" to="w_export_1432">les pousses les plus
    tendres</rs>
    entre les barrières d'épines d'acacias..
  </s>
  <s n="54" nbToks="12" nbWords="11" from="w_export_1441" to="w_export_1452" nbMentions="0"/>
  <s n="55" nbToks="38" nbWords="35" from="w_export_1453" to="w_export_1490">
    Elle saisit donc
    <rs type="mention" xml:id="u-MENTION-414" from="w_export_1456" to="w_export_1459">les pousses d'acacias</rs>
    avec sa langue, puis
    <rs type="mention" xml:id="u-MENTION-418" from="w_export_1465" to="w_export_1465">les</rs>
    guide entre ses lèvres, referme la bouche et tire la tête en arrière pour racler les feuilles grâce à ses
    dents du bas.
    <lb n="21"/>
  </s>
</div>

```

FIGURE 4.3 – Aperçu des chaînes de référence dans DEMRC : format d'analyse avec annotation intégrée

version augmentée avec des annotations sur la nature du référent est réalisée, dans les deux formats : `xml` avec annotation intégrée et avec annotation déportée, au format XML-TEI-URS.

4.2.3 Pré-traitement de la ressource DEMRC

À partir du format d'analyse, nous avons traité la ressource DemRC avec Stanza (Qi et al., 2020). Nous avons obtenu ainsi l'analyse morpho-syntaxique, les dépendances syntaxiques et les entités nommées. L'analyse morpho-syntaxique est utile pour l'attribution de la nature des maillons et de leur fonction syntaxique. Concernant les entités nommées, une fois attribuées aux maillons, elles peuvent permettre d'identifier le type de référent.

Nous avons utilisé les modèles Stanza pour le français. Le modèle utilisé pour la reconnaissance des entités nommées prend en charge quatre types : PER (*Person*), LOC (*Location*), ORG (*Organization*) et MISC (*Miscellaneous*). MISC signale qu'une entité nommée a été détectée par le système, mais sa catégorie d'appartenance n'est pas déterminée. Nous proposons un fichier distinct pour chaque chaîne de référence. Ainsi, on obtient autant de fichiers que de CR analysées. Chaque fichier est nommé avec l'identifiant DEMOCRAT des chaînes. Une dernière différence concerne l'indication des maillons. Nous proposons d'indiquer l'inclusion d'un token dans un maillon dans une colonne supplémentaire en suivant le codage BIOS utilisé pour les entités nommées :

- **S-mention** : indique que le token correspond à un maillon complet.
- **B-mention** : indique que le token est le début d'un maillon. Lorsque le maillon commence par un déterminant dont la forme est contractée, deux **B-mention** se succèdent.
- **I-mention** : indique que le token est dans un maillon contenant plusieurs tokens (sans en être le premier ou le dernier).
- **E-mention** : indique que le token est le dernier d'un maillon contenant plusieurs tokens.

À la colonne signalant les maillons suit une autre colonne indiquant leurs identifiants DEMOCRAT. En gardant les identifiants DEMOCRAT des CR et des maillons, nous rendons possible le retour au corpus d'origine. À titre d'exemple, nous proposons ci-dessous un extrait du fichier *Stanza* correspondant à la chaîne 285 du fichier *articleswiki* (*articleswiki_s-CHAINE-285*).

1 Elle	il	PRON	Gender=Fem Number=Sing Person=3 PronType=Prs	2 peut	nsbj	O	115	119		
2 peut	pouvoir	VERB	Mood=Ind Number=Sing Person=3 Tense=Pres VerbForm=Fin	0 root	root	O	120	124		
3 l'	le	PRON	Number=Sing Person=3 PronType=Prs	4 allonger	obj	O	125	127		
4 allonger	allonger	VERB	VerbForm=Inf	2 peut	xcomp:obj	O	127	135		
5 pour	pour	ADP		6 atteindre	mark	O	136	140		
6 atteindre	atteindre	VERB	VerbForm=Inf	4 allonger	advcl	O	141	150		
7 les	le	DET	Definite=Def Number=Plur PronType=Art	8 pousse	det	O	151	154 B-mention	u-MENTION-406	
8 pousse	pousse	NOUN	Gender=Fem Number=Plur	6 atteindre	obj	O	155	162 I-mention	u-MENTION-406	
9 les	le	DET	Definite=Def Number=Plur PronType=Art	11 tendres	advmod	O	163	166 I-mention	u-MENTION-406	
10 plus	plus	ADV		9 les	fixed	O	167	171 I-mention	u-MENTION-406	
11 tendres	tendre	ADJ	Gender=Fem Number=Plur	8 pousse	amod	O	172	179 E-mention	u-MENTION-406	
12 entre	entre	ADP		14 barrières	case	O	180	185		
13 les	le	DET	Definite=Def Number=Plur PronType=Art	14 barrières	det	O	186	189		
14 barrières	barrière	NOUN	Gender=Fem Number=Plur	8 pousse	nmod	O	190	199		
15 d'	de	ADP		16 épines	case	O	200	202		
16 épines	épine	NOUN	Gender=Fem Number=Plur	14 barrières	nmod	O	202	208		
17 d'	de	ADP		18 acacias	case	O	209	211		
18 acacias	acacia	NOUN	Gender=Fem Number=Plur	16 épines	nmod	O	211	218		
19 .	.	PUNCT		2 peut	punct	O	218	220		
1 La	le	DET	Definite=Def Gender=Fem Number=Sing PronType=Art	2 girafe	det	B-MISC	221	223		
2 girafe	girafe	NOUN	Gender=Fem Number=Sing	4 a	nsbj	E-MISC	224	230		
3 n'	ne	ADV	Polarity=Neg	4 a	advmod	O	231	233		
4 a	avoir	VERB	Mood=Ind Number=Sing Person=3 Tense=Pres VerbForm=Fin	0 root	root	O	233	234		
5 pas	pas	ADV	Polarity=Neg	4 a	advmod	O	235	238		
6 d'	un	DET	Definite=Ind Number=Plur PronType=Art	7 incisives	det	O	239	241		
7 incisives	incisive	NOUN	Gender=Fem Number=Plur	4 a	obj	O	241	250		
8 à	à	ADP		10 mâchoire	case	O	251	252		
9 à	le	DET	Definite=Def Gender=Fem Number=Sing PronType=Art	10 mâchoire	det	O	253	255		
10 mâchoire	mâchoire	NOUN	Gender=Fem Number=Sing	4 a	obl:arg	O	256	264		
11 supérieure	supérieur	ADJ	Gender=Fem Number=Sing	10 mâchoire	amod	O	265	275		
12 .	.	PUNCT		4 a	punct	O	275	276		
1 Elle	il	PRON	Gender=Fem Number=Sing Person=3 PronType=Prs	2 saisit	nsbj	O	277	281		
2 saisit	saisir	VERB	Mood=Ind Number=Sing Person=3 Tense=Pres VerbForm=Fin	0 root	root	O	282	288		
3 donc	donc	ADV		2 saisit	advmod	O	289	293		
4 les	le	DET	Definite=Def Number=Plur PronType=Art	5 pousse	det	O	294	297 B-mention	u-MENTION-414	
5 pousse	pousse	NOUN	Gender=Fem Number=Plur	2 saisit	obj	O	298	305 I-mention	u-MENTION-414	
6 d'	de	ADP		7 acacias	case	O	306	308 I-mention	u-MENTION-414	
7 acacias	acacia	NOUN	Gender=Fem Number=Plur	5 pousse	nmod	O	308	315 E-mention	u-MENTION-414	
8 avec	avec	ADP		10 langue	case	O	316	320		
9 sa	son	DET	Gender=Fem Number=Sing PossNumber=Sing PossPerson=3 PronType=Prs	10 langue	det	O	321	323		
10 langue	langue	NOUN	Gender=Fem Number=Sing	2 saisit	obl:mod	O	324	330		
11 .	.	PUNCT		19 referme	punct	O	330	331		
12 puis	puis	ADV		19 referme	cc	O	332	336		
13 les	le	DET	Definite=Def Number=Plur PronType=Art	14 guide	det	O	337	340 S-mention	u-MENTION-418	
14 guide	guide	NOUN	Gender=Masc Number=Sing	19 referme	nsbj	O	341	346		
15 entre	entre	ADP		17 lèvres	case	O	347	352		
16 ses	son	DET	Number=Plur PossNumber=Sing PossPerson=3 PronType=Prs	17 lèvres	det	O	353	356		

FIGURE 4.4 – Extrait du fichier *articleswiki_RC_s-CHAINE-285.csv*

Les colonnes correspondent dans l'ordre à :

- la forme ;
- le lemme ;
- la catégorie grammaticale (*Universal Dependencies tagset*) ;
- les informations morpho-syntaxiques ;
- l'index du gouverneur ;
- la forme du gouverneur ;
- la fonction syntaxique (*Universal Dependencies tagset*) ;

- h) le type d'entité nommée, le cas échéant ;
- i) l'*offset* de début du token ;
- j) l'*offset* de fin du token ;
- k) l'annotation des maillons, le cas échéant ;
- l) l'identifiant DEMOCRAT dumaillon, le cas échéant.

Un traitement de même type a été réalisé par Wilkens et al. (2020), ce qui a donné lieu à la diffusion de fichiers similaires avec ajout d'informations sur les chaînes et les mentions⁶. Cependant, les identifiants des CR, des maillons, des tokens et des phrases ont été attribués indépendamment des textes d'origine, ce qui ne facilite pas les allers-retours avec les annotations d'origine.

Dans la suite de notre travail, nous utilisons les sorties obtenues pour extraire les informations sur les critères de description, comme la nature des maillons, leur fonction syntaxique ou le type d'entité nommée. Nous nous appuyons également sur le *format d'analyse*, précédemment mentionné.

4.3 Anatomie des chaînes dans les deux ressources exploitées : AnnoDis vs DEMRC

Dans cette section, nous présentons l'anatomie des CR présentes dans les deux ressources exploitées : AnnoDis et DEMRC. Cette analyse consiste en une description des chaînes selon une approche globale. Ce type de description permet de rendre compte de la dimension des CR et de leur composition globale. Nos calculs ont porté sur trois critères parmi ceux que nous avons présentés dans le chapitre 2. Nous prenons soin d'explicitier quel type de calcul nous avons utilisé.

Chaque section correspond à un critère. Pour chaque critère retenu, nous fournissons, dans un premier temps, une description des CR et interprétons les résultats en comparant les deux ressources. Dans un deuxième temps, nous discutons des variations observées selon les facteurs suivants : type de texte et genre textuel. Pour tous les résultats présentés, lorsque cela est possible, nous proposons des comparaisons avec les résultats décrits dans la littérature (cf. chapitre 3).

Le tableau 4.3 illustre les statistiques descriptives concernant la longueur des CR et la distance entre les maillons dans les deux ressources. Nous y présentons les résultats en distinguant les types de textes et les genres textuels. Nous fournissons également les statistiques globales par type de texte, ainsi que les résultats pour les ressources dans leur ensemble. Les colonnes représentent dans l'ordre : le type de texte, le genre textuel, la longueur maximale, la longueur moyenne, l'écart-type⁷ (std.) et la médiane.

6. Les fichiers sont disponibles à l'adresse suivante : https://github.com/boberle/coreference_databases.

7. Nous rappelons que l'écart-type permet de décrire la dispersion des données autour de la moyenne. Une valeur proche de zéro indique que les données ne sont pas très dispersées autour de la moyenne. Plus la valeur est élevée, plus les données sont dispersées autour de la moyenne, ce qui signifie qu'il y a beaucoup de variance dans les données observées.

4.3.1 Longueur des chaînes

La longueur des CR est calculée en nombre de maillons. Nous avons calculé la longueur moyenne, médiane et l'écart type pour chaque ressource et pour chaque genre textuel qui y est représenté. Pour une meilleure visualisation des données, la description du [tableau 4.3](#) est accompagnée de boîtes à moustaches qui permettent de visualiser la distribution, la dispersion ainsi que la présence d'éventuels individus atypiques (*outliers*) dans les données. Les lignes à l'intérieur de chaque boîte indiquent la médiane, les losanges indiquent l'écart-type, la ligne pointillée à l'intérieur des losanges indique la moyenne, et les ronds les individus atypiques, c'est-à-dire des CR très longues qui tirent la moyenne vers le haut.

Dans AnnoDis, les CR sont globalement courtes, avec une moyenne globale de 6 maillons, un écart-type de 4,86, et une médiane de 4. Cela suggère une distribution relativement homogène autour de la moyenne. Les articles encyclopédiques ($AE_{\{W\}}$ dans [figure 4.5](#)) présentent les chaînes les plus longues dans cette ressource (moyenne de 7 maillons, médiane de 5 maillons, std. = 6,19). Nous pouvons observer que la présence d'individus atypiques ([figure 4.5](#)) tend à tirer la moyenne vers le haut, avec une longueur maximale de 40 maillons. À l'inverse, les articles scientifiques en géopolitique ($AS_{\{G\}}$) présentent les chaînes les plus courtes, avec une moyenne de 5 maillons et une faible dispersion. Les articles de linguistique se situent entre les deux. En comparai-

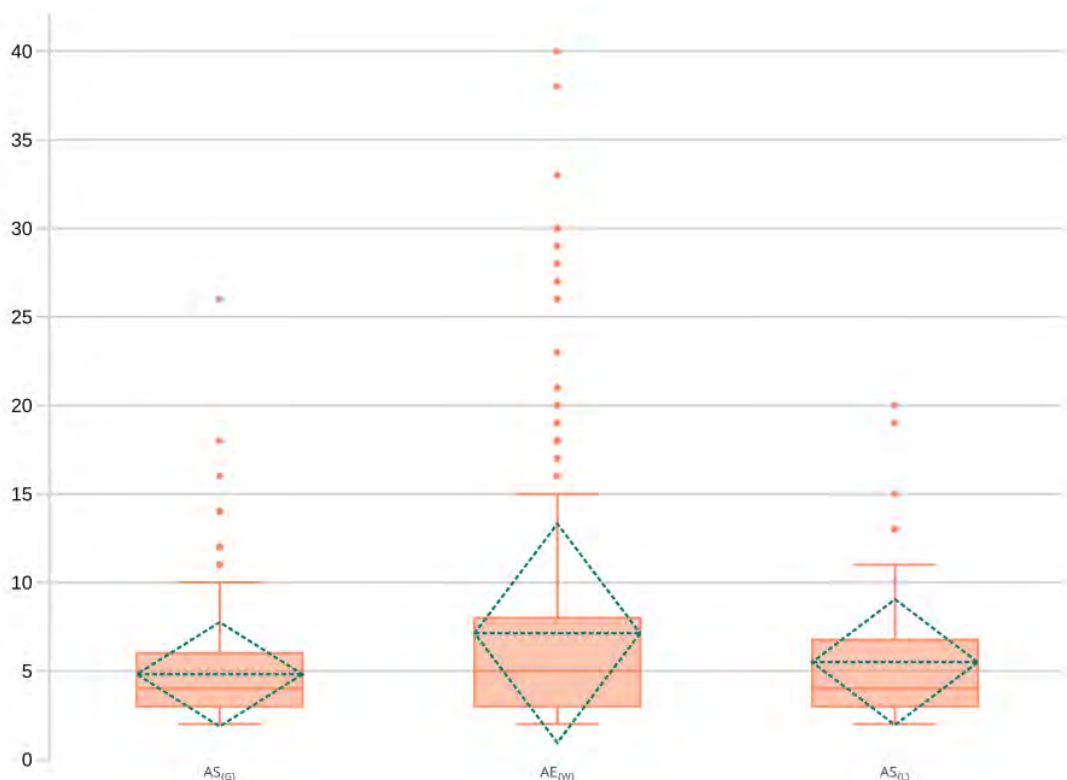


FIGURE 4.5 – Longueur des chaînes dans AnnoDis en fonction du genre textuel

son, la ressource DEMRC, qui contient à la fois des textes narratifs ($DEMRC_{\{Nar+\}}$) et non narratifs ($DEMRC_{\{Nar-\}}$), présente des CR plus longues et plus dispersées.

Globalement, les CR sont constituées en moyenne de 11 maillons, presque le double d'AnnoDis, avec un écart-type élevé (std. = 40,76), qui traduit une importante variabilité dans la longueur des chaînes. En considérant les textes non narratifs de DEMRC, nous observons que la longueur des CR reste plus proche des valeurs d'AnnoDis, avec une moyenne de 7 maillons. Toutefois, la dispersion est importante (std. = 11,74), et les chaînes peuvent atteindre jusqu'à 211 maillons. Dans les textes non narratifs de DEMRC, les articles de presse présentent les chaînes les plus courtes (moyenne 5,3, std. = 5,11), ce qui n'est pas surprenant si l'on considère leur structure et la fréquence de renouvellement des référents. En revanche, parmi les textes non narratifs, dans le pamphlet on trouve, en moyenne, les chaînes les plus longues (8,7 maillons, std. = 15,8), indiquant une plus grande dispersion.

Sans surprise, les CR dans les textes narratifs sont en moyenne beaucoup plus longues que dans les textes non narratifs (cf. 2.3.2). La moyenne atteint 15 maillons pour les textes narratifs (contre 7 pour les non narratifs), avec des écart-types très élevés (std. > 48), ce qui indique une forte dispersion autour de la moyenne. Cette variabilité est due notamment à la présence d'individus atypiques (figure 4.6) : des CR très longues qui tirent la moyenne vers le haut, comme celle du texte *Le Diable au corps* qui compte à elle seule 895 maillons. Le référent (*Je*) de la CR (disponible dans l'annexe B.1) correspond au narrateur, dont les mentions sont réparties de façon régulière tout au long du texte — une régularité qui illustre la centralité du référent (voir figure 4.8). Dans les textes narratifs, ce sont les nouvelles qui présentent les chaînes les plus longues (17 maillons) et une dispersion très importante (std. = 55,1), suivies des romans (std. = 14,6) qui présentent néanmoins une plus forte dispersion (figure 4.7), et des biographies (std. = 13,76). Ces chiffres confirment que les genres narratifs favorisent la mise en place de CR centrées autour d'un ou de quelques référents principaux.

Pour conclure, les résultats présentés confirment que les chaînes de référence sont, en moyenne, plus longues dans les textes narratifs que dans les textes non narratifs. Ces résultats sont en accord avec les observations que nous avons proposées dans Federzoni et al. (2020), en comparant les résultats obtenus pour AnnoDis aux résultats décrits dans les travaux existants, notamment ceux de Longo (2013) ; Longo et Todorascu (2014) ; Schnedecker et Landragin (2014) ; Todorascu et al. (2017). Malgré la différence sur les moyennes, la médiane reste étonnamment stable à 4 maillons, quels que soient les types et genres textuels (à l'exception des articles encyclopédiques d'AnnoDis). Cette constance suggère que, dans la majorité des cas, les chaînes sont courtes, et que ce sont quelques chaînes très longues — principalement dans les textes narratifs — qui tirent la moyenne vers le haut.

Ces constats amènent à une mise en perspective du rôle du modèle d'annotation. En comparant AnnoDis et les textes non narratifs de DEMRC, nous nous attendions à observer une forte différence sur la longueur des CR en raison des modèles d'annotation différents. Nous nous attendions à ce que les CR AnnoDis soient plus courtes que celles de DEMRC (cf. 3.4). Les résultats ont montré des différences notables sur la longueur maximale des CR (respectivement 40 et 211 maillons) et, dans une moindre mesure, sur les écarts-types. Cependant, les moyennes et les médianes restent comparables entre les deux ressources, ce qui tend à montrer que le type de texte reste le

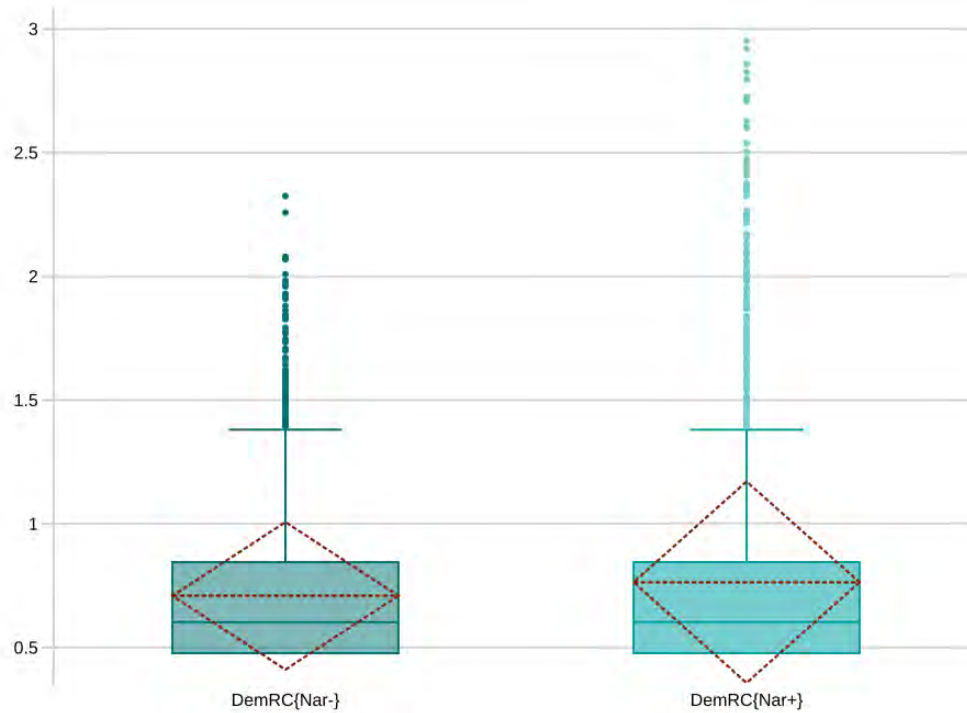


FIGURE 4.6 – Longueur des chaînes dans DEMRC en fonction du type de texte

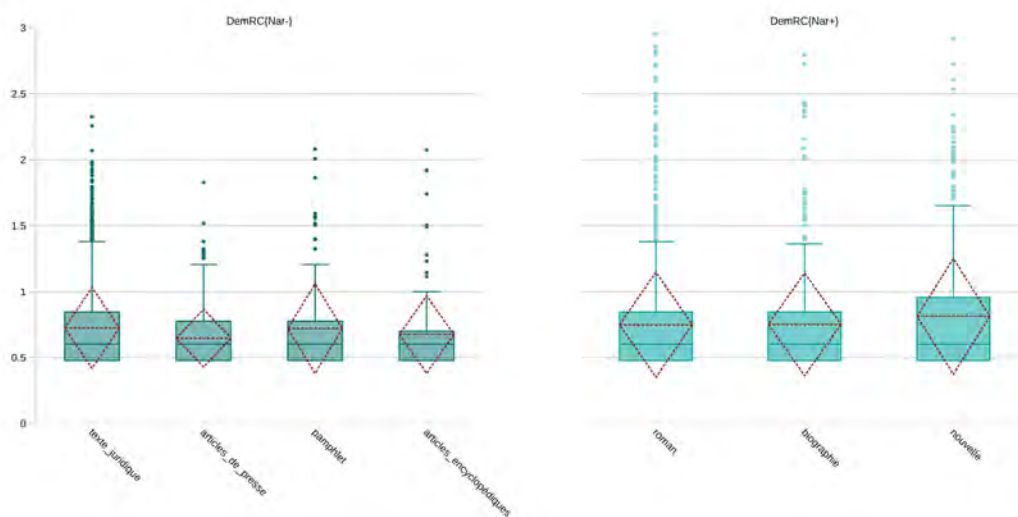


FIGURE 4.7 – Longueur des chaînes dans DEMRC en fonction du genre textuel

facteur déterminant, davantage que le modèle d'annotation. Autrement dit, le modèle d'annotation peut avoir un effet ponctuel (sur certains extrêmes), mais il n'a pas un

impact direct sur la longueur moyenne des CR.

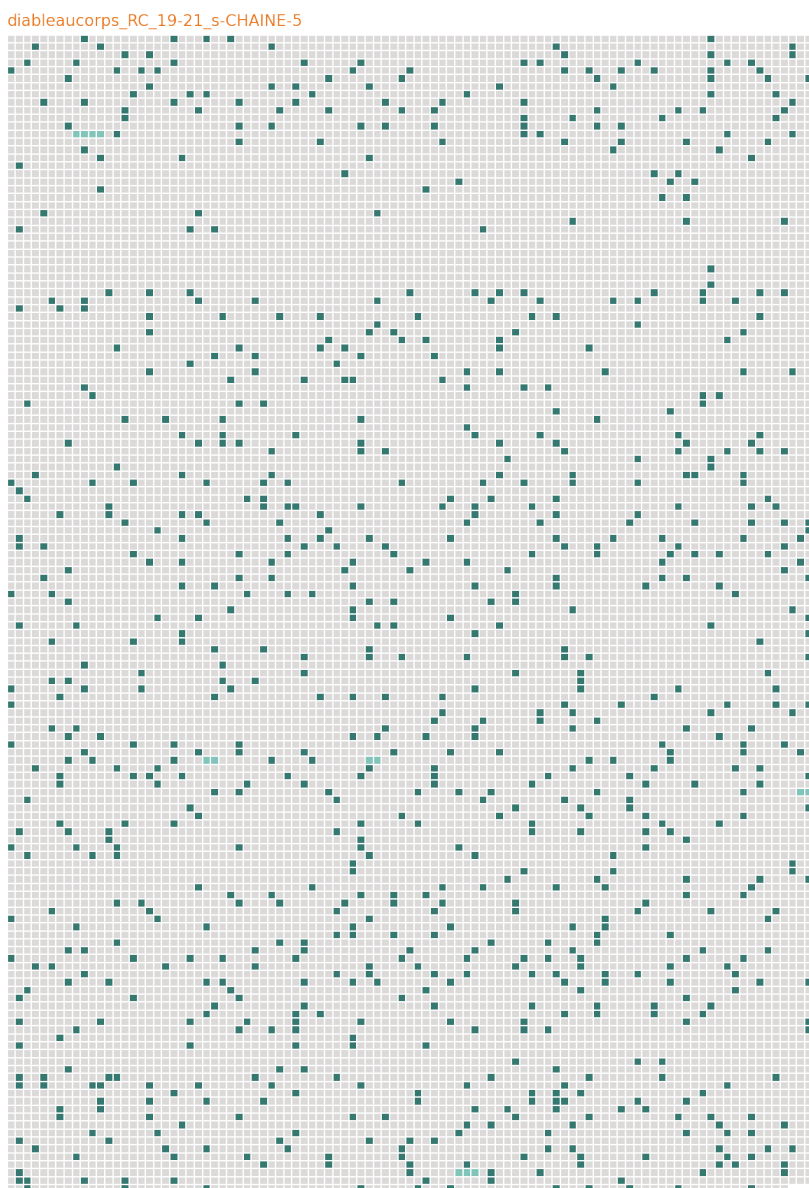


FIGURE 4.8 – Visualisation de la chaîne constituée de 895 maillons (*s-CHaine-5*) dans le texte *Le diable au corps* (sous-corpus des *romans*)

4.3.2 Distance entre les maillons

L'analyse porte ici sur la distance moyenne entre les maillons des chaînes de référence, mesurée en nombre de mots. Le calcul commence au premier mot qui suit le token de fin du maillon et s'arrête au mot qui précède le token de début du maillon suivant. La ponctuation n'est pas prise en considération dans le calcul (cf. 2.2.3). Le [tableau 4.3](#) illustre les statistiques descriptives concernant la distance des CR dans les deux ressources.

Dans le corpus AnnoDis, qui regroupe exclusivement des textes non narratifs, la moyenne des distances varie de 28 mots — dans les articles encyclopédiques — à

AnnoDis	Longueur CR				Distance			
Genre	Max.	Moy.	Std.	Méd.	Max.	Moy.	Std.	Méd.
articles scientifiques _{Geop}	26	4,8	2,95	4	112	30,20	17,23	26
articles scientifiques _{Ling}	20	5,49	3,57	4	101	32,76	17,88	28
articles encyclopédiques _{Wiki}	40	7,12	6,19	5	92	27,56	13,79	24,5
Total AnnoDis	40	5,94	4,86	4	112	29,4	16	25,6
DEMRC_{Nar-}								
Genre	Max.	Moy.	Std.	Méd.	Max.	Moy.	Std.	Méd.
articles de presse	67	5,29	5,11	4	4611	202,4	594,3	35,5
articles encyclopédiques	118	7,24	14,05	4	1360	224	305,7	76
pamphlet	120	8,68	15,83	4	3848,5	261,4	637,5	11
texte juridique	211	7,70	12,11	4	5430,5	456,8	759,7	141
Total DEMRC_{Nar-}	211	7,35	11,74	4	5430,5	390,3	716,5	88
DEMRC_{Nar+}								
biographie	625	13,76	48,82	4	4445,5	442,5	749,7	47,3
nouvelle	829	16,69	55,15	4	6088	344,8	692,6	45,3
roman	895	14,60	57,09	4	4953,5	386,9	668,9	67,4
Total DEMRC_{Nar+}	895	14,90	55,05	4	6088	388,7	691,9	59
Total DEMRC	895	11,29	40,76	4	6088	389,4	703,7	74,5

TABLEAU 4.3 – Statistiques descriptives des deux ressources exploitées

33 mots dans les articles scientifiques en linguistique. Les médianes sont proches des moyennes (entre 24 et 28), ce qui indique que dans la moitié des chaînes, les mentions sont rapprochées. Les écarts-types, compris entre 13,8 et 17,9, confirment une faible dispersion autour de la moyenne. Cela pourrait indiquer que les CR de cette ressource sont des CR plutôt compactes, où la mention des référents est régulière.

Dans le corpus DEMRC, qui combine des textes narratifs (DEMRC_{Nar+}) et non narratifs (DEMRC_{Nar-}), les résultats varient fortement selon le type de texte. Dans les textes narratifs, les distances entre les maillons sont nettement plus élevées. La moyenne dépasse 340 mots dans tous les genres. Cependant, les médianes restent très basses (entre 45 et 67 mots) avec un fort écart-type. Cela indique que certaines CR présentent des maillons très distants (comme illustré dans (17)), tirant la moyenne vers le haut. Cette forte distance entre les maillons pourrait signifier que dans les textes narratifs certaines CR sont plus éparpillées. Pour confirmer cette interprétation, il serait nécessaire de croiser ces résultats avec d'autres critères, notamment la longueur et la couverture, tout en analysant les individus atypiques à part.

- (17) La flamme éclaire la figure rouge du vieux Jean-Michel, **sa barbe blanche et rude**, son air bourru et ses yeux vifs.

[2 269 mots]

Il vint près d'elle, lui effleura le front de **sa barbe râpeuse** ; il demanda si elle n'avait besoin de rien, baissa la lumière de la lampe, et partit en heurtant les chaises, dans l'obscurité de la chambre.

[581 mots]

Pendant ce temps, le vieux Jean-Michel attendait devant la maison, sous la pluie, **la barbe mouillée de brouillard** [R_{JC} ; 

Dans les textes non narratifs de DEMRC les moyennes des distances varient entre 202 mots (articles de presse) et 456 mots (textes juridiques), mais les médianes sont systématiquement plus faibles (entre 35 et 141). L'écart-type des textes juridiques atteint des valeurs proches à celles observées dans les textes narratifs.

De manière générale, ces résultats montrent que dans les textes narratifs, les CR présentent une distance entre les maillons plus importante que celle des textes non narratifs. En comparant les CR des textes non narratifs de DEMRC à celles d'AnnoDis, on observe des différences : les distances sont plus élevées dans DEMRC, notamment pour les articles encyclopédiques. Cela suggère que les écarts pourraient être liés aux modèles d'annotation.

Des différences significatives entre les types de texte ?

Afin de déterminer si les différences observées entre textes narratifs et non narratifs sont statistiquement significatives, nous avons effectué un test de Mann-Whitney sur les longueurs moyennes des chaînes de référence (en nombre de maillons) ainsi que sur les distances moyennes entre maillons (en nombre de mots). L'ensemble des CR des deux ressources a été pris en compte, en considérant les textes d'AnnoDis comme appartenant au groupe non narratif, au même titre que les textes non narratifs de DEMRC. Les résultats révèlent une différence significative pour la longueur moyenne des chaînes ($p = 0,017$), confirmant que les chaînes présentes dans les textes narratifs sont en moyenne plus longues que celles des textes non narratifs. En revanche, aucune différence significative n'a été observée pour la distance moyenne entre maillons ($p = 0,117$). Cela suggère que, si certaines chaînes narratives peuvent être très éparpillées, cet écart n'est globalement pas significatif entre les textes narratifs et non narratifs. Les résultats concernant la distance entre les maillons diffèrent de ceux mis en évidence par [Rousier-Vercruyssen et Landragin \(2019\)](#), qui avaient observé une inter-distance statistiquement plus élevée dans les textes non narratifs. Cette différence peut être due au nombre de CR analysées et à leur longueur, les auteurs ayant analysé les 60 CR les plus longues du corpus Democrat (30 pour les textes narratifs et 30 pour les textes non narratifs).

4.3.3 Nature des maillons

Concernant la composition des CR, nous nous sommes focalisée sur la nature des maillons. L'attribution de ces catégories a été faite automatiquement sur la base de l'étiquetage morpho-syntaxique. Le [tableau 4.4](#) illustre la distribution de la nature des maillons dans les deux ressources, tout type et genre textuel confondus. Nous observons que les SN_{def} sont relativement plus fréquents dans AnnoDis et les pronoms dans

DEMRC. Nous observons la même tendance pour les Np et les Poss, et les $\emptyset$ qui sont absents dans AnnoDis. Cette différence pourrait s'expliquer par le modèle d'annotation : nous pouvons, en effet, nous attendre à ce que les expressions de haute accessibilité telles les pronoms, les possessifs et encore plus les sujets zéro soient annotées moins systématiquement dans AnnoDis en raison de l'approche descendante et de l'annotation non experte (cf. 3.4). Un autre facteur qui pourrait avoir un impact sur la distribution de la nature des maillons est le type de texte : DEMRC inclut des textes narratifs. Le genre textuel pourrait lui aussi avoir un impact sur la composition des CR (Kunz, 2010 ; Longo, 2013 ; Schnedecker, 2021).

Nature	AnnoDis		DEMRC	
	#	%	#	%
SN_{def}	1198	34,7	11 181	23,4
Pro	1026	29,7	18 566	38,8
Np	442	12,8	2 742	5,7
SN_{dem}	272	7,9	979	2,0
Poss	182	5,3	7 968	16,7
Autre	146	4,2	618	1,3
SN_{ind}	126	3,7	1 641	3,4
SN_{$\emptyset$}	64	1,9	2 730	5,7
$\emptyset$	0	0	1 396	2,9
Total	3456	100	47 821	100

TABLEAU 4.4 – Distribution de la nature des maillons dans AnnoDis et DEMRC

Afin d'évaluer l'influence du type et du genre textuel sur la nature des expressions référentielles, nous avons mené une série de tests du χ^2 d'indépendance sur la distribution des catégories des référents (SN définis, SN indéfinis, pronoms, etc.). Dans toutes ces analyses, les résidus de Pearson associés au test du χ^2 offrent une visualisation fine de la liaison, si elle existe : dans les visualisations proposées, les cellules indiquent le nombre d'occurrences de la nature des maillons en fonction de la ressource, du type de texte, ou des genres textuels. Les astérisques indiquent s'il existe une liaison significative entre la nature du maillon et le facteur analysé (« *, ** ou *** » selon la significativité). Si aucun astérisque n'est présent, cela signifie que la liaison n'est pas significative. La couleur des cellules correspond à la direction de la liaison, positive en rouge et négative en bleu.

Tout d'abord, nous avons comparé les textes AnnoDis aux textes non narratifs dans DEMRC. Le test du χ^2 indique qu'il existe une liaison significative (p-value < 0,001) entre les catégories grammaticales et ces textes. L'analyse des résidus de Pearson (figure 4.9) permet de confirmer que les SN_{def} sont plus fréquents dans DEMRC, y compris si l'on considère seulement les textes non narratifs. Alors que les Poss et les Pro sont plus fréquents dans AnnoDis, contrairement à ce que nous avons observé. Cela semble indiquer que le type de texte aurait plus d'impact que le modèle d'annotation. Nous avons alors effectué le test du χ^2 entre les textes narratifs et non narratifs dans DEMRC : celui-ci confirme une différence significative dans la répartition des catégories (p-value < 0,001). Comme nous pouvons l'observer dans la figure 4.10, les Np, les

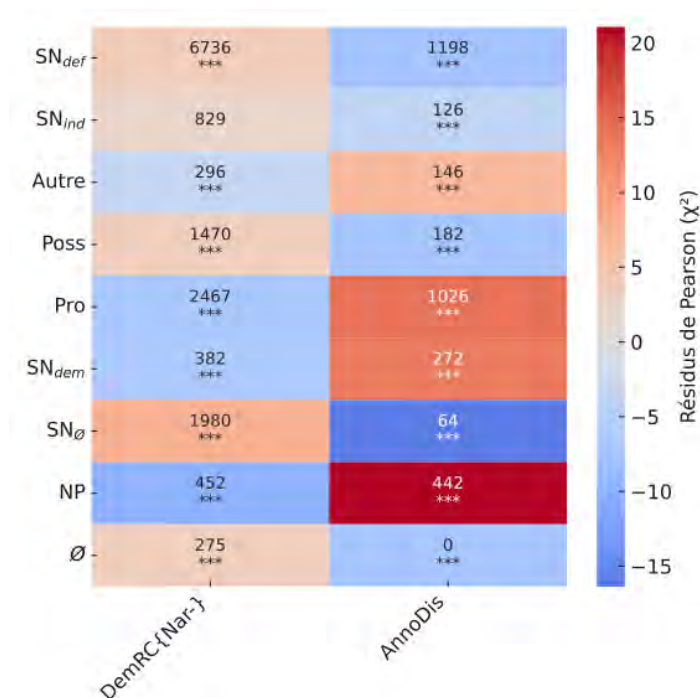


FIGURE 4.9 – Visualisation de la distribution de la nature des maillons dans AnnoDis et DEMRC_{Nar-}

Pro, les Poss et les sujets zéro sont davantage utilisés dans les textes narratifs, alors que les syntagmes nominaux définis, indéfinis, démonstratifs et sans déterminant sont plus fréquents dans les textes non narratifs. Cette opposition se confirme même lorsque nous élargissons le groupe non narratif en intégrant AnnoDis à DEMRC_{Nar-}. Enfin, afin de vérifier si les genres textuels ont un impact, nous avons répété le test pour les genres textuels AnnoDis et DEMRC.

Concernant AnnoDis, nous pouvons observer dans [figure 4.11](#) que les SN définis, démonstratifs et indéfinis sont davantage présents dans les articles scientifiques en géopolitique, alors que les Np sont davantage utilisés dans les articles encyclopédiques. Les pronoms sont moins utilisés dans les articles en géopolitique, mais les autres relations ne sont pas significatives. Passons maintenant à DEMRC. Nous pouvons observer que les SN_{def} sont très fréquents dans les textes juridiques, les articles de presse et les articles encyclopédiques. La forte présence de cette catégorie dans les articles de presse et dans les textes encyclopédiques va dans le même sens que les observations de [Schneidecker et Longo \(2012\)](#) et de [Longo et Todirascu \(2014\)](#). Nous observons également que les Np sont aussi fréquents dans les articles de presse, comme observé par [Schneidecker \(2005\)](#) (le genre étudié par l'autrice est celui des portraits journalistiques). Toutefois, les Np sont davantage fréquents dans les textes narratifs, et notamment dans les romans et les biographies. Cela pourrait s'expliquer par la tendance à instancier des référents humains dans les textes relevant de ces deux genres. Les pronoms et les possessifs sont très fréquents dans les romans, suivis par les nouvelles et les biographies. Les SN sans déterminant sont très largement utilisés dans les textes juridiques.

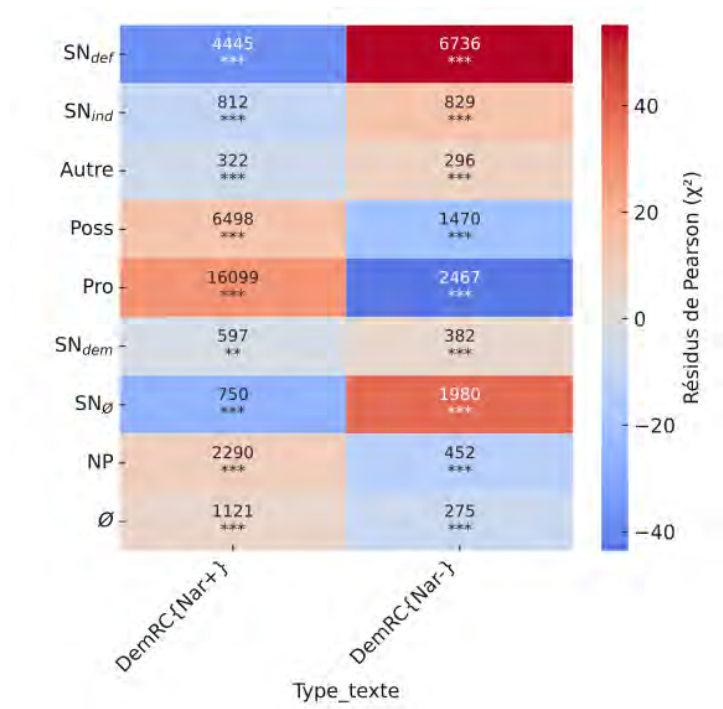


FIGURE 4.10 – Visualisation de la distribution de la nature des maillons dans DEMRC en fonction du type de texte

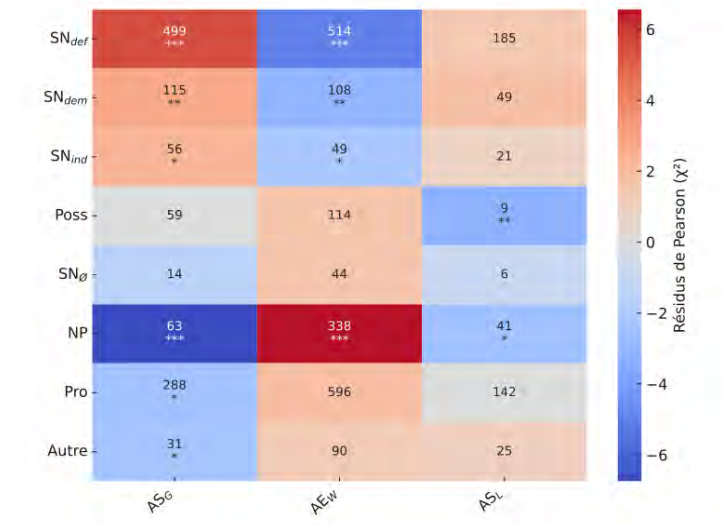


FIGURE 4.11 – Visualisation de la distribution de la nature des maillons dans AnnoDis en fonction du genre textuel

4.4 Bilan

Dans ce chapitre, nous avons présenté la première étape de notre méthodologie : la sélection et la préparation des données en vue de leur exploitation pour l’analyse des chaînes de référence. Nous avons détaillé les caractéristiques des deux ressources exploitées (AnnoDis et DEMRC) ainsi que les traitements spécifiques appliqués à chacune d’elles. Dans ce cadre, nous avons proposé un format d’analyse avec annotations

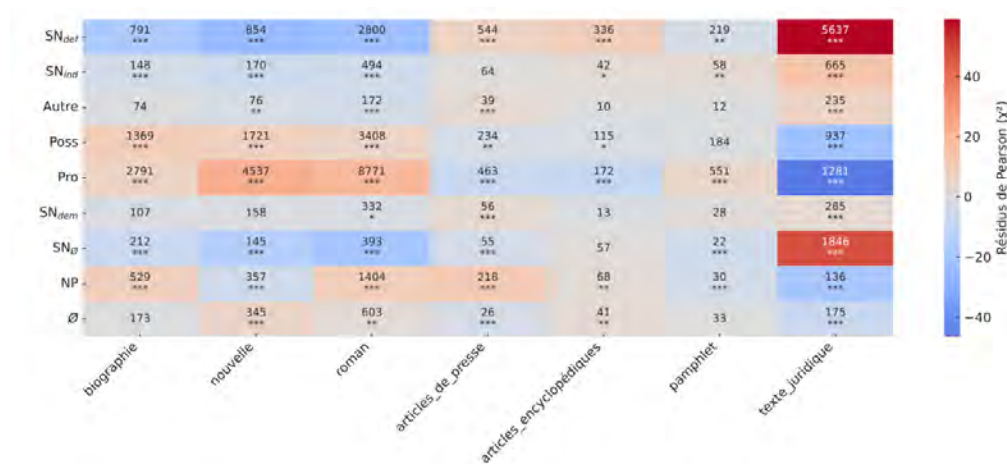


FIGURE 4.12 – Visualisation de la distribution de la nature des maillons dans DEMRC en fonction du genre textuel

intégrées, facilitant aussi bien les traitements automatiques que la visualisation qualitative des données.

Ensuite, nous avons mené une analyse exploratoire des données qui nous a permis de dresser une première anatomie des chaînes de référence : en suivant l’approche globale, nous nous sommes focalisée sur les trois critères de description les plus utilisées pour la description des CR (cf. [chapitre 2](#)), à savoir la longueur des CR, la distance entre les maillons et la nature des maillons.

Tout d’abord, les comparaisons menées entre les deux ressources ont montré, que contrairement à ce à quoi nous nous attendions, le modèle d’annotation n’a pas d’impact sur la longueur des CR. L’impact des modèles semble être plus important sur la distance entre les maillons. Ensuite, les résultats montrent que les textes narratifs tendent à instancier des chaînes plus longues et plus éparpillées, avec une plus forte présence Np et d’expressions de haute accessibilité (pronoms, possessifs et sujets zéro) ; les textes non narratifs présentent des chaînes plus courtes et plus compactes, notamment dans AnnoDis. Dans les textes non narratifs, les CR seraient davantage constituées de SN définis, indéfinis, et sans déterminant.

Ces constats nous ont permis de mieux connaître les CR que nous allons manipuler et soulignent l’importance de croiser plusieurs critères de description pour comprendre le fonctionnement des CR. Nous avons ainsi pu commencer d’identifier des tendances de composition différentes selon différents facteurs. Cependant, ce type de description n’est pas suffisant pour répondre à notre objectif : rendre compte des stratégies d’introduction et de maintien du référent et dégager des patrons des CR. Pour répondre à cet objectif, nous proposons une méthode outillée permettant d’analyser de manière systématique l’enchaînement des maillons.

Chapitre 5

Mise en œuvre de l’approche linéaire via l’analyse des séquences

Contents

5.1	Les chaînes de référence : des <i>séquences</i> de maillons	98
5.1.1	Qu’est-ce que l’analyse des séquences ?	99
5.2	L’analyse des séquences appliquée aux chaînes de référence	100
5.2.1	Création des séquences	101
5.2.2	Exploration et description des séquences	105
5.2.3	Comparaison des séquences	105
5.2.3.1	Mesures basées sur la distribution des états	109
5.2.3.2	Mesures basées sur le nombre d’attributs communs	110
5.2.3.3	Mesures d’édition	112
5.2.4	Classification des séquences	115
5.2.5	Interprétation des clusters obtenus	117
5.3	Des séquences aux patrons	119

Nous proposons de caractériser les CR suivant une approche *linéaire*. Cette approche consiste à rendre compte de « l’enchaînement, au sens littéral du terme, des expressions référentielles » (Schneidecker, 2017), c’est-à-dire comment les maillons se succèdent dans le discours. Décrire systématiquement les enchaînements des expressions référentielles a un double avantage : d’une part, cela permet de s’intéresser aux facteurs qui peuvent influencer le choix des expressions référentielles, et confronter ainsi certaines théories cognitives, comme la théorie de l’accessibilité (Ariel, 2001) et la théorie du centrage (Grosz et al., 1995 ; Walker, Walker, Joshi, Walker, & Prince, 1998) à la réalité des usages en corpus. Décrire les CR sous l’angle des enchaînements permet également de comprendre les stratégies de continuité que réalisent les CR et l’influence de l’organisation globale du discours sur ces stratégies (influence de la structure textuelle ou de la présence simultanée de plusieurs CR). L’objectif d’une telle description est de rendre compte des régularités et des différences qui caractérisent les CR pour dégager des *patrons*, c’est-à-dire des types de composition similaires (Schneidecker, 2021).

À notre connaissance, peu d’études ont proposé des patrons reflétant les modalités d’enchaînement des maillons (Guo, 2021 ; Longo & Todorascu, 2014 ; Schnedecker, 2005) et peu de CR sont prises en compte dans ces travaux. Pour le français oral, Antoine, Lefeuvre, et Schang (2016) ont proposé des patrons à partir d’un nombre plus élevé de CR. Ces études ont mis en évidence des modalités d’enchaînement des maillons (certains se focalisant sur les trois premiers maillons (Longo & Todorascu, 2014)) récurrents et ont montré que ces configurations sont influencées par la nature du référent ou le genre textuel. Ces travaux constituent une base de départ pour notre étude. Nous comparons les résultats obtenus dans ces études aux patrons que nous mettrons en évidence grâce à une méthode outillée permettant de décrire systématiquement les enchaînements des maillons sur un grand nombre de données.

Les sections suivantes portent sur la mise en œuvre de l’approche *linéaire*. Tout d’abord, nous introduisons les besoins auxquels nous devons répondre pour analyser systématiquement les enchaînements des maillons. Nous présentons ensuite les étapes de la mise en œuvre de l’approche *linéaire* et nous discutons de son application aux chaînes de référence.

5.1 Les chaînes de référence : des *séquences de maillons*

Afin de mettre en œuvre l’approche linéaire et décrire l’enchaînement des maillons, nous avons besoin d’une méthode outillée permettant de répondre aux besoins suivants :

- rester dans la linéarité, maintenir l’ordre et la position des maillons dans la chaîne ;
- exploiter les critères permettant de décrire la composition des CR
- garantir un traitement systématique d’une grande quantité de CR ;

Une telle approche doit nous permettre de :

- comparer les chaînes et identifier à la fois les régularités et les différences entre les CR ;
- regrouper les chaînes similaires en formant des groupes qui révèlent des types de composition similaires ;
- interpréter les groupes et comprendre quels éléments sont *représentatifs* des différents types ;
- dégager des patrons pertinents qui refléteraient des types de composition différents ;

Dans cette optique, nous proposons une méthode outillée qui repose sur l’analyse des séquences (cf. 5.1.1). Comme le soulignent Gabadinho, Ritschard, Müller, et Studer (2011), cette analyse exploratoire consiste à trier, comparer et regrouper les séquences similaires. L’objectif principal de ce type d’analyse est de dégager des régularités dans les données séquentielles et d’extraire des informations simplifiées et opérationnelles à partir de ces données (Gabadinho et al., 2011). Un des avantages de ce type d’analyse concerne la possibilité de mener une étude exploratoire des données sans avoir des *a priori*, comme explicité par Aisenbrey et Fasang (2010) :

« Methodologically, sequence analysis provides the possibility to approach data without any distributional assumptions, and its exploratory potential is particularly well suited to analyzing nonstandard and “outlier” life courses. » (Aisenbrey & Fasang, 2010, 450)

Par ailleurs, l'analyse des séquences s'avère être un outil pertinent permettant de visualiser les CR sous forme de *séquences*, facilitant ainsi l'identification des enchaînements les plus fréquents et des patterns récurrents. Enfin, une fois l'exploration des données terminée, la méthode facilite l'étude des relations entre les types identifiés et divers facteurs, dans notre cas les facteurs de variation sélectionnés (comme le type de texte ou la nature du référent).

Dans ce qui suit, nous présentons en quoi consiste l'analyse des séquences et comment nous proposons de l'appliquer à l'étude des chaînes de référence.

5.1.1 Qu'est-ce que l'analyse des séquences ?

L'analyse des séquences, introduite par Abbott (1983) en sciences sociales et généralement utilisée dans cette discipline pour étudier les trajectoires de carrières professionnelles ou biographiques des individus, est définie comme l'étude systématique des séquences (Blanchard, 2019). Une séquence est définie comme une succession d'états ou d'événements. La succession d'états peut donner lieu à des sous-séquences, certaines d'entre elles sont récurrentes et peuvent constituer des *motifs* significatifs (Blanchard, 2019). L'objectif principal de l'analyse des séquences est « d'identifier — dans la diversité d'un corpus de séquences constituées de séries d'états successifs — les régularités, les ressemblances puis le plus souvent de construire des typologies de « séquences-type » » (Robette, 2018). Pour répondre à cet objectif, les recherches qui appliquent l'analyse des séquences suivent un protocole standard de traitement, un *core program* (Blanchard, 2019 ; Blanchard, Bühlmann, & Gauthier, 2014), ce qui garantit une cohérence méthodologique. Ce *core program* est composé de six étapes : (a) la conceptualisation, qui consiste à poser les questions de recherche ; (b) le recueil des données ; (c) la préparation des données, qui comporte le codage des données recueillies sous forme de séquences ; (d) l'exploration des séquences, permettant d'explorer les données à l'aide de visualisations et de décrire la distribution des états, l'entropie, la complexité des séquences (C. Elzinga, 2010) ; (e) la comparaison des séquences par paires, qui s'effectue en calculant la distance entre les paires de séquences ; (f) la synthèse et l'interprétation des résultats de cette comparaison, qui repose sur l'analyse des matrices de distances obtenues lors de l'étape précédente, par l'application de méthodes de classification (clustering) ou de *multidimensional scaling*.

Parmi ces six étapes, nous assimilons la première à la définition de notre problématique (cf. II) et la deuxième aux traitements (préparation et enrichissement) des ressources annotées (cf. chapitre 4). Dans ce qui suit, nous présentons les quatre dernières étapes, que nous considérons comme spécifiques à l'analyse des séquences. Pour appliquer l'analyse des séquences à l'étude des chaînes de référence, nous allons suivre les étapes typiques du *core program*, tout en constatant que des adaptations sont nécessaires. Dans ce qui suit, nous détaillons ces différentes étapes et les ajustements que nous proposons pour l'étude des CR.

5.2 L'analyse des séquences appliquée aux chaînes de référence

La [figure 5.1](#) illustre le déroulement de notre démarche, en mettant en correspondance chaque étape avec son équivalent dans le *core program* (indiqué par une lettre minuscule sur le côté). Comme illustré dans cette figure, la première étape consiste à coder nos données sous forme de séquences, ce qui correspond à l'étape (c) du *core program*. Nous poursuivons donc avec l'exploration et la comparaison des séquences, avant d'effectuer une classification et d'interpréter les classes obtenues. Enfin, nous ajoutons une étape spécifique à notre étude : l'identification de patrons de CR. Cette dernière, propre à l'analyse des chaînes, ne trouve pas d'équivalent direct dans le *core program*.

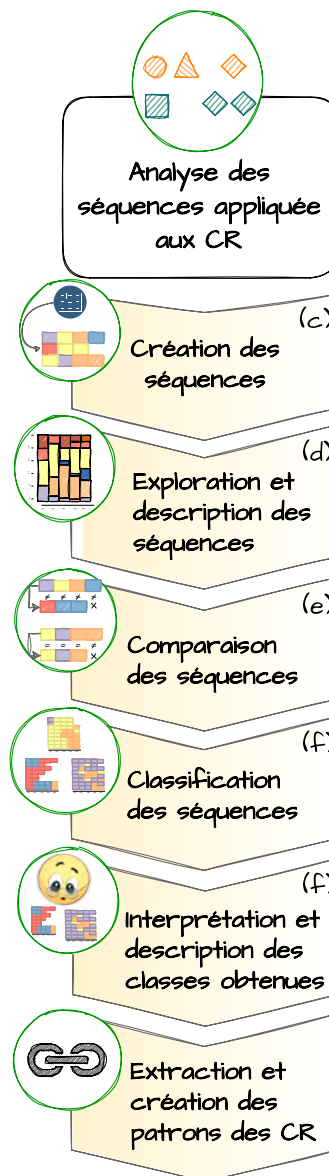


FIGURE 5.1 – Les étapes de l'analyse des séquences appliquée aux chaînes de référence

Pour mener à bien l'analyse des séquences, nous utilisons TraMineR ([Gabadinho, Ritschard, Studer, & Muller, 2010](#))¹. Cette boîte à outil, disponible sur R, offre la possibilité d'implémenter des méthodes statistiques, des méthodes de visualisation spécifiques aux séquences et des méthodes de clustering dans un seul et même environnement. Cela permet de garantir une homogénéité dans les traitements des données et dans le rendu visuel des graphiques ([Gabadinho et al., 2011](#)). TraMineR offre également la possibilité de traiter les séquences sous des formats différents, selon la nature des données et la manière dont les informations sont organisées. Pour une description détaillée des formats disponibles, voir [Gabadinho, Ritschard, Studer, et Muller \(2010\)](#).

5.2.1 Création des séquences

Une séquence est définie par [Brzinsky-Fay, Kohler, et Luniak \(2006\)](#) comme une liste ordonnée d'états, d'objets ou d'événements :

« A sequence is defined as an ordered list of elements, where an element can be a certain status (e.g., employment or marital status), a physical object (e.g., base pair of DNA, protein, or enzyme), or an event (e.g., a dance step or bird call). The positions of the elements are fixed and ordered by elapsed time or by another more or less natural order » ([Brzinsky-Fay et al., 2006](#), 435).



²En suivant cette définition, nous considérons les chaînes comme des séquences d'états. Chaque *état* correspond à un maillon caractérisé par un critère de description, comme illustré schématiquement dans la [figure 5.2](#), où chaque état équivaut à la nature du maillon (cf. [2.1.1](#)). Le même principe s'applique à tout critère de description utilisé, qu'il s'agisse de la fonction syntaxique ou du type de reprise, par exemple. Cependant, nous avons choisi d'adopter dès à présent la nature du maillon pour illustrer nos propos et présenter les exemples à venir, puisque c'est sur ce critère que reposent les résultats que nous allons examiner en détail. Nous aborderons brièvement les autres critères dans la discussion finale.

L'analyse descriptive des séquences est généralement utilisée en sciences sociales pour l'analyse de données longitudinales, afin d'étudier, par exemple, l'évolution des carrières professionnelles des individus sur une fenêtre temporelle donnée. Dans ces travaux, chaque état est relevé à un même moment pour l'ensemble des individus (séquences). Chaque état représente la situation professionnelle d'un individu à une date donnée ou sur une période déterminée. Pour illustrer, considérons les années 2019 et 2020. Supposons que nous examinons deux individus, notés « A » et « B ». Le premier était étudiant (E) durant ces deux années, tandis que le deuxième était au chômage (Ch) en 2019 et employé en CDD (Emp.CDD) en 2020. Donc, les états relevés pour

1. Pour les autres boîtes à outils développées pour l'analyse des séquences, voir [Liao et al. \(2022\)](#).

2. Dans ce chapitre, nous employons cette notation chaque fois que nous montrons comment l'analyse des séquences s'applique aux chaînes de référence et clarifions la signification de certains aspects dans ce contexte. Cela nous permet d'expliquer notre interprétation, d'illustrer nos propos par des exemples parlants pour le lecteur et de mettre en évidence les adaptations nécessaires à l'étude des CR.

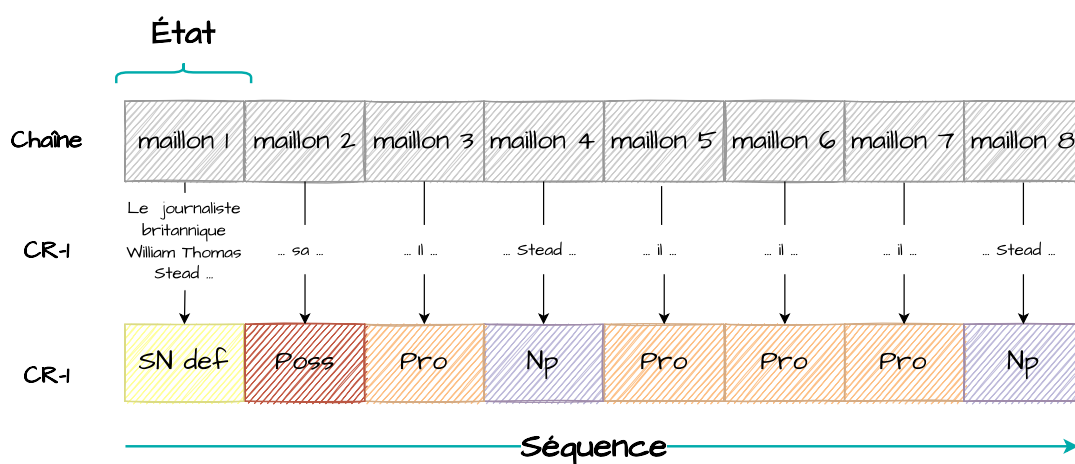


FIGURE 5.2 – Une chaîne de référence en tant que séquence d'états. La chaîne représentée ici correspond à la CR_1 dans le [tableau 5.1](#)

ces deux individus se traduisent respectivement par « E – E » et « Ch – Emp.CDD ». Ce principe s'étend à toute fenêtre temporelle prise en compte dans l'étude.



Dans le cas des chaînes de référence où les états correspondent aux maillons, la configuration diffère. En effet, si nous assimilons le texte à une fenêtre temporelle, les maillons apparaissent à des moments distincts. Nous proposons d'établir une analogie entre la position des maillons dans la chaîne et la fenêtre temporelle étudiée. Ainsi, chaque position peut être interprétée comme une période temporelle ou une date précise, commune à tous les individus. Par exemple, le troisième maillon d'une CR représente un état de la séquence à un moment donné (la position trois dans la chaîne), le même pour l'ensemble des individus. Bien sûr, cette analogie doit être nuancée, car malgré une position identique dans la chaîne, il ne faut pas oublier le texte et les autres paramètres qui entrent en jeu, comme la distance entre les maillons.

Comme nous l'avons mentionné précédemment, les séquences peuvent être constituées sous des formats différents en fonction de la nature des données et de la manière dont les informations sont organisées (cf. 5.2). Dans un premier temps, nous allons utiliser le format *STate-Sequences* (STS), qui se révèle être le format interne habituel dans TraMineR et le format le plus utilisé en analyse des séquences ([Gabadinho, Ritschard, Studer, & Muller, 2010](#)). Dans ce format, les informations que l'on prend en compte pour créer les séquences sont organisées comme suit : chaque individu correspond à une ligne ; les informations concernant chaque individu sont représentées en colonne, ainsi chaque colonne va correspondre à un état : donc, pour chaque séquence, on aura autant de colonnes que d'états successifs.



L'organisation des données que nous donnons en entrée à TraMineR correspond au format STS. En effet, lors de l'extraction des critères de description, nous avons associé à chaque chaîne (qui correspond à une ligne) un ensemble de critères. Chaque critère correspond à une colonne. Lorsque les critères portent sur le niveau *micro* et donc sur les maillons, on aura autant de colonnes consécutives que de maillons dans la chaîne. Les données en entrée se présentent alors comme indiqué dans le [tableau 5.1](#).

Id_CR	Type texte	Nature ref.	Longueur	M1	M2	M3	M4	...
CR_1	Narr{-}	Hum{+}	8	SN _{def}	Poss	Pro	Np	...
CR_2	Narr{+}	Hum{+}	7	Np	Pro	SN _{def}	Np	...
...	...	...	...	...	...	...	...	...
CR_27	Narr{-}	Hum{-}	3	SN _{ind}	SN _{dem}	SN _{dem}		

TABLEAU 5.1 – Organisation des informations pour les données utilisées pour l'analyse des séquences

À partir de ces données, que nous donnons en entrée à TraMineR, nous créons des séquences. Pour ce faire, d'un point de vue technique, nous avons besoin de gérer deux aspects : les états et la longueur des séquences.

Les états : Avant de créer les séquences à proprement parler, il faut établir l'alphabet, c'est-à-dire la liste de tous les états distincts présents dans l'ensemble des séquences. Dans cet alphabet, les valeurs possibles sont rangées par ordre alphabétique. Une fois l'alphabet établi, les données en entrée peuvent être transformées en séquences d'états et représentées sous le format choisi pour l'analyse.



Établir l'alphabet à partir de nos données signifie sélectionner le critère de description des CR que nous voulons prendre en compte, dans notre cas la nature des maillons. Techniquement, il s'agit alors de sélectionner parmi toutes les colonnes des données en entrée, celles qui correspondent à la nature des maillons. Notre alphabet sera alors constitué de neuf états distincts : $\{\emptyset, \text{Autre}, \text{Np}, \text{Poss}, \text{Pro}, \text{SN}_{\text{def}}, \text{SN}_{\text{dem}}, \text{SN}_{\text{ind}}, \text{SN}_{\emptyset}\}$. Une fois l'alphabet établi, nous pouvons générer les séquences au format STS. Nous obtenons alors des séquences avec autant d'états que de maillons dans la chaîne. Par exemple, pour les chaînes présentées dans le [tableau 5.1](#), nous obtenons les séquences suivantes³ :

CR_1 SN_{def} - Poss - Pro - Np - ...

CR_2 Np - Pro - SN_{def} - Np - ...

...

CR_27 SN_{ind} - SN_{dem} - SN_{dem}

3. Pour des raisons de lisibilité, nous avons décidé de ne pas représenter les chaînes dans leur intégralité.

Une fois la gestion des états terminée, il reste à contrôler la longueur des séquences, c'est-à-dire le nombre d'états à prendre en compte pour l'analyse.

La longueur des séquences : Dans la plupart des études sur les trajectoires professionnelles, pour chaque séquence (individu) on a un même nombre d'états (périodes de temps pris en compte). Néanmoins, comme le souligne [Robette \(2011\)](#), les séquences peuvent être de longueur différente. L'auteur souligne que cette variation doit être appréhendée différemment selon qu'elle découle des modalités de collecte des données ou qu'elle reflète la nature même de ces dernières. Lorsque la différence de longueur est liée aux modalités de collecte des données, elle n'apporte aucune information pertinente sur les types de séquences qui résulteraient de l'analyse. Au contraire, lorsqu'elle reflète une caractéristique intrinsèque des données, la variation de longueur peut révéler une information importante pour l'analyse. Ainsi, la prise en compte de séquences de longueur différente peut s'avérer pertinente pour mettre au jour une typologie des trajectoires, considérant « les [individus] « à trajectoire courte » [...] comme relativement similaires – et relativement distincts des [individus] « à trajectoire longue » » ([Robette, 2011](#), 15).



Dans l'étude des chaînes de référence, constituées d'un nombre varié de maillons, la différence de longueur entre les séquences est une caractéristique intrinsèque des données analysées. Dans ce sens, la prise en compte de cette variation apporte des informations importantes à notre analyse. C'est pour cette raison que nous jugeons pertinent de conserver des séquences de longueurs différentes. Dans cette optique, on aurait envie de prendre en compte les CR dans toute leur longueur, mais le choix n'est pas anodin.

Il faut prêter attention aux différences de longueur dans les séquences : [Robette \(2011\)](#) nous met en garde sur le fait de bien réfléchir au sens d'une comparaison entre des séquences de longueurs différentes, ce qui va dépendre de l'objectif de recherche et l'objet étudié. Par ailleurs, les variations peuvent influencer de manière significative les résultats obtenus lors de la comparaison des séquences entre elles (étape (e) du *core program*). En effet, le calcul de la similarité entre les paires de séquences peut être fortement impacté par les variations de longueur, comme nous le verrons en détail dans 5.2.3. Sans oublier que certaines mesures utilisées pour estimer la similarité entre les séquences n'autorisent pas la prise en compte des longueurs différentes ([Robette, 2011](#)).



Étant donné la variabilité dans la longueur des chaînes, il ne faut pas sous-estimer l'impact de la longueur sur les résultats. Elle varie plus ou moins considérablement selon la ressource utilisée : par exemple, dans AnnoDis la longueur peut varier entre 2 et 40 maillons et dans DEMRC, elle varie entre 3 et 895 maillons (cf. [chapitre 4](#)). Face à une telle variabilité, nous sommes confrontée à un choix délicat : d'une part, les différences de longueur peuvent apporter des informations importantes pour l'analyse ; d'autre part, une telle variation de longueur aura un fort impact sur les comparaisons des séquences et risque de fausser les résultats obtenus.



Par ailleurs, nous pouvons nous demander si comparer des séquences de longueurs aussi disparates a du sens : deux chaînes composées d'un même enchaînement d'états, mais de longueurs différentes, doivent-elles – et peuvent-elles – être considérées comme similaires ([Robette, 2011, 15](#)) ? Par exemple, une chaîne de 4 maillons et une autre de 28, toutes deux constituées d'un Np suivi d'une succession de pronoms, peuvent-elles être jugées similaires ? Sont-elles plus proches l'une de l'autre que deux chaînes de même longueur présentant une variation dans l'enchaînement de leurs maillons ? Dans ce contexte, nous optons pour la conservation des séquences de longueurs différentes afin de ne pas perdre d'informations potentiellement importantes pour l'identification des différents types des chaînes. Toutefois, afin de limiter l'impact de la variation des longueurs sur le calcul de similarité entre les séquences, nous choisissons de fixer une limite en termes de nombre de maillons à prendre en compte, c'est-à-dire le nombre maximum d'états pouvant constituer une séquence. La méthode pour fixer cette limite est présentée dans le [chapitre 6](#) ([tableau 6.4](#)).

5.2.2 Exploration et description des séquences

La deuxième étape de notre analyse consiste à explorer les données à traiter. Elle permet de proposer une première description des caractéristiques des séquences de manière globale. Il est possible d'observer la fréquence et la distribution des différents états qui constituent les séquences.

Nous proposons d'effectuer cette exploration sur les séquences coupées selon la longueur fixée. Cela nous permettra de voir les tendances globales qui se dégagent après la coupure des CR. Pour ce faire, nous exploitons le graphique de distribution ([figure 5.3](#)), qui permet de visualiser la nature des maillons (ou tout autre critère de description) dans les différents états qui composent les séquences (de M1 à M7 ici), chaque couleur correspondant à une catégorie grammaticale.

5.2.3 Comparaison des séquences

La troisième étape de l'analyse des séquences prévoit la comparaison des séquences. Cette comparaison s'effectue par le biais de mesures de (dis)similarité (ou distance)

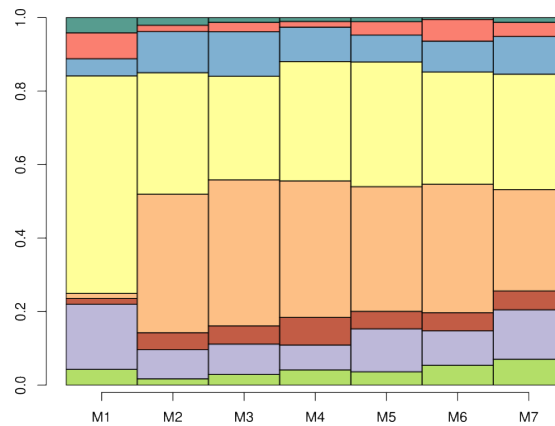


FIGURE 5.3 – Exemple de graphique de distribution obtenu pour la ressource AnnoDis

entre les séquences ([Studer & Ritschard, 2016](#)). Ces mesures sont nombreuses et peuvent être divisées en trois catégories, comme proposé par [Matthias et Gilbert \(2014\)](#) :

- les mesures basées sur la distance entre les distributions des états, comme la distance euclidienne ou la mesure χ^2 ;
- les mesures basées sur le nombre d'attributs communs entre deux séquences, comme la distance basée sur le calcul de la plus longue sous-séquence commune (*Longest Common Subsequence* (LCS)) et celle basée sur le plus long préfixe commun (*Longest Common Prefix* (LCP))
- les mesures basées sur une distance d'édition, comme le *Optimal Matching* (OM) et ses variantes ou encore la distance de Hamming (*Generalized Hamming* (HAM) et *Dynamic Hamming* (DHD)).

Le choix de la mesure de distance est un choix important, car les différentes mesures captent différents aspects des séquences, ayant ainsi un impact sur les résultats obtenus. Mais alors, comment savoir quelle catégorie de mesure utiliser ? Parmi ces mesures, laquelle est la plus adaptée pour l'étude des chaînes de référence ?

Pour pouvoir répondre à ces questions, il faut tout d'abord considérer les aspects selon lesquels les séquences peuvent différer. [Studer et Ritschard \(2016\)](#) proposent les aspects suivants⁴ :

- les états observés : les états distincts présents dans l'ensemble des séquences (i.e. l'alphabet) ;
- la distribution : la distribution de chaque état dans la séquence ;

4. Nous proposons une traduction qui facilite l'application aux chaînes de référence. Les catégories proposées par les auteurs sont :

- experienced states* — the distinct elements of the alphabet present in the sequence ;
- distribution* — the within-sequence state distribution (total time) ;
- timing* — the age or date at which each state appears ;
- duration* — the spell lengths in the distinct successive states ;
- sequencing* — the order of the distinct successive states.

- la position : l'âge ou la date à laquelle un état donné est observé.



Dans le cas des chaînes de référence, cela correspond à la position des maillons dans la CR : premier maillon, deuxième maillon, etc.

- la durée : durée dans des états successifs distincts. Si deux états successifs sont distincts, la durée sera égale à 1 pour chaque état. Si deux ou plusieurs états successifs sont identiques, la durée sera égale au nombre d'états identiques.



Prenons, par exemple, la séquence de la CR_1 présentée dans la [figure 5.2](#) où les états successifs sont les suivants : SN_{def} - Poss - Pro - Np - Pro - Pro - Pro - Np. La durée pour cette séquence correspond à : SN_{def} (1) - Poss (1) - Pro (1) - Np (1) - Pro (3) - Np (1).

- l'enchaînement : l'ordre des différents états successifs.



Parmi ces aspects, ceux qui nous intéressent le plus pour la caractérisation linéaire des chaînes de référence sont la position, la durée et l'enchaînement. En effet, ces trois aspects permettent de rester dans la linéarité et de rendre compte des modalités d'enchaînement des maillons, alors que les deux premiers écrasent la dimension linéaire. Étant donné que nous nous intéressons aux stratégies d'introduction et de maintien des référents, prendre en considération ces aspects nous permet de distinguer le premier maillon (introduction) des maillons suivants (maintien).

Maintenant que nous avons clarifié quels sont les aspects que nous souhaitons prendre en compte, il faut comprendre en quoi consistent les mesures de distance et, surtout, en quoi consiste leur application aux chaînes de référence. C'est ce que nous allons voir dans les sections qui suivent. Nous nous focalisons exclusivement sur les mesures intégrées dans TraMineR et nous en proposons une explication adaptée à l'étude des CR. Pour une explication bien plus détaillée et exhaustive, nous renvoyons aux travaux de [Matthias et Gilbert \(2014\)](#), [Studer et Ritschard \(2016\)](#) et [Liao et al. \(2022\)](#).



Pour illustrer nos propos, nous utiliserons trois chaînes que nous avons sélectionnées dans les ressources exploitées. Les CR_1 et CR_2, et la CR_3. Nous donnons le texte des trois chaînes ci-dessous ; les maillons sont signalés en gras.

CR_1

- (18) **Le journaliste britannique William Thomas Stead** a mené, durant **sa** carrière, de nombreux combats par le biais d'articles et de nouvelles. L'un d'entre eux concerne le manque de moyens de sauvetage à bord des paquebots. **Il** publie une première nouvelle en 1886 intitulée *Comment le Paquebot Poste sombra au milieu de l'Atlantique*, par un Survivant, racontant une collision entre deux navires dont les passagers ne sont pas tous sauvés, faute de moyens de sauvetage. **Stead** conclut : « C'est exactement ce qui se produira si les paquebots sont lancés avec trop peu de canots ». Six ans plus tard, **il** publie *De l'Ancien Monde au Nouveau*, nouvelle dans laquelle **il** raconte un voyage fictif qu'**il** aurait fait à bord du paquebot (bien réel) *Majestic* de la White Star Line sous le commandement d'Edward Smith. Au cours de la traversée, le navire s'arrête pour repêcher les naufragés d'un paquebot ayant heurté un iceberg. **Stead** conclut cette fois-ci en disant que « les océans parcourus par de rapides paquebots sont jonchés des os blanchis de ceux qui ont embarqué comme nous et qui ne sont jamais arrivés à bon port ». [...]

[TE_{w} ; 

CR_2

- (19) — Dis donc, **Marcel**, vends-tu pour cent sous, dis ? **L'homme**, enfoui dans une limousine, ne répondait pas, et la jeune femme, au bout de cinq grandes minutes, reprenait : — Dis **Marcel**, cent sous ce panier-là, et quatre francs l'autre, ça fait-il neuf francs qu'il faut **te** donner ? Un nouveau silence se fit : — Alors qu'est-ce qu'il faut **te** donner ? dix francs, tu le sais bien, **je** te l'ai dit ...

[R_{vdp} ; 

CR_3

- (20) Alternance Est-ce **Barak** qui a gagné ou Netanyahu qui a perdu ?? On ne peut résumer le triomphe **du candidat travailliste**, à cette question.. [...] La raison principale du succès de **M. Barak** repose sans doute sur **sa** grande capacité d'adaptation.. Au cours des derniers mois, **cet ancien chef d'état-major connu pour ses hauts faits d'armes et ses actions d'éclat**, a su mettre un pied partout. . Contrairement à Benyamin Netanyahu, qui a joué uniquement sur l'impact psychologique de la sécurité d'Israël, **il** a fait de la situation économique et de la progression du chômage, des thèmes de campagne. **Le leader travailliste** a aussi observé et tenu compte de l'évolution de la société israélienne, qu'**il** a préféré accompagner plutôt que de lui courir après, avec un train de retard. Ainsi, **il** a rassuré les laïcs sans se couper des religieux, flatté les électeurs russes, fédéré l'ensemble des mécontents de la politique et des méthodes de Netanyahu. Concernant l'avenir de la paix, **il** s'est efforcé de montrer **sa** volonté de créer un climat propice, à la reprise des discussions. Bref, s'**il** a une stratégie en dehors de la sécurité de la nation qu'**il** ne bradera jamais, c'est celle de l'éventail. [...] Mais après les acclamations, les parfums du triomphe, les discours et les serments, **Ehud Barak** va trouver les dossiers les plus épineux sur **son** bureau. [...] Et là encore, **M. Barak** semble assuré d'une majorité, qui **lui** évitera de se comporter en metteur en scène de génie pour gouverner. . Pourtant, compte tenu des décisions fondamentales qu'**il** **lui** faudra prendre, et dont certaines nécessiteront un véritable consensus, le plus difficile commence pour **M. Barak**. [$AP_{\{ER\}}$; ]

5.2.3.1 Mesures basées sur la distribution des états

Les mesures du premier type se basent sur l'analyse de la distribution des états dans chaque séquence. Autrement dit, ces mesures quantifient le temps total passé dans chaque état au sein d'une séquence (Matthias & Gilbert, 2014, 8).



Pour les chaînes de référence, cela revient à calculer la distribution des différentes valeurs du critère de description, soit les neuf catégories pour le critère de la nature. Considérons donc la CR_1, dont la séquence correspondante est : SN_{def} - Poss - Pro - Np - Pro - Pro - Pro - Np. La distribution des états de cette chaîne sera donc (0, 0, 2/9, 1/9, 4/9, 1/9, 0, 0, 0). Les distributions sont présentées dans le vecteur selon l'ordre alphabétique des neuf catégories (cf. 5.2.1).

Chaque séquence possède son vecteur de distribution des états. La (dis)similarité entre les séquences correspond à la distance entre leurs vecteurs de distribution. Cette

distance peut être calculée via :

- la distance euclidienne (EUCLID dans TraMineR) : cette mesure rend compte des différences absolues dans la proportion du temps passé dans un état donné.
- la distance du χ^2 (CHI2 dans TraMineR) : cette mesure effectue une pondération des différences au carré pour chaque état par l'inverse de la proportion globale du temps passé dans l'état. Cela permet de donner plus d'importance à un état peu fréquent.

Ces deux mesures rendent compte de la distribution des états de manière globale.



Appliquer ces mesures aux CR signifierait effacer les effets de linéarité qui nous intéressent particulièrement, puisqu'elles ne rendent pas compte des aspects de position et d'enchaînement. Par conséquent, nous ne prenons pas en considération ces deux mesures.

5.2.3.2 Mesures basées sur le nombre d'attributs communs

Le deuxième groupe de mesures est basé sur le nombre d'*attributs* communs entre les séquences. Ainsi, comme le soulignent [Aisenbrey et Fasang \(2010\)](#) et [C. Elzinga \(2007\)](#), plus le nombre d'attributs communs est élevé, plus les séquences sont considérées comme similaires. La distance entre elles est alors faible. Les attributs des séquences peuvent être appréhendés de manière différente, comme le souligne ([Aisenbrey & Fasang, 2010, 440](#)) :

« Possible sequence attributes are the number of common states in two sequences or the length of the first or last common spell in a sequence, which is also referred to as a common prefix or postfix. » ([Aisenbrey & Fasang, 2010, 440](#))

Les différentes possibilités ont donné lieu à différentes mesures. Nous allons en présenter trois. Nous nous sommes appuyée sur les mesures présentées dans [Gabadinho et al. \(2011\)](#)⁵ :

1. *longest common prefix* (LCP) ;
2. *reversed longest common prefix* (RLCP) ou *longest common suffix* ;
3. *longest common subsequence* (LCS).

Longest common prefix Cette mesure part de l'idée que les séquences qui commencent avec un même état sont plus similaires entre elles que celles qui commencent avec des états différents ([C. Elzinga, 2007, 9](#)).

5. Dans cet article, les auteurs présentent une quatrième mesure, *Simple Hamming distance* ([Hamming, 1950](#)). Cette mesure ne peut être appliquée qu'à des séquences de longueur identique, ce qui n'est pas notre cas. Pour cette raison, nous ne la prenons pas en considération.



Comme son nom l'indique, la distance LCP est calculée en prenant en compte le plus grand nombre d'états successifs communs entre les séquences à partir du premier état. Prenons les trois séquences suivantes (CR_1 (18), CR_2 (19) et CR_3 (20)) :

CR_1 : SN_{def} - Poss - Pro - Np - Pro - Pro - Pro - Np

CR_2 : Np - Pro - SN_{def} - Np - Pro - Pro - Pro

CR_3 : Np - Pro - SN_{def} - ØSN - Poss - SN_{dem} - Pro

Le premier état de la CR_1 et de la CR_2 et les deux suivants sont différents : les deux séquences n'ont pas de préfixe commun et elles sont donc considérées comme différentes entre elles. Il en va de même si l'on compare la CR_1 et la CR_3. Au contraire, la CR_2 et la CR_3 seront considérées comme similaires, car elles ont en commun les trois premiers états, qui constituent donc le plus long préfixe commun entre les deux séquences. Cette mesure pourrait être intéressante pour comparer les CR lorsqu'on s'intéresse aux modalités d'introduction des référents et au maintien sur une courte distance. L'application de cette mesure permettrait d'identifier des stratégies similaires d'introduction des référents.

La limite d'une telle mesure est qu'elle se focalise uniquement sur une sous-séquence commune qui commence avec le premier état des deux séquences. C. Elzinga (2007) suggère que cette mesure considère comme différentes des séquences dont le début diffère, bien que des états communs puissent être rencontrés plus loin dans la séquence. Il propose de prendre l'exemple de l'étude des carrières infructueuses : si un individu n'arrive pas à s'intégrer au marché du travail, cela n'apparaîtra que plus tard dans sa carrière (C. Elzinga, 2007). Cela est reflété par des états qui se situent donc vers le milieu ou la fin de la séquence.



Nous pouvons faire un parallèle avec les CR : deux chaînes peuvent commencer avec un maillon de nature différente et présenter une composition similaire par la suite. Cela ne sera pas pris en compte par la mesure LCP. Sans oublier le cas contraire : deux chaînes peuvent avoir un même début, mais présenter des enchaînements complètement différents.

Reversed longest common prefix Cette mesure a été proposée par C. Elzinga (2007) pour pallier la limite de la mesure LCP que nous avons présentée précédemment. Il s'agit de la mesure LCP inversée, dans le sens où elle se focalise sur le plus long suffixe commun. Cela signifie qu'elle prend en compte les états communs à partir de la fin de la séquence. Au regard de l'utilisation de cette mesure pour la comparaison des CR, nous pouvons émettre des réserves encore plus importantes que celles présentées pour la mesure LCP étant donné que les derniers maillons de la chaîne ne sont jamais pris en compte de manière privilégiée pour décrire la composition des CR.

Longest Common Subsequence Cette mesure prend en compte la sous-séquence commune la plus longue entre les séquences. Une sous-séquence est formée d'un ou plusieurs états qui constituent la séquence. Calculer la sous-séquence commune la plus

longue entre deux séquences revient donc à prendre en considération toutes les sous-séquences possibles, vérifier lesquelles figurent dans les deux séquences comparées et retenir la sous-séquence la plus longue, c'est-à-dire celle qui contient le plus grand nombre d'états présents dans les séquences comparées⁶. Comme le soulignent [Gabadinho et al. \(2011\)](#), les états dans la sous-séquence commune ne sont pas nécessairement consécutifs dans les deux séquences que l'on compare. Cela signifie que cette mesure fait abstraction de l'ordre des états. Dans ce sens, cette mesure ne rend pas compte de l'aspect d'enchaînement ni de la position des états dans la séquence.



Par exemple, prenons la CR_1 : nous pouvons la diviser en plusieurs sous-séquences, signalées entre crochets : [SN_{def}] ; [SN_{def} - Poss] ; [Poss - Pro] ; [SN_{def} - Poss - Pro] et ainsi de suite. À partir de cela, dans la CR_1 et la CR_2 la sous-séquence commune la plus longue est [Pro - Np - Pro - Pro - Pro], avec 5 états. Cet exemple montre bien que les états sont consécutifs dans la CR_1, mais pas dans la CR_2. En comparant la CR_2 avec la CR_3, la sous-séquence commune la plus longue est constituée de 3 états : [Np - Pro - SN_{def}]. Cette fois, les états sont consécutifs dans les deux séquences. Entre la CR_1 et la CR_3 la sous-séquence la plus longue est constituée de 3 états, non consécutifs dans la CR_3, à savoir [SN_{def} - Poss - Pro].

CR_1 : SN_{def} - Poss - Pro - Np - Pro - Pro - Pro - Np

CR_2 : Np - Pro - SN_{def} - Np - Pro - Pro - Pro

CR_3 : Np - Pro - SN_{def} - SN_∅ - Poss - SN_{dem} - Pro

Sur la base de ces observations, CR_1 et CR_2 sont considérées comme plus similaires entre elles que CR_2 et CR_3 et CR_1 et CR_3. Peut-on vraiment considérer que les CR_1 et CR_2 sont plus similaires que CR_2 et CR_3 ? D'un point de vue de leur composition globale, oui, nous en sommes d'accord. Par contre, nous l'avons déjà dit (cf. [partie II](#)), lorsqu'on reste dans la linéarité, on remarque que les stratégies mises en place sont différentes. Pour ces raisons, nous sommes réticente vis-à-vis de cette mesure.

5.2.3.3 Mesures d'édition

Les mesures de ce type calculent la distance entre des séquences sur la base des opérations nécessaires pour transformer une séquence en une autre. Cette transformation se réalise par des opérations d'insertion, de suppression et de substitution ([Gabadinho et al., 2011](#)). Toutes les mesures de distance d'édition n'autorisent pas les trois types d'opérations. Par exemple, la distance de Hamming généralisée (HAM) et la distance dynamique de Hamming (DHD) ([Lesnard & de Saint Pol, 2006](#)) n'autorisent que les opérations de substitution. De ce fait, ces deux mesures ne peuvent pas être appliquées à des séquences de longueur variable.

6. Pour une description détaillée des algorithmes, voir [C. Elzinga \(2007\)](#) ; [C. H. Elzinga \(2003\)](#) (entre autres).



Lorsqu'elles sont appliquées aux séquences, les deux distances HAM et DHD permettent de donner plus d'importance à la position et à l'enchaînement des états. Ces aspects sont précisément ceux qui nous intéressent pour l'étude des CR. Cependant, en raison de leur inapplicabilité aux séquences de longueur variable, nous ne les retenons pas. À l'inverse, l'*Optimal Matching* (OM), qui autorise les trois types d'opérations, peut être utilisé pour comparer des séquences de longueurs différentes. C'est pourquoi nous nous intéressons à cette mesure dans ce qui suit.

Optimal matching L'*Optimal Matching* (OM) - Appariement Optimal en français - est basé sur la mesure d'édition proposée par [Levenshtein \(1966\)](#) et a été introduit en sciences sociales par [Abbott et Barman \(1997\)](#). Cette mesure repose sur le calcul du coût minimum nécessaire pour rendre deux séquences identiques via les opérations de substitution, insertion et suppression des états ([Matthias & Gilbert, 2014](#)). Dans l'analyse des séquences, on distingue généralement deux types d'opérations : d'un côté, les opérations d'insertion et de suppression (indel), de l'autre, les opérations de substitution. Les premières favorisent l'identification d'enchaînements communs, en accordant moins d'importance à la position exacte des états. À l'inverse, la substitution valorise la position dans la séquence, en comparant prioritairement les états situés aux mêmes points dans les séquences observées ([Lesnard & de Saint Pol, 2006](#)). À chaque opération est assigné un coût qui doit être paramétré par l'utilisateur. [Studer et Ritschard \(2016\)](#) soulignent que la possibilité de paramétrer les coûts de ces opérations rend la mesure particulièrement flexible et applicable à différents cas de figure. Cependant, plusieurs auteurs ([Abbott et Tsay \(2000\)](#) ; [Aisenbrey et Fasang \(2010\)](#) ; [Lesnard \(2010\)](#), entre autres) ont critiqué l'OM en raison de la difficulté à donner un sens, notamment en sciences sociales, à ces opérations et aux coûts qui leur sont attribués ([Studer & Ritschard, 2016](#)).



Il en va de même pour l'analyse des CR : il est difficile de savoir comment paramétrer les coûts de ces opérations.

Un aspect fondamental concerne les conséquences de ces choix de paramétrage : pour les opérations d'insertion-suppression, fixer un coût faible revient à simplifier la dimension de la position des états. Cela se traduit par un alignement des enchaînements similaires, quelle que soit leur position dans la séquence. À l'inverse, plus le coût de ces opérations est élevé, plus on contraint la prise en compte de la position. Ce réglage conditionne donc le type de proximité que l'on fait ressortir dans les comparaisons. Les coûts de substitution concernent, quant à eux, des états situés à la même position dans les deux séquences comparées, mais dont le degré de similarité peut être modulé. Dans TraMineR, les coûts de substitution sont définis à l'aide d'une matrice de substitution que l'utilisateur peut générer avec la fonction `seqsubm()`. Pour construire cette matrice, il existe plusieurs méthodes :

- **CONSTANT** : consiste à attribuer un coût fixe et identique pour toute opération de substitution effectuée dans la séquence ;
- **FUTURE** : base le coût sur la similarité des états qui suivent ;

- INDEL : génère une matrice où le coût est égal au double du coût d'insertion-suppression pour toute opération de substitution ;
- INDELSLOG : attribue un coût ajusté logarithmiquement en fonction de la longueur des séquences ;
- TRATE : consiste à calculer le coût des opérations de substitution à partir de l'observation des pourcentages de transition d'un état à l'autre (Lesnard & de Saint Pol, 2006). L'idée sous-jacente est d'assigner un coût plus élevé quand la transition entre deux états est peu fréquente et un coût plus faible lorsque les transitions entre deux états sont fréquentes (Matthias & Gilbert, 2014).



Ainsi, avec la méthode TRATE, si nous prenons les trois chaînes nous servant d'exemple, le coût de substitution du Poss au Pro entre le CR_1 et la CR_2 sera plus bas que celui entre la CR_2 et la CR_3 pour substituer un Np en SN_∅, car l'enchaînement Poss >Pro (# 596) est plus fréquent que l'enchaînement Np >SN_∅ (# 24).

Il est important de noter que, malgré leur nom, les méthodes *INDEL* et *INDELSLOG* utilisées dans `seqsubm()` ne concernent pas les opérations d'insertion et de suppression. Elles génèrent des matrices de substitution selon des logiques spécifiques, mais n'affectent pas directement les coûts d'insertion-suppression. En effet, les coûts des opérations d'insertion et de suppression sont paramétrés indépendamment : l'utilisateur de TraMineR fixe ce coût lors du calcul de la distance entre les séquences avec la fonction `seqdist()`. L'argument *indel* permet d'attribuer manuellement un coût qui est en général constant et identique pour les deux opérations.

Dans le [tableau 5.2](#) nous résumons les méthodes et leur principe.

Méthode	Type d'opération	Principe
<i>Constant</i>	Substitution	Coût identique pour toutes les substitutions.
<i>Future</i>	Substitution	Basée sur la similarité des futurs (états suivants).
<i>Trate</i>	Substitution	Basée sur les taux de transition observés entre états.
<i>Indel (submat)</i>	Substitution	Coût de 0 pour identité, double du coût indel pour les autres.
<i>Indelslog</i>	Substitution	Coût ajusté logarithmiquement selon la longueur moyenne des séquences.
<i>Indel (paramètre)</i>	Insertion / Suppression	Coût constant fixé manuellement dans <code>'seqdist()'</code> .

TABLEAU 5.2 – Méthodes de calcul des coûts dans l'Optimal Matching (TraMineR)

Des variantes de l'OM ont été proposées et implémentées dans TraMineR, comme le OMslen, OMloc, OMspell. La première prend en considération les variations de longueur, la deuxième permet de donner plus d'importance à l'ordre des états successifs,

et la troisième permet de repérer des motifs récurrents (ou *spells*) dans l'ensemble des séquences.



L'OM nous semble une mesure pertinente pour plusieurs raisons : tout d'abord, autorisant les trois types d'opération, elle permet de comparer des séquences de longueurs différentes, ce qui est notre cas. Ensuite, elle permet de prendre en compte la position, la durée et l'enchaînement, caractéristiques qui nous intéressent particulièrement. Concernant le paramétrage des coûts des opérations, notre étude étant la première à appliquer l'analyse des séquences aux CR, nous ne pouvons pas nous appuyer sur des expérimentations existantes. Néanmoins, la méthode TRATE nous paraît être la plus pertinente, puisqu'elle se base sur les taux de transition observés dans les données, ce qui rend le choix moins arbitraire que celui du coût constant, difficile à fixer.

Pour résumer, les séquences peuvent différer selon plusieurs dimensions : les états observés, leur distribution, leur position, leur durée et leur enchaînement. Comme indiqué précédemment, les trois derniers aspects nous intéressent particulièrement dans le cadre de cette analyse, dont l'objectif est d'identifier des patrons des chaînes de référence. Nous avons donc retenu les mesures qui permettent de prendre en compte la position, la durée et l'enchaînement des états, tout en autorisant la comparaison entre séquences de longueurs différentes. Parmi ces mesures, nous avons retenu les méthodes d'OM, dont le paramétrage est détaillé dans le [chapitre 6](#) (cf. 6.2). Cette méthode repose sur deux types d'opérations : insertion-suppression et substitution, chacune associée à un coût. Comparer deux séquences revient à identifier l'ensemble des opérations de coût minimal permettant de transformer l'une en l'autre. La distance ainsi calculée correspond à ce coût minimal. L'ensemble des comparaisons est représenté sous forme d'une matrice de distances, qui constitue le point d'entrée pour la classification des séquences ([Lesnard & de Saint Pol, 2006](#)).

5.2.4 Classification des séquences

Cette étape consiste à exploiter la matrice de (dis)similarité obtenue afin de synthétiser et interpréter les résultats des comparaisons, en appliquant des méthodes de clustering ou de *multidimensional scaling*. Dans ce travail, nous avons recours à des méthodes de clustering afin de regrouper les séquences similaires. Comme le disent [Batool et Hennig \(2021\)](#), le choix des méthodes de classification dépend fortement des objectifs visés :

« There is a wide variety of clustering methods (Hennig et al., 2015), and it has been argued that there is no universally best approach, and that the cluster analysis approach needs to be chosen taking into account what kinds of clusters are required, which depends on domain knowledge and the aim of clustering (Von Luxburg et al., 2012; Hennig, 2015). » ([Batool & Hennig, 2021](#))

Comme le rappellent [Gabadinho et al. \(2011\)](#), le clustering permet d'obtenir automatiquement des classes (ou clusters) regroupant des séquences similaires.

« Clustering is an exploratory data analysis method aimed at finding automatically homogeneous groups or clusters in the data (Kaufman and Rousseeuw 2005). In life course studies (e.g., McVicar and Anyadike-Danes 2002 ; Widmer and Ritschard 2009), the method has typically been used in combination with OM distances to identify distinct groups of sequences with similar patterns ; that is, to define a typology of sequences. » (Gabadinho et al., 2011)



Dans notre cas, l'objectif est d'identifier des types différents d'enchaînements de maillons. En appliquant le clustering à nos données, les CR ayant des enchaînements similaires sont regroupées dans un même cluster. L'idée est d'obtenir un nombre x de clusters, chacun correspondant à un type d'enchaînements différent, ce qui permettra de dégager des régularités et de mettre au jour des types de CR différents.

Il existe de nombreux algorithmes de clustering dont nous ne discutons pas ici en détail. Nous renvoyons à [Ezugwu et al. \(2022\)](#) pour un état des lieux de ces algorithmes. Pour notre étude, nous nous sommes focalisée sur le clustering hiérarchique, pour deux raisons. La première concerne la reproductibilité de nos résultats : nous souhaitons avoir des résultats stables afin de garantir la validité des interprétations que nous proposons pour les clusters obtenus. Le clustering hiérarchique répond à cette exigence : il permet d'obtenir les mêmes regroupements à chaque fois que l'analyse est lancée sur les données, contrairement aux méthodes telles que K-means ([Lloyd, 1982](#) ; [Macqueen, 1967](#)) et PAM ([Reynolds, Richards, & Rayward-Smith, 2004](#)), qui sont non déterministes. La deuxième raison porte sur le nombre de clusters que l'on veut obtenir. Certaines méthodes, comme K-means, nécessitent que l'on établisse ce nombre à l'avance. Or, nous ne connaissons pas à priori le nombre de clusters qui serait pertinent pour les séquences. Le clustering hiérarchique permet de prendre cette décision dans un second temps, en se basant sur des outils tels que l'observation d'un *dendrogram*, ou l'analyse des sauts d'inertie, du coefficient de silhouette, entre autres. Nous présentons le protocole retenu pour déterminer le nombre de clusters dans le [chapitre 6](#) (cf. [6.2](#)). Le clustering hiérarchique peut être effectué suivant deux approches : l'approche agglomérative (ou ascendante) qui consiste à fusionner de façon prégressive les éléments les plus proches (ou similaires) ; l'approche divisive (ou descendante), qui part du groupe total pour le diviser en sous-groupes. Dans notre étude, nous adoptons l'approche agglomérative, couramment utilisée dans les études sur les séquences. À chaque étape de fusion, il est nécessaire de calculer la distance entre deux clusters. Pour ce faire, il existe plusieurs méthodes de calcul, appelées *méthodes de liaison* (*linkage*). Ici, nous présentons les méthodes propres à l'approche agglomérative ([Ackerman & Ben-David, 2016](#) ; [Angur Mahmud Jarman, 2020](#) ; [Ward, 1963](#)) :

- *Single* : la distance entre deux clusters correspond à la distance minimale entre les éléments appartenant aux deux clusters ;
- *Complete* : la distance entre deux clusters correspond à la distance maximale entre les éléments appartenant aux deux clusters ;
- *Average* : la distance entre deux clusters correspond à la distance moyenne entre tous les éléments de chaque clusters ;

- *Centroid* : la distance entre les deux clusters est calculée à partir de la distance entre les points centraux (*centroid*) des deux clusters ;
- *Ward* : elle consiste à minimiser la variance intra-cluster à chaque étape de fusion. Autrement dit, cette méthode donne une importance particulière à l'homogénéité (ou inertie) interne des clusters.



Dans notre étude, nous exploitons le clustering hiérarchique agglomératif avec la méthode de *Ward*. Celle-ci tend à produire des clusters compacts, bien séparés et de taille relativement équilibrée ([Aisenbrey & Fasang, 2010](#)). Ces propriétés en font une méthode particulièrement adaptée lorsqu'on cherche à faire émerger des patrons typiques, comme c'est le cas dans notre analyse des chaînes de référence. Le choix de cette méthode, ainsi que la comparaison avec d'autres combinaisons de paramètres que nous avons également testées, sont détaillés dans le [chapitre 6](#) (cf. 6.2).

Une fois les clusters obtenus, l'étape suivante de notre méthode consiste à interpréter les regroupements obtenus. Pour ce faire, nous nous appuyons sur différentes visualisations disponibles dans TraMineR. Comme nous allons le voir dans ce qui suit, ces représentations, propres à l'analyse des séquences, constituent un appui essentiel pour mettre en évidence les régularités et les différences dans la composition des chaînes de référence.

5.2.5 Interprétation des clusters obtenus

Après avoir effectué le clustering, la visualisation des données constitue un aspect important pour l'interprétation des résultats obtenus, comme le souligne [Liao et al. \(2022\)](#) :

« Within the explorative method of SA, visualization is of high importance because it makes use of our brain's visual capacity to find patterns in seemingly chaotic data. The visualization of sequences requires the display of three categorical dimensions : observational units, time points, and states. [...] When SA researchers visualize sequences, they typically want to highlight the three kinds of regularities we introduced in the introduction, sequencing, timing, and duration of states that may show the dynamics and the paths taken by individuals. Different types of visualization may show one or more of these regularities » ([Liao et al., 2022](#)).

Dans cette optique, nous mobilisons plusieurs types de visualisation pour guider l'interprétation des clusters et identifier les différents modes d'enchaînement des maillons. Chaque représentation graphique met en évidence un aspect particulier des séquences — hétérogénéité, enchaînements — et permet d'explorer visuellement les régularités propres à chaque classe. Cela nous permet d'interpréter les résultats en fonction de nos connaissances sur le domaine et en fonction des travaux existants sur les CR.

Nous commençons par une visualisation des *index plot*, illustrés dans la [figure 5.4](#). Ce type de graphique permet d'avoir une vision d'ensemble des clusters obtenus et d'identifier visuellement les régularités et les différences sur l'ensemble des données,

en privilégiant la dimension individuelle des séquences. Sur l’axe des x nous trouvons le nombre d’états qui composent la séquence, du premier (M1) au dernier (M7). L’axe des y représente les séquences individuelles. Les séquences sont ordonnées sur cet axe par leur distance. Cette distance correspond au score obtenu entre paires de séquences (cf. 5.2.3.3). Chaque couleur correspond dans la figure 5.4 à la catégorie grammaticale du maillon (la nature du maillon). Quel que soit le critère de description utilisé, chaque couleur correspond à une caractéristique distincte des maillons.

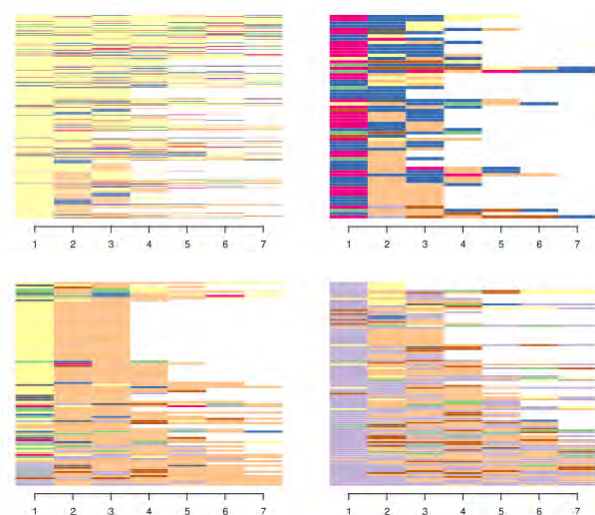


FIGURE 5.4 – Exemple des *index plots* obtenus après un clustering en 4 classes sur les données AnnoDis

Ensuite, nous nous intéressons à l’entropie transversale (Lesnard, 2010 ; Studer & Ritschard, 2016), qui permet de mesurer comment l’homogénéité de la classe évolue selon les états. Autrement dit, pour chaque position dans les séquences d’un cluster donné, l’entropie permet d’évaluer le degré de diversité des états observés. Cette mesure est dérivée de l’entropie de Shannon (Shannon, 1948), utilisée pour mesurer la diversité (ou l’incertitude) d’une distribution. Le score obtenu varie entre 0 (homogénéité maximale : un seul état dominant) et 1 (hétérogénéité maximale : forte dispersion des états).

Enfin, afin de conforter notre interprétation des tendances principales, nous identifions, pour chaque cluster, les cinq séquences représentatives. Il s’agit ici des cinq séquences modales, c’est-à-dire les plus fréquentes dans chaque cluster⁷. Les représentations graphiques de ces séquences, illustrées dans figure 5.5, offrent une synthèse visuelle des types d’enchaînements dominants dans chaque cluster. L’interprétation des clusters est accompagnée systématiquement d’un retour aux données et d’une observation qualitative des CR appartenant à chaque classe. Pour guider cette démarche, nous nous appuyons sur les séquences représentatives, qui facilitent la sélection d’exemples pertinents, mobilisés par la suite dans la présentation des résultats afin d’illustrer nos observations (cf. chapitre 7 et chapitre 8).

7. D’autres critères d’identification des séquences représentatives sont disponibles dans *TraMineR*, comme la séquence médiane ou la plus centrale.

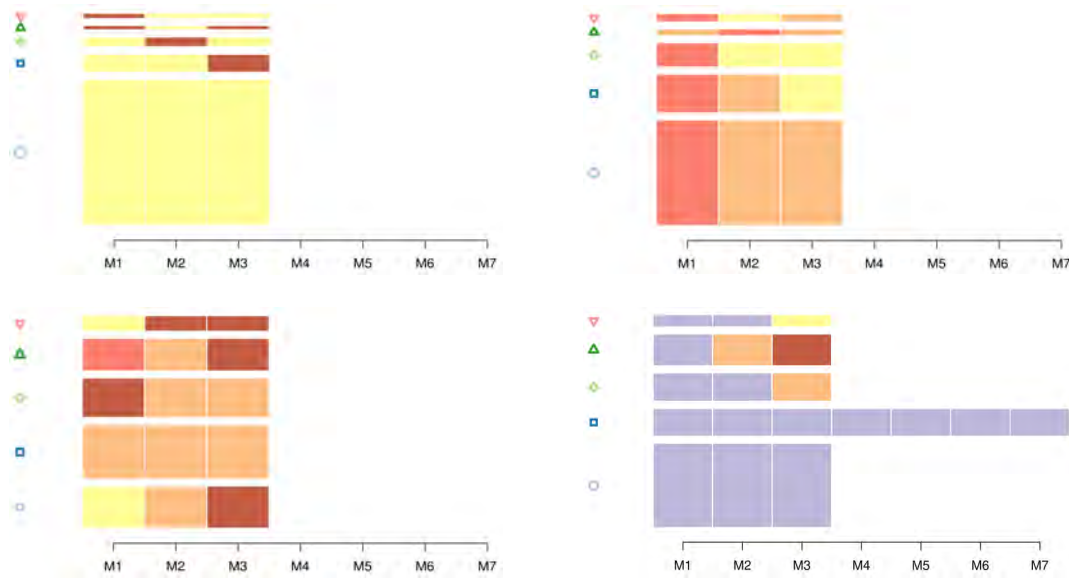


FIGURE 5.5 – Exemple des séquences représentatives obtenues après un clustering en 4 classes sur les données DEMRC

Pour résumer, notre interprétation des clusters repose sur un ensemble de visualisations graphiques. Bien que de nombreux autres outils et mesures soient disponibles (comme la complexité, la turbulence ou les matrices de transition), nous avons fait le choix de concentrer notre analyse sur des représentations directement exploitables pour l'interprétation qualitative des CR. Par exemple, à partir d'un algorithme de fouille de motifs séquentiels ([Massegia, Teisseire, & Poncelet, 2004](#)), il est possible d'identifier les sous-séquences discriminantes caractéristiques de chaque cluster. Cependant, nous n'avons pas retenu cette option dans le cadre de cette analyse. Nous avons privilégié des visualisations synthétiques globales centrées sur les dynamiques internes des séquences et leur régularité dans les clusters.

5.3 Des séquences aux patrons

La dernière étape de notre méthode consiste à dégager et proposer des patrons des CR qui reflètent les types d'enchaînements mis au jour grâce à l'analyse des séquences. Ces patrons sont pour nous une façon de formaliser les observations faites dans les étapes précédentes et de fournir une représentation systématique des modalités d'enchaînement que nous avons pu identifier. Cela permet de mieux rendre compte des stratégies mises en place pour introduire et maintenir le référent.

Pour dégager des patrons, nous utilisons les séquences au format *State-Permanence-Sequence* (SPS), proposé par [Aassve, Billari, et Piccarreta \(2007\)](#). Suivant ce format, chaque état distinct est associé à sa durée, c'est-à-dire au nombre d'occurrences consécutives dans la séquence. Cela permet de construire des séquences du type :

CR_1 (SN_{def}, 1)-(Poss, 1)-(Pro, 1)-(Np, 1)-(Pro, 3)-(Np, 1)

CR_2 (Np, 1)-(Pro, 1)-(SN_{def}, 1)-(Np, 1)-(Pro, 3)

Nous extrayons les séquences, codées dans ce format, pour chacun des clusters obtenus. Nous trions ensuite ces séquences par leur fréquence. Le nombre de séquences proposé par l'outil est élevé puisque dès que la durée d'un état est différente les séquences sont considérées comme différentes. Par exemple, une séquences de type $(SN_{def}, 3)$ et une autre de type $(SN_{def}, 4)$ sont considérées comme différentes. Afin de proposer des patrons plus génériques, nous regroupons les états associés avec le même type d'information, même si la durée est différente. Le patron est proposé dans le format suivant : $CAT\{n\}$, où CAT est le critère de description associé aux états et n le nombre d'états consécutifs ayant la même caractéristique. Par exemple, en prenant en considération la nature des maillons, une séquence composée de 3, 4 ou 5 syntagmes nominaux définis correspond au patron $SN_{def}\{3 : 5\}$. Étant donné que nous nous intéressons aux stratégies d'introduction et de maintien des référents, il est important de distinguer le premier maillon (introduction) des maillons suivants (maintien). Pour cette raison, nous proposons de distinguer le premier maillon des autres. Par conséquent, nous avons des patrons de type $CAT > CAT\{n1 : n2\}$. Par exemple, pour les CR composées de 3, 4 ou 5 SN_{def} nous aurons le patrons suivant : $SN_{def} > SN_{def}\{2 : 4\}$. Le symbole $>$ indique la transition d'une catégorie à l'autre.

Bilan

Dans ce chapitre, nous avons posé les fondements méthodologiques d'une approche inédite pour décrire les chaînes de référence en tenant compte de leur linéarité, c'est-à-dire de la manière dont les maillons s'enchaînent dans les CR. Nous avons proposé une méthode outillée pour caractériser systématiquement ces enchaînements à grande échelle.

Nous avons fait le choix d'appliquer l'analyse des séquences, une méthode issue des sciences sociales, aux chaînes de référence. Cette méthode permet de traiter les CR comme des séquences d'états, chaque état correspondant à un maillon caractérisé par un critère de description (comme la nature du maillon, sa fonction syntaxique, etc.). Cette méthode nous permet d'explorer les enchaînements de manière systématique afin d'identifier des régularités et des différences dans la composition des chaînes de référence. Pour ce faire, nous exploitons des méthodes de clustering qui permettent de regrouper automatiquement les chaînes similaires. Une fois les séquences regroupées en classes par le clustering, nous souhaitons évaluer si ces types de chaînes empiriquement identifiés sont associés à certains facteurs de variation : le type textuel, le genre textuel, et la nature du référent. Pour cela, nous appliquons un test du χ^2 , permettant d'évaluer statistiquement l'existence d'une dépendance entre les classes et les facteurs considérés. Cependant, l'analyse de la nature du référent nécessite un enrichissement des données, car cette information n'est pas disponible dans tous les corpus. Enfin, à partir de l'interprétation des clusters obtenus, nous proposons des patrons qui reflètent les types de composition les plus fréquents dans les deux ressources.

Cette démarche constitue, à notre connaissance, une première tentative d'adapter l'analyse des séquences à l'étude des chaînes de référence. Elle implique plusieurs choix de paramétrage essentiels, en particulier concernant la gestion des longueurs de chaînes, le format des séquences, ou encore les critères à retenir. Ces décisions sont

fondamentales pour garantir la cohérence et la pertinence des résultats. Le [chapitre 6](#) présente donc l'enrichissement des ressources (notamment l'annotation de la nature des référents), et les choix des paramètres requis pour l'analyse.

Chapitre 6

Explorer les séquences : enrichissement des données, paramétrage et visualisation

Contents

6.1	Enrichissement des ressources	124
6.1.1	La nature des référents dans AnnoDis	124
6.1.2	La nature des référents dans DEMRC	127
6.2	Paramétrage de l'analyse des séquences	129
6.2.1	Création des séquences	130
6.2.2	Comparaison et classification des séquences	133
6.2.3	Choix du nombre de clusters	135
6.3	Bilan	140

Ce chapitre présente le protocole expérimental élaboré pour appliquer l'approche linéaire à l'analyse des CR. Avant d'entamer l'analyse des séquences, nous avons finalisé la préparation des données en les enrichissant par des annotations. Cet enrichissement était nécessaire pour disposer de l'ensemble des informations requises pour mener notre analyse à bien. En particulier, nous avons ajouté des annotations sur la nature du référent, ces informations n'étant pas initialement présentes dans les corpus utilisés (cf. [chapitre 3](#)). Ensuite, comme mentionné dans le chapitre précédent, appliquer l'analyse des séquences aux chaînes de référence nécessite certaines adaptations méthodologiques. Nous allons donc présenter en détail les ajustements que nous avons adoptés, notamment pour la création des séquences et pour le paramétrage de la classification de celles-ci, en reconnaissant que certains choix peuvent demeurer arbitraires.

6.1 Enrichissement des ressources

Dans cette section, nous expliquons la démarche adoptée pour enrichir les données et obtenir les informations nécessaires à notre analyse. Comme mentionné dans le [chapitre 4](#), les deux ressources exploitées dans le cadre de cette thèse ne contiennent pas d’annotation sur la nature des référents. Or, cette information constitue l’un des facteurs de variation dont nous souhaitons étudier l’impact sur les types de composition des CR. Il a donc été nécessaire d’ajouter ces annotations pour rendre possible cette analyse. Les modalités d’annotation diffèrent pour AnnoDis et DEMRC en raison de leurs caractéristiques intrinsèques, ainsi que des contraintes de temps et de ressources humaines auxquelles nous avons été soumise aux différents moments où l’annotation a été réalisée au cours de la thèse.

6.1.1 La nature des référents dans AnnoDis

La notion de nature du référent est analysée dans cette thèse sous deux aspects : d’une part, le degré de *concrétude* du référent, c’est-à-dire s’il s’agit d’une entité concrète ou abstraite ; d’autre part, l’animacité du référent, en distinguant les humains (individuels ou collectifs) des référents non humains (cf. [2.3.3](#)).

Dans la ressource AnnoDis, seul le deuxième aspect est pris en charge pour les annotations, que nous avons effectuées de manière semi-automatique. Nous avons fait le choix de nous focaliser sur la nature ontologique étant donné que l’annotation automatique du degré de concrétude s’avère difficile à mettre en place. L’annotation a été effectuée en deux étapes : une première annotation automatique à base de règles, suivie d’une étape de vérification manuelle des résultats obtenus. Nous étions seule sur cette tâche de vérification et correction des annotations automatiques, ainsi, par souci de fiabilité des annotations, nous avons opté pour une annotation binaire de la nature ontologique du référent : référents humains vs référents non humains ¹.

Concernant l’annotation automatique, nous avons appliqué des règles très simples en nous appuyant sur la nature des maillons, harmonisée lors du prétraitement d’AnnoDis (cf. [4.1](#)). Cette information était la seule exploitable sans recours à des ressources externes. En effet, l’étiquetage Talisman utilisé pour cette ressource ne fournit pas l’annotation des entités nommées, ce qui aurait permis une caractérisation plus précise. Nous avons appliqué les règles suivantes :

- pour chaque chaîne, on considère tous les maillons qui la constituent ;
- si parmi les maillons il y a des noms propres, alors la chaîne est annotée *référent humain* ;
- sinon, la chaîne est annotée *référent non humain*.

Le choix d’appliquer des règles aussi grossières se justifie par la possibilité de vérifier manuellement les annotations, garantissant ainsi la qualité du processus malgré sa première approche simplifiée. La vérification manuelle était possible étant donné le nombre relativement faible de CR.

1. Une annotation plus fine de ce trait ainsi que du caractère concret ou abstrait des référents peut être envisagée pour des travaux futurs.

Pendant la phase de vérification manuelle, plusieurs questions se sont présentées, dont certaines étroitement liées au modèle d'annotation AnnoDis. Certaines annotations pouvaient parfois soulever des doutes quant à la cohérence du référent de la chaîne. En d'autres termes, il était possible qu'un maillon d'une chaîne semble inadéquatement rattaché à cette chaîne spécifique, car il pouvait renvoyer à un autre référent. Pour illustrer cela, prenons l'exemple (21), où le dernier maillon de la chaîne est celui qui apparaît comme un intrus. En effet, seul le déterminant possessif devrait être considéré, car c'est celui-ci qui coréfère à *la Cour. sa légitimité* est un autre référent pouvant donner lieu à une autre chaîne.

- (21) Ceci est d'autant plus porteur d'incertitudes que les mesures présidentielles n'entament pas sa crédibilité - bien au contraire - et que **la Cour Suprême** actuelle n'est pas portée le moins du monde à remettre en cause cette évolution, et ce pour une série de raisons. D'abord, en règle générale, **la Cour** est certainement une des institutions nationales les plus à l'écoute des évolutions de l'opinion publique, et ce contrairement à ce que voudrait une interprétation étroite de la difficulté contre majoritaire. Étant potentiellement soumise à une critique de type majoritaire, **elle** se doit de faire preuve d'une grande prudence, et tenter de refléter dans une certaine mesure les opinions dominantes, sous peine d'entamer **sa légitimité** [AS_{G} ; 

Toutefois, nous rappelons que les modalités de délimitation des maillons n'ont pas été explicitées dans le guide d'annotation AnnoDis (cf. 3.3). Par conséquent, nous considérons que le maillon est bien rattaché à la chaîne en cours de traitement, en nous basant sur le fait que le déterminant possessif *sa* serait normalement annoté seul. Nous interprétons donc la présence du nom qui l'accompagne comme la conséquence du modèle d'annotation adopté.

Une question récurrente concerne le traitement des toponymes. Ceux-ci, étant étiquetés comme Np, ont été considérés comme des référents humains. Cette annotation est cohérente si l'on considère que dans certains contextes, les toponymes peuvent désigner des référents humains collectifs (Lecolle, 2006, 2019). Cependant, parfois au sein d'une même chaîne, certains maillons renvoient à un humain collectif, tandis que d'autres font référence à l'entité géographique, comme le montre l'exemple (22)².

- (22) Plus qu'un choix, c'était un devoir pour **le Kazakhstan et l'Azerbaïdjan** d'ouvrir leurs économies. En effet, **ces deux pays** avaient cruellement besoin de capitaux. La capacité de financement domestique étant trop limitée, un système autarcique était impossible **dans ces deux pays**. **Ils** avaient besoin de capitaux pour reconvertir l'appareil de production et développer l'exploitation des hydrocarbures ou développer la production métallurgique. **Le Kazakhstan** a toujours été une économie fondée sur les industries extractives. Ainsi, en 1990, plus des deux tiers de la **production industrielle kazakh** provenait de ces industries. [AS_{G} ; 

2. Dans cet exemple, nous avons aussi un rattachement inhabituel : la mention *Le Kazakhstan* est rattachée à la CR dont le référent est *le Kazakhstan et l'Azerbaïdjan*. Nous mettons encore une fois en évidence que cela est une conséquence du modèle d'annotation adopté.

La question est alors de savoir comment traiter ces cas de manière cohérente. Afin de simplifier le processus d’annotation, nous avons fait le choix de les considérer comme des référents humains, en tenant compte du fait qu’ils peuvent, dans certains contextes, désigner des groupes humains collectifs. De plus, pour valider ou corriger les annotations, compte tenu des outils disponibles à ce moment-là, ainsi que des contraintes de temps, nous avons décidé de nous appuyer sur le premier maillon des chaînes. La nature du référent est ainsi validée ou corrigée si nécessaire sur la base de ce premier maillon, nous permettant ainsi de gérer les ambiguïtés dont nous venons de parler.

L’étape de vérification manuelle nous a permis de corriger 30% des CR (soit 169 / 581 CR). Les exemples (23a) et (23b)³ montrent des cas d’erreurs que nous avons corrigé manuellement : respectivement une CR dont le référent a été classé comme humain en raison de la présence de noms propres, alors que ces derniers ne renvoient pas à des êtres humains et une CR où le référent humain a été classé comme non humain en raison de l’absence de noms propres.

- (23) a. **Linux**, ou GNU / Linux, est un système d’exploitation compatible POSIX. Linux est créé en 1991 par Linus Torvalds pour ordinateur. Développé sur Internet par des dizaines de milliers d’informaticiens bénévoles et salariés, **Linux** fonctionne maintenant sur du matériel allant du modem au supercalculateur. Il existe de nombreuses distributions Linux indépendantes, destinées aux ordinateurs personnels et serveurs informatiques, pour lesquels **Linux** est très populaire. Elles incluent des milliers de logiciels, notamment ceux du projet GNU, d’où la dénomination GNU / Linux. **Linux** est également populaire sur système embarqué. La mascotte de Linux est le manchot Tux. [AE_{w} ; ⚡]
- b. **Le cycliste** est aussi moins exposé aux polluants de l’air que les autres usagers de la route. Même sur une route circulante, **il** aspire moins de gaz que les automobilistes parce qu’il roule sur le côté de la chaussée et parce que la prise d’air des véhicules est généralement plus proche des pots d’échappement que le nez du cycliste - où les taux de pollution sont plus faibles. **Sa position** surélevée lui permet d’échapper à certains polluants qui sont plus lourds que l’air. Sans compter que **les cyclistes** peuvent profiter de leur flexibilité pour explorer des parcours qui évitent les grands axes de circulation. [AE_{w} ; ⚡]

L’annotation de la nature du référent, après vérification et correction manuelle, aboutit à une répartition équilibrée des chaînes dans l’ensemble du corpus, comme illustré dans le [tableau 6.1](#). Cependant, il existe des divergences entre les sous-corpus. Comme nous pouvions nous y attendre, les chaînes à référent humain (AnnoDis_{Hum+}) sont sous-représentées dans les articles de recherche en linguistique. Dans les autres sous-corpus, la distribution des chaînes à référent humain et celle à référent non humain (AnnoDis_{Hum-}) est plus équilibrée, avec un nombre de CR à référent humain légèrement supérieur à celui de CR à référent non humain.

3. Le lecteur remarquera que, dans les deux exemples, certaines mentions du référent ne sont pas annotées, ou alors, la délimitation des maillons n’est pas celle à quoi on s’attendrait. Nous rappelons que cela dépend du modèle d’annotation adopté.

Genre	AnnoDis _{Hum+}		AnnoDis _{Hum-}		Nb CR
	#	%	#	%	
Articles Scientifiques _{Geop}	136	58	98	42	234
Articles Scientifiques _{Ling}	22	25	65	75	87
Articles Encyclopédiques _{Wiki}	138	53	122	47	260
AnnoDis	296	51	285	49	581

TABLEAU 6.1 – Distribution de la nature du référent dans AnnoDis

À la découverte des référents annotés : leçons d’une annotation semi-automatique Cette première expérience d’annotation semi-automatique, malgré l’utilisation de règles très simples, a permis d’annoter correctement 70% des CR. Cependant, une telle approche n’est envisageable que si l’annotation automatique peut être vérifiée manuellement, tâche qui demeure relativement complexe lorsqu’il s’agit de traiter un grand nombre de données.

Les exemples (23a) et (23b) illustrent bien les limites des règles que nous avons appliquées. Celles-ci se basent uniquement sur la présence ou l’absence d’un nom propre parmi les maillons de la CR. Or, l’attribution de la nature des maillons repose sur une harmonisation automatique issue de l’étiquetage réalisé par Talisman, qui peut, par ailleurs, comporter des erreurs. Néanmoins, lorsque nous utilisons ces outils pour traiter un grand nombre de données, nous devons accepter la possibilité que des erreurs soient présentes. La question est donc de trouver des solutions pour limiter leur impact sur les analyses. Lorsqu’il s’agit d’annoter la nature des référents, l’utilisation de l’annotation en entités nommées ainsi que de ressources externes peut constituer une approche pertinente.

Un autre aspect qui pourrait être mis en discussion, concerne l’étape de vérification manuelle. En effet, la vérification et la correction des annotations automatiques a été effectuée par une seule personne, ne permettant pas de calculer un accord inter-annotateur. Cependant, comme nous l’avons dit, pour garantir une certaine fiabilité, nous avons rendu la tâche la plus simple possible, en nous focalisant sur une annotation binaire de la nature ontologique des référents. Par ailleurs, il serait sans doute plus judicieux de distinguer les cas où les toponymes correspondent à des référents humains (collectifs) des cas où ils correspondent à des référents non humains. Toutefois, en corrigeant les annotations, nous nous sommes rendue compte que la plupart des cas correspondent à des humains collectifs. Par ailleurs, il pourrait être intéressant de proposer une analyse plus fine de la nature ontologique, en distinguant les référents humains collectifs des référents humains individuels.

Nous avons tiré parti des enseignements de notre première expérimentation pour l’annotation de la nature du référent dans DEMRC.

6.1.2 La nature des référents dans DEMRC

Suite à l’expérience sur AnnoDis, nous avons opté pour une annotation plus fine et complète de la nature des référents dans DEMRC. Étant donné que cette ressource contient un plus grand nombre de CR, une annotation plus fine semble pertinente et

pourrait fournir des effectifs suffisamment importants pour réaliser des tests statistiques. Nous avons donc décidé d'effectuer une annotation sur deux aspects : le degré de concrétude et la nature ontologique des référents. Concernant la nature ontologique, nous voulions faire une distinction entre les référents non humains, les référents humains collectifs et les humains individuels.

Nous avons mené une campagne d'annotation afin d'obtenir des annotations manuelles sur l'ensemble des données. Nous détaillons ici les étapes de cette campagne, les résultats obtenus et l'accord inter-annotateur. Nous présentons également la méthode d'adjudication en cas de désaccord.

La première étape a consisté à rédiger un guide d'annotation, testé par nos directrices (Cécile Fabre et Lydia-Mai Ho-Dac) et une enseignante-chercheuse (Josette Rebeyrolle) sur une dizaine de chaînes, et nous avons ensuite discuté des cas de désaccord, ce qui nous a permis d'améliorer le guide. Avant de lancer la campagne d'annotation, une seconde phase de test a été menée avec une doctorante en deuxième année sur 160 CR, ce qui nous a permis d'affiner encore les critères d'annotation. L'accord inter-annotateur calculé via le kappa de Cohen (Cohen, 1960) a montré des résultats satisfaisants :

- Distinction humain / non humain : $k = 0,9$
- Humain collectif / individuel : $k = 0,7$
- Degré de concrétude : $k = 0,75$

Ces scores nous ont permis de valider le guide et de lancer la campagne d'annotation. Six annotateurs⁴ ont participé, dont trois étudiants de L3 en Sciences du Langage à l'Université de Toulouse Jean Jaurès effectuant leur stage au laboratoire CLLE, et trois experts (Cécile Fabre, Josette Rebeyrolle et nous-même). Une phase d'entraînement a été mise en place : nous avons formé trois binômes, constitués d'un expert et d'un étudiant. Chaque groupe a annoté 50 CR, soit 150 chaînes en tout. Afin d'éviter des biais liés aux types de texte ou aux genres textuels, nous avons dispatché les CR de manière aléatoire. Lors de cette première phase, les annotateurs ne pouvaient pas échanger entre eux. Ces annotations ont été analysées pour identifier les désaccords, qui ont été discutés lors des réunions afin de clarifier le guide. Cette étape a été essentielle pour finaliser la version définitive du guide d'annotation.

Ensuite, nous avons lancé la campagne d'annotation. Chaque annotateur a reçu environ 720 chaînes à annoter. Dans le guide d'annotation (disponible en [annexe D](#)), il est demandé aux annotateurs d'identifier deux caractéristiques des référents :

- Concret *vs* abstrait, avec une échelle de tangibilité ;
- Humain *vs* non humain, avec distinction entre humain individuel et humain collectif, selon les définitions de (Lecolle, 2016). Une option « doute » était disponible mais à utiliser avec parcimonie.

4. Suite au désistement d'un annotateur, une septième annotatrice a dû être contactée afin de pouvoir réaliser l'annotation de la totalité des CR. Il s'agit d'une ancienne étudiante qui a fait son master en Science du Langage (master LiCo) à l'Université de Toulouse Jean-Jaurès. Nous remercions Mathilde Lala d'avoir accepté de prendre en charge une partie de l'annotation et de nous avoir ainsi permis de couvrir la totalité des chaînes.

Genre	DEMRC _{Nar-}											
	Abstrait		Concret		DEMRC _{Hum-}		DEMRC _{HumInd+}		DEMRC _{HumColl+}		Total	
	#	%	#	%	#	%	#	%	#	%	#	%
articles de presse	43	13,4	278	86,6	108	33,6	111	34,6	59	18,4	321	100
articles encyclopédiques	22	18,6	96	81,4	80	67,8	13	11,0	3	2,5	118	100
pamphlet	36	27,5	95	72,5	29	22,1	48	36,6	18	13,7	131	100
texte juridique	670	46,0	785	54,0	318	21,9	294	20,2	173	11,9	1455	100
Total DEMRC_{Nar-}	771	38,1	1254	61,9	535	26,4	466	23,0	253	12,5	2025	100
Genre	DEMRC _{Nar+}											
	Abstrait		Concret		DEMRC _{Hum-}		DEMRC _{HumInd+}		DEMRC _{HumColl+}		Total	
	#	%	#	%	#	%	#	%	#	%	#	%
biographie	129	28,7	321	71,3	154	34,2	155	34,4	12	2,7	450	100
nouvelle	62	12,4	439	87,6	196	39,1	216	43,1	27	5,4	501	100
roman	187	14,9	1072	85,1	685	54,4	350	27,8	37	2,9	1259	100
Total DEMRC_{Nar+}	378	17,19	1832	82,9	1035	46,8	721	32,6	76	3,4	2210	100

TABLEAU 6.2 – Distribution de la nature des référents dans DEMRC

Les chaînes ont été annotées dans leur contexte pour éviter les erreurs d'interprétation.

À la fin de la campagne, toutes les 4 235 chaînes de la ressource DEMRC ont pu être annotées. 27% des CR (soit 1143) ont été annotées par deux ou plus annotateurs. Nous avons calculé l'accord inter-annotateur pour les différentes annotations. Pour les 805 chaînes qui ont été annotées par 2 annotateurs, nous avons calculé le kappa de Cohen. Nous obtenons un kappa de 0,76 pour le trait concret-abstrait et 0,78 pour la distinction humains-non humains. L'accord obtenu pour le trait humain individuel est de 0,89 ; $k = 0,80$ pour le trait non humain et $k = 0,60$ pour humain collectif.

Afin de procéder à l'adjudication de l'annotation, nous avons fait le choix d'attribuer la caractéristique qui était majoritaire pour chaque CR. Par exemple, si la chaîne a été annotée par 4 personnes, dont 3 ont sélectionné abstrait et 1 concret, le référent sera considéré comme abstrait. En cas d'égalité dans les annotations, nous avons dû trancher : pour ce faire, nous sommes retournée analyser la chaîne en contexte. Cette procédure a ses limites, l'idéal aurait été de pouvoir solliciter un troisième avis. Dans le [tableau 6.2](#) nous présentons la distribution de la nature des référents dans DEMRC. Nous pouvons observer que les référents concrets sont plus fréquents que les référents abstraits. Les référents humains individuels sont également les plus fréquents, suivi des non humains et des collectifs, qui représentent la catégorie la moins représentée.

6.2 Paramétrage de l'analyse des séquences

Dans cette section, nous détaillons les différentes étapes suivies pour appliquer l'analyse des séquences à l'étude des CR. La première étape de notre protocole expérimental porte sur la création des séquences, ce qui implique le choix de critères de description caractérisant les états et la longueur des séquences, un aspect qui demande une attention particulière (cf. [chapitre 5](#)). La deuxième étape concerne le choix des paramètres pour effectuer la classification des séquences. Pour ce faire, plusieurs éléments doivent être pris en considération : la mesure de similarité utilisée pour comparer les séquences entre elles, les paramètres qu'elle implique, ainsi que l'algorithme de clustering, avec les paramètres à ajuster avant de procéder à la classification des données. Les paramètres que nous présentons ici sont utilisés à la fois pour l'analyse des chaînes AnnoDis et pour les chaînes DEMRC.

6.2.1 Création des séquences

En considérant les chaînes comme des séquences d'états, où chaque état correspond à un maillon (cf. [chapitre 5](#)), la première étape de notre méthode consiste en la sélection du critère de description à exploiter pour établir l'*alphabet* et créer les séquences (cf. [5.2.1](#)). Dans cette expérimentation, le critère pris en compte est celui de la nature des maillons, comme illustré schématiquement dans la [figure 6.1](#).⁵

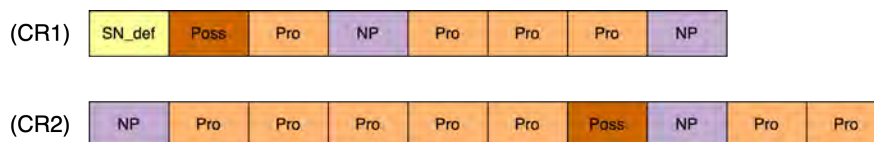


FIGURE 6.1 – Exemple de deux chaînes AnnoDis encodées en tant que séquences d'états caractérisés par la nature du maillon

À partir de la nature des maillons, notre *alphabet* est alors composé de neuf catégories différentes (huit pour AnnoDis, car les sujets zéro ne sont pas annotés dans cette ressource, cf. [chapitre 4](#)). Les catégories sont les suivantes : $\emptyset$, SN_{def} , SN_{dem} , SN_{ind} , $SN_{\emptyset}$, Poss, Pro, Np, Autre.

Créer les séquences, signifie également considérer leur longueur. Concernant cet aspect, nous rappelons que des fortes variations de longueur peuvent avoir un impact sur le calcul de la distance entre les séquences, et par conséquent un impact sur leur classification. Réduire cet impact est d'autant plus important lorsque les données présentent de très fortes variations, comme c'est le cas pour les CR de la ressource DEMRC : leur longueur varie entre 3 et 895 maillons. Il est donc important de fixer une limite en termes de nombre d'états, c'est-à-dire de nombre de maillons, à prendre en compte. Cette limite est établie à partir de l'observation de la distribution des chaînes en fonction de leur longueur (cf. [5.2.1](#)). Celle-ci peut varier selon les données (types de texte, modèle d'annotation, etc.), c'est pourquoi nous observons ce comportement pour chacune des ressources que nous utilisons. Comme le montre la [figure 6.2](#), près de 80% des chaînes de AnnoDis et de DEMRC sont constituées de sept maillons ou moins. Ces résultats nous permettent de fixer la limite à sept maillons pour l'approche en séquences des CR dans les deux ressources. Ainsi, dans AnnoDis 21,3% des CR, composées de plus de sept maillons, ont été coupées. Dans DEMRC ce pourcentage s'élève à 22,8%. Par conséquent, toutes les séquences AnnoDis ont une longueur qui varie entre 2 et 7 états, celle de DEMRC varie entre 3 et 7. Cette différence s'explique par le modèle d'annotation (cf. [3.3.1](#)). Le [tableau 6.5](#) indique la distribution des séquences selon leur longueur une fois la réduction à sept maillons opérée. Concernant les séquences composées de sept états, le chiffre indiqué dans le tableau correspond à la somme des CR dont la longueur d'origine est de sept maillons (156, soit 5,5% des CR dans AnnoDis et 174, soit 4,1% des CR dans DEMRC) et de celles qui ont été coupées à sept (124, soit 21,3% dans AnnoDis et 967, soit 22,8% CR dans DEMRC).

5. Nous avons testé également deux autres critères (fonction syntaxique et type de reprise), mais comme les résultats ne sont pas directement exploités dans le cadre de cette thèse, nous ne les présentons pas ici.

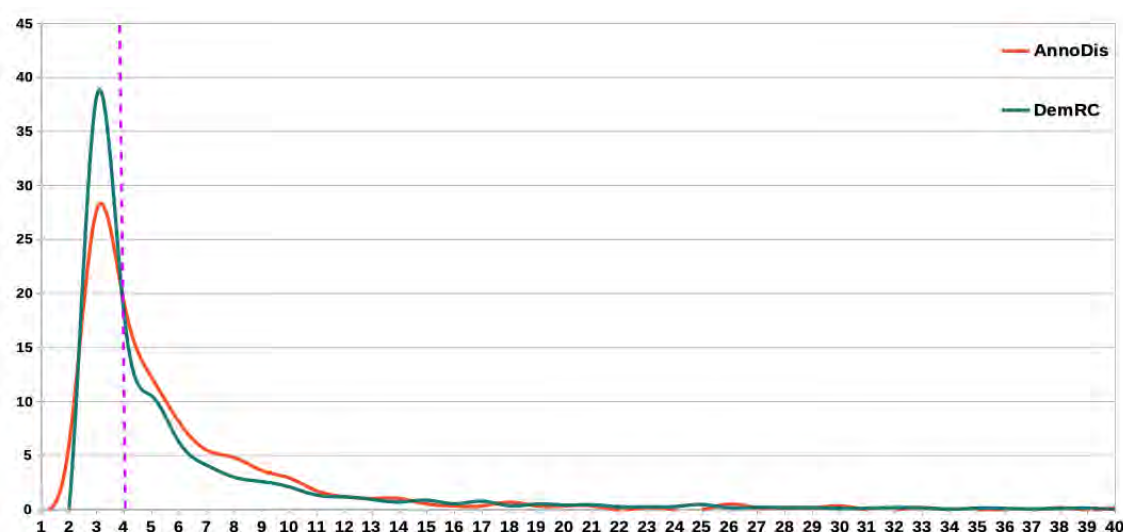


FIGURE 6.2 – Distribution des chaînes AnnoDis et DEMRC en fonction de leur longueur en nombre de maillons

Nombre de séq.	#	%
= 2 états	35	6
= 3 états	161	27,7
= 4 états	111	19,1
= 5 états	71	12,2
= 6 états	47	8,1
= 7 états	156	26,9
Total	581	100

TABLEAU 6.3 – AnnoDis

Nombre de séq.	#	%
= 2 états	0	0
= 3 états	1 620	38,3
= 4 états	764	18
= 5 états	446	10,5
= 6 états	264	6,2
= 7 états	1141	26,9
Total	4 235	100

TABLEAU 6.4 – DEMRC

TABLEAU 6.5 – Distribution des séquences selon leur longueur après coupe à sept états dans les deux ressources

Le [tableau 6.6](#) et le [tableau 6.7](#) indiquent respectivement la distribution dans AnnoDis et dans DEMRC, par ordre de fréquence, de la nature des maillons dans les CR avant et après la coupe. Les colonnes *après coupe* indiquent dans l'ordre : les effectifs retenus après coupe selon la nature du maillon ; le pourcentage calculé sur le nombre de maillons de la même nature avant la coupe (par exemple dans DEMRC, après la coupe, il reste 4 998 pronoms sur les 18 566 présents avant la coupe, soit 26,9% des pronoms) ; le pourcentage de maillons retenus sur le nombre total des maillons (par exemple, dans DEMRC 25,3% des maillons retenus après coupe (19 717) sont des pronoms). La dernière colonne illustre la quantité d'information exclue par la coupe à sept maillons. Par exemple dans DEMRC, on supprime 73% des pronoms de départ.

Dans AnnoDis, nous remarquons qu'après la coupe, on retient 78,8% des maillons de départ. Ainsi, la coupe semble ne pas avoir un impact trop important sur la quantité d'information exclue. En revanche, dans DEMRC, nous observons qu'après la

Nature	Avant coupure		Après coupure			Information exclue	
	#	%	#	%	% du total	#	%
SN _{def}	1198	34,7	1019	37,38	29,48	179	14,9
Pro	1026	29,7	794	29,13	22,97	232	22,6
Np	442	12,8	292	10,71	8,45	150	33,9
SN _{dem}	272	7,9	247	9,06	7,15	25	9,2
Poss	182	5,3	123	4,51	3,56	59	32,4
Autre	146	4,2	99	3,63	2,86	47	32,2
SN _{ind}	126	3,7	99	3,63	2,86	27	21,4
SN _∅	64	1,9	53	1,94	1,53	11	17,2
Total	3456	100	2726	100	78,88	730	21,12

TABLEAU 6.6 – Distribution de la nature des maillons dans AnnoDis avant et après la coupure des chaînes

Nature	Avant coupure		Après coupure			Information exclue	
	#	%	#	%	% du total	#	%
Pro	18 566	38,8	4 998	26,9	25,3	13 568	73,1
SN _{def}	11 181	23,4	6 695	59,9	34,0	4 486	40,1
Poss	7 968	16,7	2 530	31,8	12,8	5 438	68,2
Np	2 742	5,7	863	31,5	4,4	1 879	68,5
SN _∅	2 729	5,7	1 663	60,9	8,4	1 066	39,1
SN _{ind}	1 641	3,4	1 313	80,0	6,7	328	20,0
∅	1 396	2,9	510	36,5	2,6	886	63,5
SN _{dem}	979	2,0	701	71,6	3,6	278	28,4
Autre	618	1,3	444	71,8	2,3	174	28,2
Total	47 821	100	19 717	41,2	100	28 104	58,8

TABLEAU 6.7 – Distribution de la nature des maillons dans DEMRC avant et après la coupure des chaînes

coupure, on retient 41,2% des maillons de départ. Le pourcentage de maillons exclus (58,8%) peut sembler donc élevé, mais il convient de rappeler que le pourcentage de CR coupées représente environ 22% des CR analysées. En effet, la plupart des CR présentes dans cette ressource ne dépassent pas le 7 maillons. Ce qui peut expliquer la quantité de maillons exclus est la présence de CR très longues : bien que les CR très longues soient rares, le nombre élevé de maillons impacte fortement la quantité d'information exclue. Par exemple, les cinq chaînes les plus longues (de 626 à 895 maillons) contiennent à elles seules 8% des maillons exclus après la coupure. Pour cette raison, nous pensons que, malgré la coupure, nos analyses ne devraient pas être biaisées. En considérant la distribution de la nature après coupure, nous observons que la fréquence de certaines catégories est impactée.

Nous pouvons constater que, dans les deux ressources, les catégories les plus affectées par la coupure correspondent à des expressions de haute accessibilité, tels les pronoms, les possessifs et les sujets zéro. Ces expressions, généralement utilisées pour maintenir un référent déjà introduit, occupent en effet les positions de maillon-n dans

les CR (cf. 1.1.1) et on peut s'attendre à ce que celles-ci aient plus de probabilité d'être exclues lorsque l'on coupe les CR à sept maillons : par exemple, dans (24), les cinq derniers maillons coupés correspondent à des pronoms, possessifs et sujets zéro.

- (24) La Rochefoucauld aima Mlle de Hautefort sans oser le lui dire ; mais quelque temps après, étant à l'armée et à la veille d'une bataille, il alla trouver **le marquis de Hautefort** avec **lequel** il servait, **lui** fit confiance de sa passion, et **lui** donna une lettre pour **sa** sœur, en **lui** faisant promettre que, s'**il** périssait dans le combat, **il** la lui remettrait et lui **dirait** de **sa** part ce qu'il ne lui avait jamais dit, et que, s'il n'était pas tué, **il** lui rendrait sa lettre à lui-même et lui **garderait** fidèlement son secret. [B_{MdH} ; ⚡]

Nous remarquons que parmi les expressions de faible accessibilité, les Np se singularisent quelle que soit la ressource : leur nombre est fortement réduit, dans les mêmes proportions qu'une expression de haute accessibilité comme le sujet zéro. Cette réduction est bien plus importante dans DEMRC que dans AnnoDis. Une explication possible est que les CR très longues se trouvent dans les textes narratifs et il est possible que le Np soit utilisé pour référer à un personnage lorsqu'il y a compétition de référents ou lorsque la distance entre les maillons est très élevée.

6.2.2 Comparaison et classification des séquences

Pour mener à bien l'analyse, plusieurs paramètres sont à régler, notamment : le type de clustering à utiliser, la manière de calculer la (dis)similarité entre les séquences, le nombre de clusters le plus pertinent (cf. 5.2). Cette première expérimentation a été effectuée en utilisant le paramétrage suivant :

- pour la comparaison des séquences :
 - la distance entre les séquences est calculée en utilisant la mesure d'édition *Optimal Matching* (OM), utilisée dans la plupart des études sur l'analyse des séquences. Nous avons retenu cette méthode car elle permet de comparer des séquences de longueurs différentes, ce qui est le cas dans notre corpus : rappelons que les séquences que nous analysons ont une longueur qui varie entre 3 et 7 états. Par ailleurs, l'OM préserve les aspects d'enchaînement et de position des états, qui sont centraux pour notre analyse (cf. 5.2). Nous rappelons que pour calculer cette distance, il est nécessaire de calculer également le coût des opérations de substitution (matrice de substitution) et d'insertion-suppression ;
 - La matrice de substitution est construite à l'aide de la méthode *TRATE*, fondée sur les probabilités de transition observées entre états. Ce choix permet de fixer les coûts de substitution en limitant l'arbitraire lié à l'affectation de coûts constants (cf. 5.2.3).
 - Le coût des opérations d'insertion et de suppression (*indel*) est fixé à 1. Ce paramètre modéré permet de ne pas survaloriser la position exacte des états, tout en tenant compte de la variabilité de longueur des séquences. Nous avons ainsi choisi une valeur équilibrée, ni trop basse (ce qui conduirait à ignorer les décalages), ni trop élevée (ce qui pénaliserait artificiellement les séquences plus courtes ou plus longues).

6.2.3 Choix du nombre de clusters

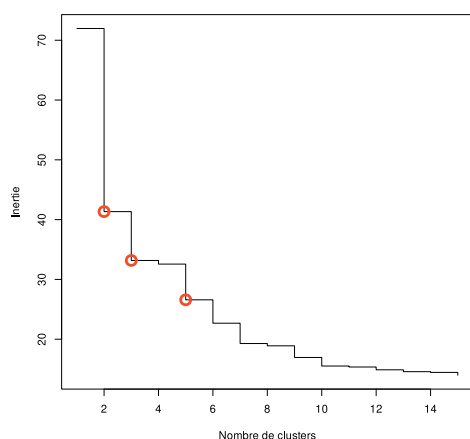
Déterminer le nombre optimal de clusters constitue l'un des enjeux méthodologiques majeurs dans l'analyse des séquences. Comme le soulignent [Brzinsky-Fay \(2007\)](#), les tests statistiques classiques utilisés pour établir le nombre de cluster sont difficilement transposables dans le cadre de l'analyse des séquences. Ces auteurs expliquent que ces critères reposent sur des hypothèses (comme l'indépendance des dimensions ou l'utilisation de distances euclidiennes) qui ne sont pas adaptées à la nature séquentielle des données, où la position, la durée et l'enchaînement jouent un rôle central. Dans ce contexte et face à cette limite, [Aisenbrey et Fasang \(2010, 433\)](#) souligne que le raisonnement s'appuie davantage sur des outils visuels et une réflexion théorique sur la validité des clusters obtenus. La notion de validité est importante : il s'agit d'évaluer la pertinence des clusters en fonction de leur *interprétabilité théorique* et de leur cohérence avec les connaissances du domaine. C'est dans cette perspective que nous présentons les outils que nous avons mobilisés pour guider le choix du nombre de clusters dans notre analyse.

Dans un premier temps, nous nous sommes appuyée sur l'observation des sauts d'inertie, une méthode recommandée dans le cadre du clustering hiérarchique pour guider l'interprétation exploratoire ([Kaufman & Rousseeuw, 2009](#) ; [Studer, 2012b](#)). Ces sauts correspondent à l'augmentation de l'inertie intra-cluster (ou perte d'homogénéité) observée à chaque étape d'agglomération entre clusters. Autrement dit, à chaque fusion, on calcule la différence d'inertie avant et après regroupement : un saut important indique que deux classes relativement dissemblables ont été fusionnées, ce qui présage plus de difficultés pour interpréter les clusters étant donné la plus grande diversité des séquences regroupées. Les sauts d'inertie sont visualisés sous forme de courbe ou de diagramme en escalier. L'objectif est d'identifier des ruptures nettes dans la progression, susceptibles d'indiquer le nombre de clusters à privilégier.

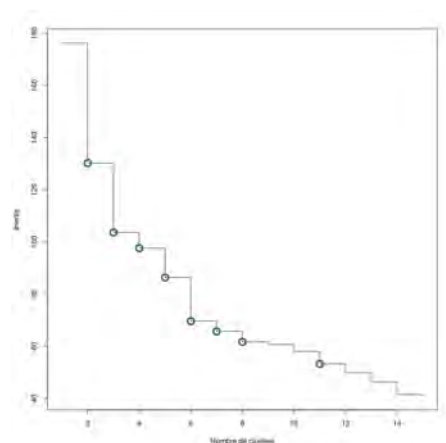
La [figure 6.4_a](#) permet d'identifier trois sauts assez nets pour la ressource AnnoDis : 2, 3 et 5 classes. Deux sauts ultérieurs sont présents : 6 et 7 classes. Étant donné que le nombre de CR de la ressource AnnoDis n'est pas très élevé, nous faisons l'hypothèse que les configurations en trois et cinq classes devraient suffire pour distinguer des types de CR interprétables. Concernant la ressource DEMRC, la [figure 6.4_b](#) montre quatre sauts nets : 2, 3, 5, 6 classes. Des sauts ultérieurs sont présents : 4, 6, 8 et 11 classes. Étant donné le nombre important de CR dans cette ressource, nous pensons qu'une division en deux ou trois classes n'est pas assez fine pour dégager des types de CR.

Si les sauts d'inertie constituent un indicateur utile pour repérer des ruptures potentielles dans la structure des données, ils ne permettent pas, à eux seuls, de déterminer avec certitude le nombre optimal de clusters. Comme nous l'avons vu (cf. [figure 6.4](#)), plusieurs découpages semblent être pertinents dans les deux ressources. Il nous semble difficile de prendre une décision uniquement à partir de ces graphiques.

Un indicateur formel couramment utilisé pour évaluer la qualité du clustering est le score de silhouette ([Rousseeuw, 1987](#)). Il prend en considération deux dimensions : l'homogénéité intra-cluster, à savoir la proximité entre les éléments (nos séquences) dans un cluster ; la séparation, c'est-à-dire à quel point les clusters sont éloignés les uns des autres. En d'autres termes, à quel point les séquences diffèrent entre les deux



(a) Sauts d'inertie AnnoDis



(b) Sauts d'inertie DemRC

FIGURE 6.4 – Sauts d'inertie résultant du clustering sur AnnoDis et DemRC

clusters. Ce score varie entre -1 et 1. Une valeur proche de 1 signifie que les éléments ont été bien repartis dans les clusters, une valeur négative indique une mauvaise répartition, une valeur proche de 0 suggère que la séparation entre les clusters est floue. En appliquant ce test au clustering de nos données, nous obtenons des scores peu discriminants, avec des valeurs relativement basses et peu contrastées entre les différents regroupements (par exemple, 0,21 et 0,15 pour les regroupements en 3 et 5 clusters dans AnnoDis). Cela présage que la similarité entre les séquences de chaque cluster sera parfois difficile à dégager. Dans ce contexte, nous préférons ne pas nous appuyer sur ce score pour fixer le nombre de clusters.

Une option alternative Compte tenu des limites du score de silhouette et des sauts d'inertie, nous privilégions une approche qualitative des résultats pour estimer la pertinence des regroupements. Ainsi, notre démarche s'inscrit dans une approche avant tout exploratoire, fondée sur l'observation des ruptures visuelles et sur la cohérence interprétative des groupes identifiés. Dans cette perspective, nous proposons de compléter l'analyse pour décider du nombre de clusters par une visualisation du dendrogramme obtenu, visualisation qui permet de « visualiser les logiques de regroupement des séquences » (Studer, 2012a) et d'analyser les regroupements à partir de nos connaissances sur les chaînes de référence. Nous présentons le graphique en forme d'arbre dans la figure 6.5. L'image, à lire du haut vers le bas, est une visualisation graphique des logiques de regroupement selon le nombre de clusters choisis. Les chiffres situés à droite indiquent le nombre de clusters envisagés à chaque étape. Le niveau 1 correspond au niveau de départ : aucune classification n'a été effectuée. Le niveau 2 montre les regroupements effectués avec une division en deux classes. Et ainsi de suite. Tous les sauts d'inertie que nous avons présentés dans la section précédente sont représentés dans le graphique. Pour une meilleure compréhension des logiques de regroupement, la figure illustre tous les niveaux, y compris ceux qui correspondent à des sauts d'inertie moins nets. Les graphiques qui composent l'arbre sont des graphiques de distribution,

qui permettent de visualiser la distribution de la nature des maillons dans les différents états qui composent les séquences (de M1 à M7), chaque couleur correspondant à une catégorie grammaticale. Chaque carré dans l'arbre correspond à un cluster : les deux

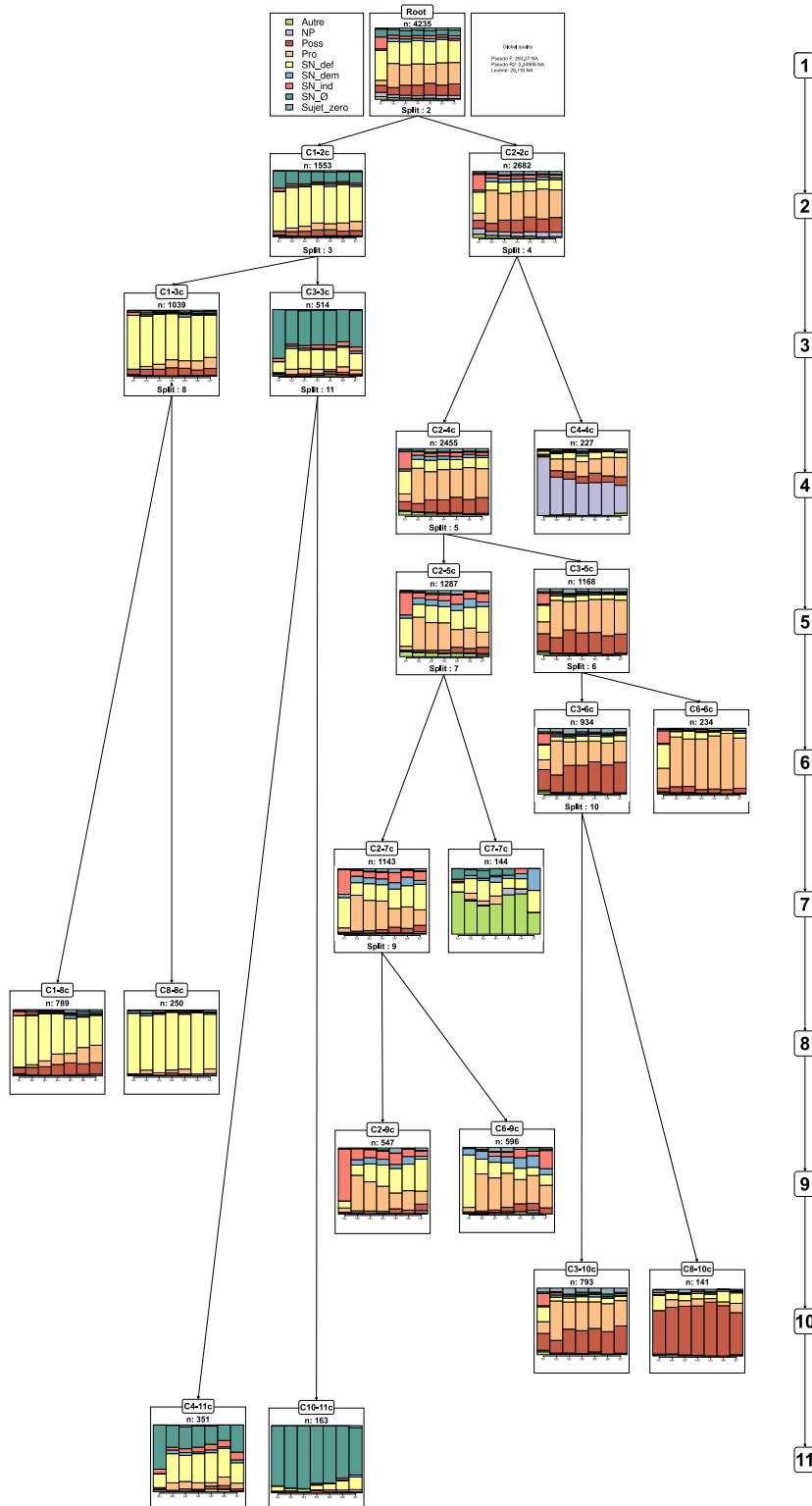


FIGURE 6.5 – Visualisation graphique des logiques de regroupement pour le clustering dans DEMRC

graphiques au niveau 2 correspondent aux 2 classes obtenues avec une classification en 2 clusters. Au dessus du graphique nous trouvons :

- le nom que nous avons attribué au cluster. Nous utilisons la notation suivante : CX-Xc pour indiquer le numéro du cluster (C1 indique le cluster 1) et à quelle configuration il appartient (2c ou 5c pour la configuration en deux ou cinq classes). Par exemple, C1-2c correspond au premier cluster de la classification en deux classes ;
- le nombre de séquences regroupées dans chaque cluster, indiqué avec *n*. Par exemple, *n* : 1553 pour le C1-2c et *n* : 2682 pour le C2-2c.

Au-dessous du graphique de distribution, *Split* indique à quel niveau le cluster est divisé ultérieurement. Par exemple, *Split 3* signifie que le cluster en question est divisé au niveau 3. Cela veut dire que l'on obtient la distinction illustrée au niveau 3 si on décide de classer les séquences en trois clusters. Lorsque aucune information sur le *Split* n'est présente, cela signifie que le cluster reste le même, quel que soit le nombre de clusters sélectionné.

Par exemple, dans la [figure 6.5](#), au niveau 3, seulement deux graphiques sont présentés. Cela indique que le C1-2c a été divisé en deux clusters C1-3c et C3-3C au niveau 3. Le C2-2c, n'étant pas reproduit, reste le même au niveau 3. Comme nous pouvons l'observer dans la [figure 6.5](#), certains clusters se stabilisent assez tôt, comme le cluster C4-4c, qui ne subit aucune division ultérieure après ce niveau-là ; d'autres clusters peuvent au contraire être divisés à des niveaux plus bas : c'est le cas des clusters C3-3c et C3-6c qui ne sont divisés qu'aux niveaux 10 et 11.

Nous abordons maintenant l'étape décisive du choix du nombre de clusters à retenir pour notre analyse. Ce choix repose sur une interprétation guidée par notre connaissance des CR et par les objectifs de typologie de cette étude. Afin de mieux identifier les distinctions susceptibles de révéler différents modes de composition des chaînes, nous avons examiné, pour chacune des ressources, les dendrogrammes obtenus à l'issue du clustering. Nous présentons ci-dessous les observations clés qui ont orienté nos décisions. Dans le cadre d'AnnoDis, cette étape s'est révélée plus aisée que pour DEMRC, en grande partie en raison de la taille plus réduite du corpus et d'une plus grande homogénéité apparente des CR. Comme le montre la [figure 6.6](#), la première division sépare les séquences constituées principalement de SN définis et celles constituées principalement de noms propres et pronoms. Le premier groupe fusionne deux tendances : des séquences constituées presque exclusivement de SN définis et celles où les pronoms sont également assez fréquents (regroupement de niveau 5). La division en trois groupes permet d'identifier deux tendances au sein du deuxième groupe : les séquences composées principalement de SN indéfinis, démonstratifs et de pronoms sont distinguées des séquences constituées principalement de noms propres et pronoms. Toutefois, lorsqu'on passe de trois à quatre classes, nous observons que la classe composée principalement de noms propres et pronoms fusionne deux tendances différentes : d'une part, les séquences qui commencent par un SN défini et une forte présence de pronoms dans les états successifs ; de l'autre part, les séquences dont le premier état est un nom propre et les états suivants ont une répartition presque équilibrée de noms propres et pronoms. Sur la base de ces observations, nous avons identifié des distinctions pertinentes au niveau 3 et 5, ce qui vient confirmer les observations

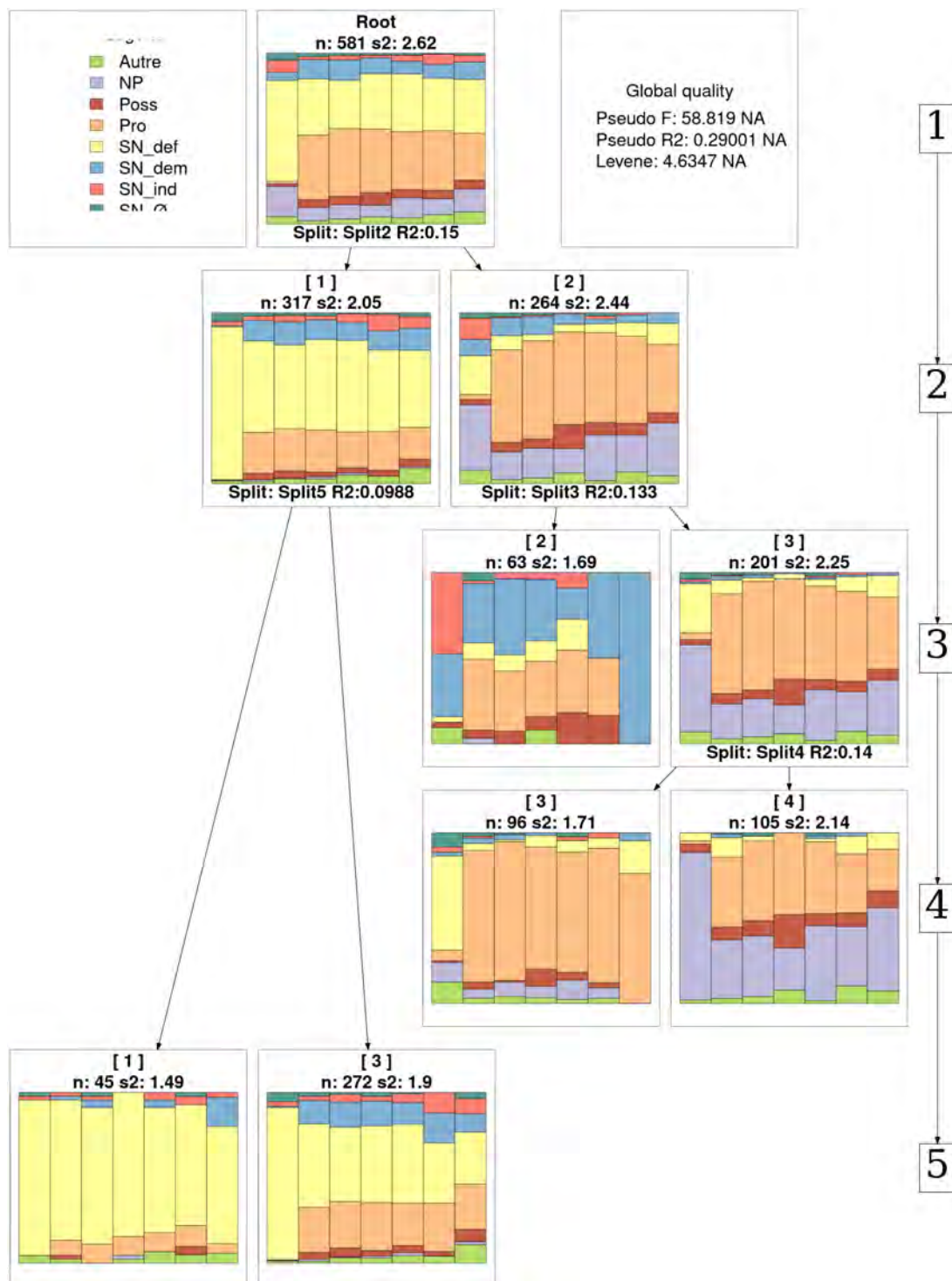


FIGURE 6.6 – Visualisation graphique des logiques de regroupement pour le clustering dans AnnoDis

des sauts d'inertie. Cela nous semble être un bon compromis entre un niveau de simplification trop élevé des données et un niveau de détail trop fin.

Dans la [figure 6.5](#), nous avons analysé les logiques de regroupement pour la ressource DEMRC. Étant donné le nombre plus important de données, nous avons distingué deux types de divisions :

- des distinctions majeures, qui font émerger des classes nettement différenciées, révélant des types de chaînes de référence difficilement identifiables autrement ;
- des distinctions secondaires, qui mettent en évidence des tendances plus fines.

Nous observons qu’une première distinction, au niveau 2, oppose des CR constituées principalement de SN_{def} et d’autres composées principalement de pronoms. Cette partition en deux classes nous semble toutefois trop générale au vu du nombre de CR analysées. Les divisions aux niveaux 3 et 4 affinent cette opposition. La solution en quatre clusters, issue de ces regroupements, s’est révélée particulièrement pertinente. Elle permet d’identifier quatre types de chaînes de référence aux structures clairement différenciées, que nous considérons comme les tendances principales de la ressource. Par exemple, C1-4c regroupe des séquences dominées par des SN_{def} sur l’ensemble des positions ; C3-4c contient des chaînes composées principalement de $SN_{\emptyset}$; C2-4c se caractérise par des débuts en SN_{def} ou SN_{ind} , suivis de pronoms fréquents ; C4-4c rassemble des chaînes fortement marquées par des Np à toutes les positions. Ce cluster est resté stable jusqu’au niveau 11, ce qui confirme sa cohérence interne. Au niveau 7, une nouvelle division apparaît avec l’émergence d’un cluster centré sur l’état “Autre” (C7-7c). Ce type ne se dégageait pas aux niveaux inférieurs, ce qui en fait une distinction que nous retenons comme un cinquième type principal. Les distinctions observées au niveau 9 permettent de différencier les séquences en fonction de leur premier état, ce qui correspond à des stratégies d’introduction des référents différentes.

Sur la base de ces observations, nous retenons deux niveaux d’analyse pour DEMRC :

- la classification en quatre clusters, qui permet de décrire les principaux types de composition des chaînes ;
- la classification en onze clusters, qui permet de considérer l’ensemble des distinctions observées, sans devoir passer en revue chaque niveau individuellement. Cette configuration facilite ainsi l’étude des variations internes à chaque type de CR, permettant d’identifier des stratégies différentes.

6.3 Bilan

Dans ce chapitre nous avons présenté les étapes nécessaires à l’application de l’analyse des séquences aux chaînes de référence. Le chapitre est structuré autour de deux volets. Le premier porte sur l’enrichissement des ressources AnnoDis et DEMRC via l’annotation de la nature des référents. Cette étape est essentielle pour pouvoir analyser l’impact de ce facteur de variation sur la composition des CR. Le deuxième détaille le paramétrage de l’analyse des séquences. Nous avons présenté la création des séquences proprement dites, avec la gestion du critère de description et de la variabilité de la longueur des CR. Rappelons que notre étude se focalise sur la nature des maillons et que nous avons réduit la variabilité de longueur en fixant une taille maximale des séquences à sept états. Ce seuil a été fixé sur la base de la distribution des CR en fonction de leur longueur. Nous avons ensuite présenté nos choix pour les mesures de

distance afin de comparer les séquences : nous avons justifié l’usage de l’OM ainsi que les critères retenus pour le calcul des coûts de substitution et d’insertion-suppression. Enfin, nous avons expliqué l’utilisation du clustering hiérarchique. Nous avons porté une attention particulière au choix du nombre de clusters, étape décisive pour obtenir une classification pertinente qui mette en lumière des types de chaînes différents. Nous avons montré que ce choix ne peut s’appuyer uniquement sur les indicateurs quantitatifs classiques, tels les sauts d’inertie et le score de silhouette, dont nous avons souligné les limites spécifiques dans l’analyse des séquences (Aisenbrey & Fasang, 2010). Nous avons pu confirmer que ces mesures ne sont pas discriminantes dans nos données. Par exemple, nous obtenons un score de silhouette proche de 0 pour le clustering en trois et en cinq classes dans AnnoDis, ce qui signifie que la distinction entre les clusters est floue. Néanmoins, par la suite, grâce à la visualisation des classes, nous avons vu que la classification en trois et cinq s’avère pertinente : elle permet de mettre en évidence des types de chaînes interprétables. Pour cette raison, nous avons adopté une démarche exploratoire et qualitative, fondée sur l’observation des regroupements hiérarchiques par le biais de visualisations graphiques. Cette approche nous permet d’intégrer nos connaissances linguistiques sur les chaînes de référence pour guider l’interprétation des regroupements et fixer ainsi le nombre de clusters qui nous semble le plus pertinent. Ce positionnement constitue un apport méthodologique central dans notre travail : il permet de dépasser une interprétation purement formelle des distances entre les séquences.

Notre démarche exploratoire et qualitative nous a amenée à distinguer deux niveaux d’analyse :

- une classification à gros grain, permettant de mettre en évidence les tendances générales et les principaux types de composition des CR ;
- une classification à petit grain, permettant d’analyser des distinctions plus fines pour chaque type de CR identifié avec la classification à gros grain.

Le choix du nombre de clusters est propre à chaque ressource sur la base des observations des logiques de regroupement. Les résultats présentés dans ce qui suit sont obtenus en utilisant la combinaison de paramètres résumée dans le [tableau 6.8](#).

Création des séquences	
Critère de description	Nature des maillons
Longueur des séquences	entre 2 et 7 (AnnoDis) ; entre 3 et 7 (DemRC)
Comparaison des séquences	
Matrice de substitution	TRATE
<i>Indel</i>	1
Mesure de (dis)similarité	Optimal Matching (OM)
Classification des séquences	
Type de clustering	Hiérarchique agglomératif
Méthode d’agrégation	Ward
Nombre de clusters	3 et 5 (AnnoDis) ; 4 et 11 (DEMRC)

TABLEAU 6.8 – Paramètres utilisés pour l’analyse des séquences dans les ressources AnnoDis et DEMRC

Troisième partie

Vers une typologie des chaînes de référence

Dans cette dernière partie, nous nous dirigeons vers une typologie des chaînes de référence, en passant par la présentation des résultats obtenus en appliquant l'analyse des séquences aux ressources AnnoDis et DEMRC. Dans le [chapitre 7](#), nous explorons les séquences dans AnnoDis. Nous montrons que l'analyse en séquences se révèle pertinente et permet d'obtenir des résultats parlants. Dans le [chapitre 8](#), nous présentons les résultats de l'application de cette même approche à la ressource DEMRC. Dans un premier temps, nous proposons une description des classes obtenues, dont nous fournissons une interprétation dans un deuxième temps. Par la suite, nous nous focalisons sur les liens entre les classes obtenues et les facteurs de variation analysés. Nous concluons par un bilan.

Chapitre 7

Explorer les enchaînements : application de l'analyse des séquences à la ressource AnnoDis

Contents

7.1	Exploration des séquences	148
7.2	Description des clusters obtenus	148
7.2.1	Cluster C1-3c AnnoDis : SN _{def} ++	151
7.2.2	Cluster C2-3c AnnoDis : SN _{ind} , SN _{dem} et Pro	152
7.2.3	Cluster C3-3c AnnoDis : Np et Pro	153
7.3	Lien entre les clusters AnnoDis et les facteurs de variation	155
7.3.1	Impact de la nature du référent sur la composition des chaînes	155
7.3.2	Impact du genre textuel sur la composition des chaînes	156
7.4	Les patrons des chaînes AnnoDis	158
7.5	Bilan	162

Ce chapitre présente les résultats obtenus lors d'une première étude qui a consisté à appliquer l'approche linéaire sur les chaînes AnnoDis¹. Nous discutons notamment des résultats obtenus par la classification des chaînes, considérées comme des séquences d'états (cf. [chapitre 5](#)). Nous présentons les résultats en trois étapes : tout d'abord, nous montrons que les classes obtenues sont linguistiquement interprétables et peuvent être mises en correspondance avec différents types de CR. Nous fournissons une description des différents clusters obtenus, nous proposons des interprétations et des liens avec la littérature. Ensuite, nous analysons le lien entre les types de CR identifiés et les facteurs de variation (type de texte, genre textuel et type de référent) (cf. [7.3](#)). Enfin, nous présentons les patrons des CR les plus fréquents qui émergent de la ressource AnnoDis (cf. [7.4](#)). Nous montrons que ces patrons correspondent à différentes stratégies d'introduction et de maintien du référent.

1. Les résultats que nous présentons dans ce chapitre ont fait l'objet d'une publication : [Federzoni, Ho-Dac, et Fabre \(2023\)](#)

7.1 Exploration des séquences

L'exploration des séquences constitue la première étape de l'analyse des séquences (cf. 5.2.2). Rappelons qu'elle permet de décrire les caractéristiques globales de l'ensemble des séquences en entrée. Nous nous intéressons notamment à la distribution de la nature des maillons. Cette étape est importante pour connaître les données que nous allons manipuler. Dans cette section, nous décrivons les caractéristiques de l'ensemble des séquences AnnoDis. Rappelons que les séquences analysées ont entre deux et sept

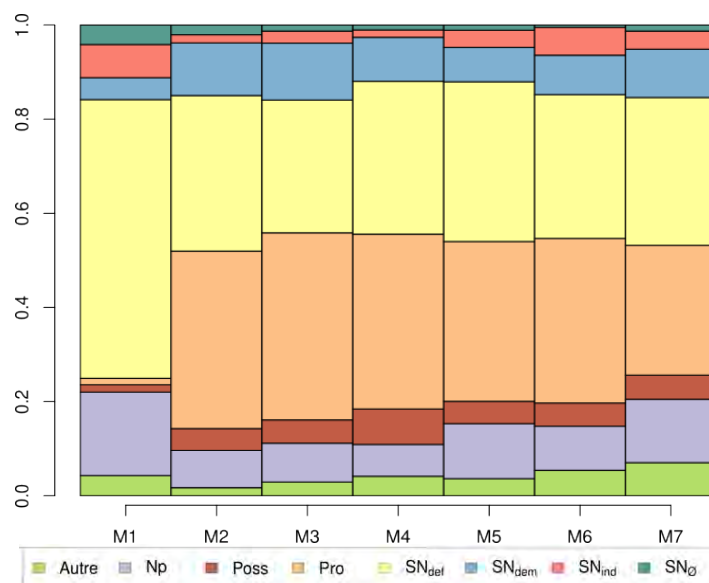


FIGURE 7.1 – Distribution de la nature des maillons dans AnnoDis après la coupure des chaînes à sept maillons

états (cf. 6.2.1).

La figure 7.1 illustre la distribution des différentes natures de maillons selon leur position dans la chaîne. Nous pouvons observer que, comme attendu, la nature du premier maillon correspond principalement à des expressions de faible accessibilité, comme les SN_def et les noms propres. Les SN_ind sont également présents en première position. En observant la distribution des maillons qui suivent le premier, nous pouvons observer que les pronoms et les SN_def sont les plus fréquents, expressions de haute et de faible accessibilité respectivement. Leur présence dans des positions autres que la première dans la CR pourrait être due à différents facteurs : nature du référent, genre textuel, référents en compétition. Une analyse approfondie des stratégies d'introduction et de maintien du référent à l'aide des clusters de séquences nous permettra de mieux décrire ces tendances.

7.2 Description des clusters obtenus

Nous présentons ici les résultats de l'application de l'analyse des séquences à la ressource AnnoDis. Les résultats sont obtenus en deux temps : une classification à

gros grain en trois clusters nous permet de décrire les tendances générales ; ensuite, une classification à petit grain en cinq clusters permet d'analyser des distinctions plus fines à l'intérieur de certaines classes (cf. 6.2). La classification des 581 séquences, obtenue avec la configuration à trois clusters (3c), permet d'identifier trois classes qui différencient très clairement des types de chaînes du point de vue de leur caractérisation linéaire. La [figure 7.2](#) permet d'avoir une vision d'ensemble des clusters obtenus. Rappelons que ce type de graphique (appelé *indexplot*) permet d'identifier visuellement les régularités et les différences sur l'ensemble des données (cf. 5.2.5). Sur l'axe des x nous trouvons la séquence d'états qui composent la séquence, du premier (M1) au dernier (M7). L'axe des y représente les séquences individuelles. Les séquences sont ordonnées sur cet axe par leur distance. Cette distance correspond au score obtenu entre paires de séquences lors du calcul de l'OM (cf. 5.2.3.3).

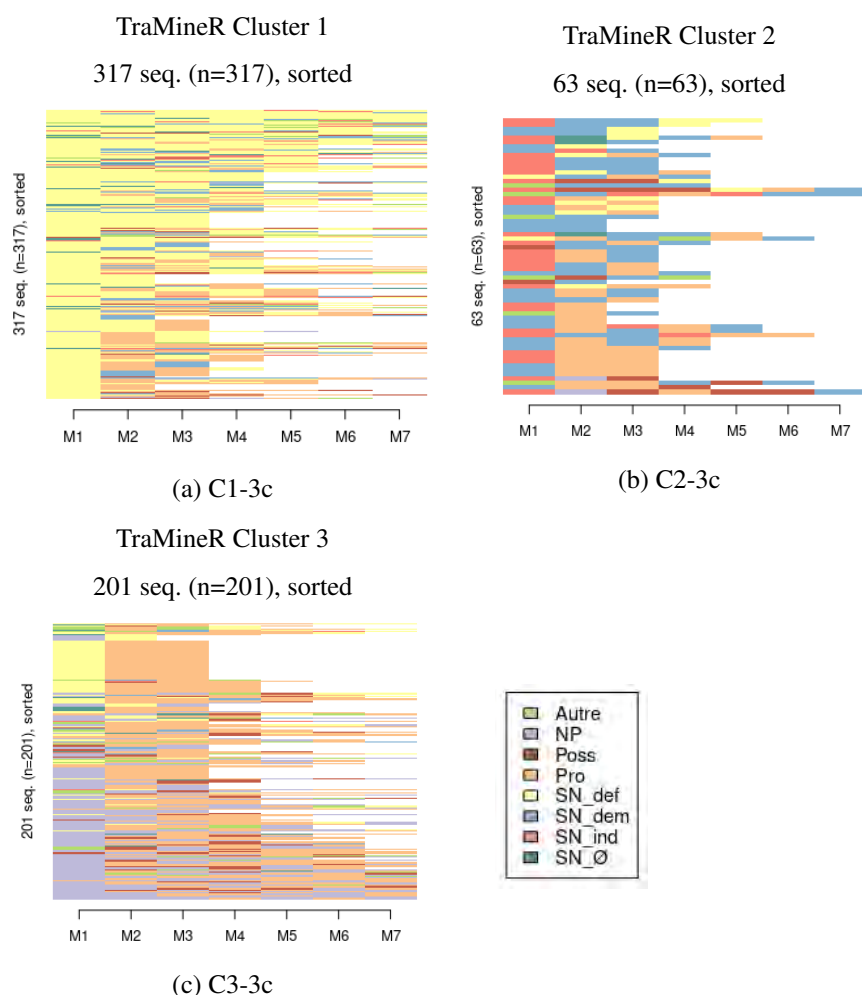


FIGURE 7.2 – Visualisation des résultats de la classification des séquences AnnoDis en 3 classes par des *index plots*

Chaque catégorie grammaticale est différenciée par une couleur.

Visuellement, nous identifions les tendances suivantes :

- des séquences composées principalement de SN_{def} (C1-3c², figure 7.2_a);
- des séquences composées principalement de SN_{ind} , SN_{dem} et pronoms (C2-3c, figure 7.2_b);
- des séquences composées principalement de Np , SN_{def} et pronoms (C3-3c, figure 7.2_c).

Nous remarquons que les trois classes sont déséquilibrées en nombre de séquences : le premier cluster regroupe 54% des séquences, le deuxième seulement 11% et le troisième 35%. L'analyse de l'entropie transversale (figure 7.3) — qui, nous le rappelons, permet de décrire la façon dont l'homogénéité de la classe évolue selon les états (cf. chapitre 5) — montre que globalement les trois clusters sont caractérisés par une certaine hétérogénéité, qui évolue différemment selon le cluster. Le C1-3c est homogène pour les premiers maillons, mais il présente de fortes variations (entropie entre 0,6 et 0,8) dans les positions suivantes. Le cluster C2-3c est plutôt hétérogène, avec une forte diminution de l'entropie pour les deux derniers états des séquences (M6 et M7). En observant le cluster obtenu (figure 7.2_b), nous remarquons en effet que les maillons qui occupent ces positions ne présentent pas beaucoup de variation : en M6 nous retrouvons soit un pronom soit un SN_{dem} et en M7 uniquement des SN_{dem} . Le C3-3c présente moins de variations importantes, avec une entropie élevée pour les premiers maillons qui diminue légèrement dans les positions suivantes et remonte pour la dernière position.

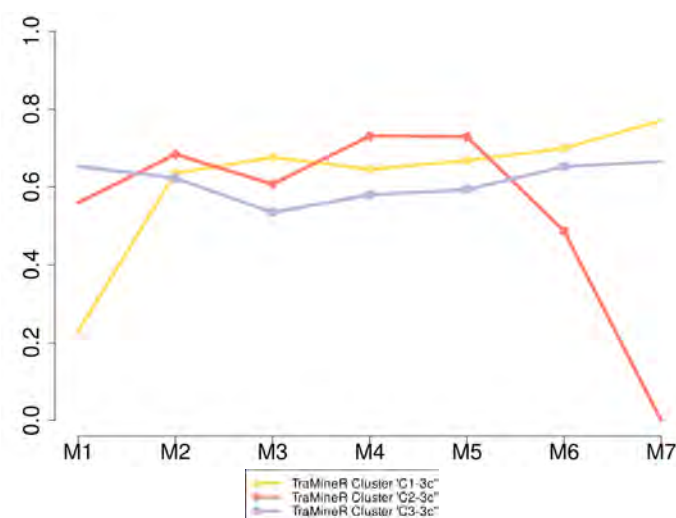


FIGURE 7.3 – Entropie transversale des séquences AnnoDis selon les trois clusters obtenus

Nous discutons plus en détail la composition des clusters obtenus dans les sections qui suivent. Nous analysons les classes obtenues en prenant en compte d'une part, la nature du premier maillon et d'autre part, la nature des maillons qui suivent.

2. Nous utilisons cette notation pour indiquer le numéro du cluster (C1) et à quelle configuration il appartient : configuration en trois ou en cinq classes (3c ou 5c).

7.2.1 Cluster C1-3c AnnoDis : SN_{def}++

Le premier cluster (C1-3c) regroupe le plus grand nombre de séquences (317, soit 54,6% des séquences). Les CR du cluster C1-3c (figure 7.2_a) commencent généralement avec un SN défini (284 CR, 89,6% des séquences du cluster), suivi globalement d'autres SN définis, comme le montre l'exemple (25) où la chaîne est composée exclusivement de SN définis³. Il s'agit de CR dont la composition est globalement homogène : tous les maillons tendent à relever de la même nature.

- (25) Juste avant le début de la révolution française, le vignoble champenois s'étendait sur quelques 50 000 hectares. Dans la seconde moitié du xix^e siècle, le vignoble connaît avec 65 000 hectares son expansion maximale comprenant aussi 2 500 hectares dans le département des Ardennes. Après le fléau du phylloxéra et de la Grande guerre, le vignoble s'est réduit à 12 000 hectares. Aujourd'hui, en 2007, le vignoble champenois s'étend sur 32 341 hectares. [AE_{w} ; 

Les CR dont le premier maillon est un SN défini peuvent néanmoins avoir une composition différente, avec une succession de pronoms ou de SN démonstratifs.

Les tendances en détail Cette distinction est confirmée par la classification en cinq clusters. Comme le montre la figure 7.4, les séquences du C1-3c sont divisées en deux clusters différents :

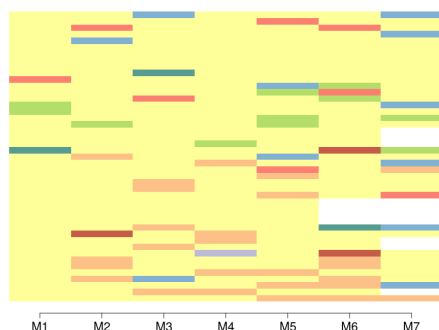
- C1a-5c qui regroupe 45 séquences, la plupart composées exclusivement de SN définis (comme dans l'exemple (25)) ;
- C1b-5c qui regroupe 272 séquences qui commencent par un SN défini et qui présentent une composition plus hétérogène par la suite, avec une variété de catégories grammaticales dans les maillons suivant le premier. L'exemple (26) illustre une CR présentant ce type de composition : introduction avec un SN défini suivi d'une alternance de SN démonstratifs, de possessifs et de pronoms.

- (26) L'affaire Dreyfus a pour origine une erreur judiciaire sur fond d'espionnage et d'antisémitisme, dont la victime est le capitaine Alfred Dreyfus (1859 - 1935), français et alsacien d'origine, et Juif. Cette affaire a bouleversé la société française pendant douze ans, de 1894 à 1906 La révélation de ce scandale, dans « J'Accuse... ! », un article d'Émile Zola en 1898, provoque une succession de crises politiques et sociales uniques en France. À son paroxysme en 1899, elle révèle les clivages de la France de la Troisième République. Elle divise profondément et durablement les Français en deux camps opposés, dreyfusards et anti-dreyfusards. Cette affaire est le symbole moderne et universel de l'iniquité au nom de la raison d'État. Enfin, elle suscite de très violentes polémiques nationalistes et antisémites diffusées par une presse influente. [AE_{w} ; 

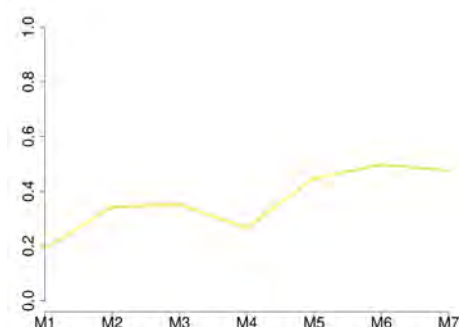
3. Dans les exemples, nous soulignons les maillons avec la couleur utilisée dans le graphique pour indiquer leur nature.

TraMineR Cluster : C1a-5c

45 seq. (n=45), sorted



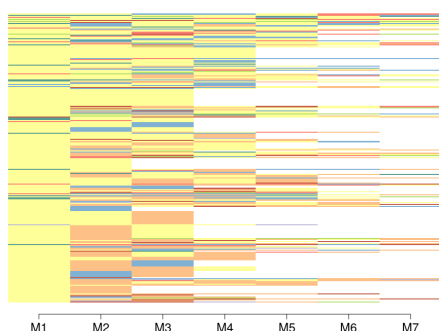
(a) C1a-5c



(b) Entropie C1a-5c

TraMineR Cluster : C1b-5c

272 seq. (n=272), sorted



(c) C1b-5c



(d) Entropie C1b-5c

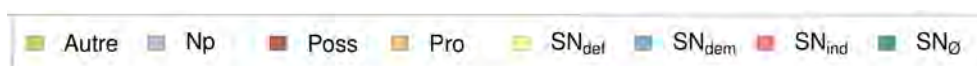


FIGURE 7.4 – Clusters résultant de la division du C1-3c divisé en C1a-5c et C1b-5c dans AnnoDis

7.2.2 Cluster C2-3c AnnoDis : SN_{ind} , SN_{dem} et Pro

Le deuxième cluster (C2-3c) est le plus petit cluster. Il regroupe 63 séquences parmi lesquelles 47,6% (30) commencent par un SN indéfini et 36,5% (23) par un SN démonstratif, suivi principalement par d'autres SN démonstratifs ou des pronoms. L'exemple (27) illustre une chaîne qui commence par un SN indéfini (*une connotation oppositionnelle*) suivi de deux démonstratifs (*Ceci, Cette dichotomie*). Ce cluster reste stable lors de la classification en 5 clusters, c'est-à-dire que ce cluster reste exactement le même dans les deux configurations.

- (27) [...] « le " communicatif " présente ainsi fréquemment une connotation oppositionnelle (sinon contradictoire) avec le linguistique. Ceci est particulièrement crucial lorsqu'on traite de la " compétence de communication " [...] Cette dichotomie pose problème au psychologue [...]. [AS_L] ; ⚡

7.2.3 Cluster C3-3c AnnoDis : Np et Pro

Le troisième cluster (C3-3c) regroupe 201 séquences (soit 34,6% des CR). 50,7% des séquences commencent par un nom propre et 28,8% par un SN défini suivi d'autres noms propres ou de pronoms. Contrairement au C1-3c où les séquences commencent également par des SN définis, nous observons que les séquences commençant par des SN définis dans le C3-3c ne sont pas constituées de SN définis dans la suite. On remarque la présence presque exclusive de pronoms.

- (28) Chez F. de Saussure, l'analogie [...], il pose que les facteurs de trouble [...]. Pour lui, cette tendance à l'irrégularité est heureusement contrebalancée par l'analogie [...]. Comme H. Paul, il ramène le concept au calcul de l'équation de la quatrième proportionnelle. [AS_L] ; ⚡

Les tendances en détail Les tendances que nous venons de mentionner sont davantage visibles lorsque nous observons les distinctions plus fines (figure 7.5).

Comme le montre la figure 7.5, le C3-3c est divisé en deux classes plus fines :

- C3a-5c qui regroupe 96 CR, la plupart commencent par un SN défini ou NP, suivi principalement de pronoms.
- C3b-5c qui regroupe 105 CR, commençant par un NP, mais contrairement au C3a-5c, la suite de ces séquences présente une alternance de noms propres, pronoms et possessifs.

La forte présence de pronoms dans le C3a-5c suggère que le référent est très accessible et qu'il n'y a pas d'autres référents en compétition, comme l'illustre l'exemple (29).

- (29) Le colonel Picquart est lui aussi réhabilité officiellement et réintégré dans l'armée au grade de général de brigade. Il est même ministre de la Guerre de 1906 à 1909 dans le premier gouvernement Clemenceau. Il meurt en 1914 d'un accident de cheval. [AE_W] ; ⚡

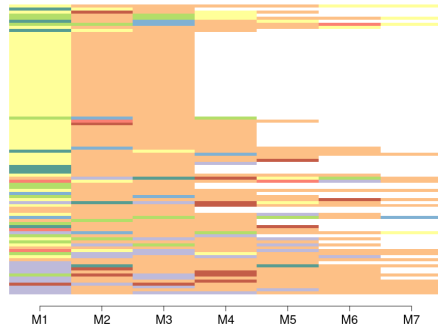
Au contraire, l'alternance entre noms propres, pronoms et possessifs dans le C3b-5c suggère plutôt que le référent n'est pas très accessible. Cela peut s'expliquer de plusieurs manières : à cause de la présence d'autres référents en compétition ou à cause d'une distance entre maillons très importante.

Pour résumer

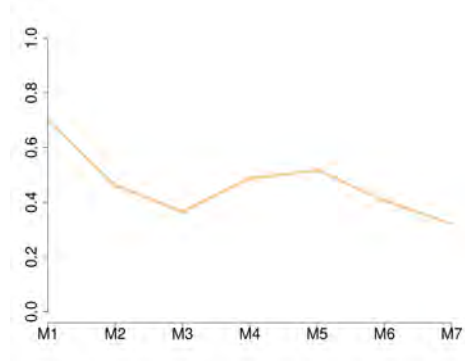
Les trois clusters obtenus dans AnnoDis semblent révéler trois stratégies différentes pour l'introduction et le maintien du référent. Les CR regroupées dans le C1-3c sont composées principalement de SN définis, utilisés aussi bien pour l'introduction du référent que pour son maintien. Les chaînes dans le C3-3c utilisent un nom propre

TraMineR Cluster : C3a-5c

96 seq. (n=96), sorted



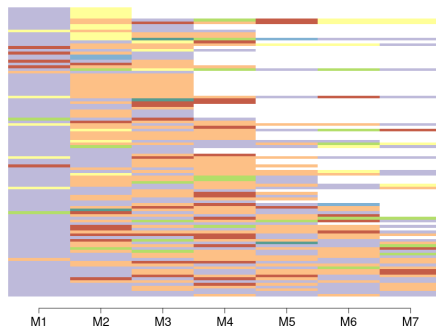
(a) C3a-5c



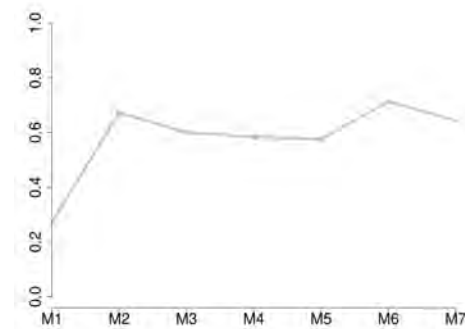
(b) Entropie C3a-5c

TraMineR Cluster : C3b-5c

105 seq. (n=105), sorted



(c) C3b-5c



(d) Entropie C3b-5c

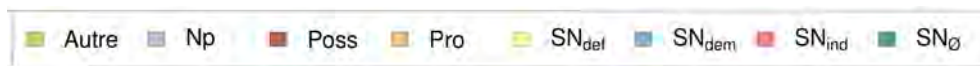


FIGURE 7.5 – Clusters résultant de la division du C3-3c divisé en C3a-5c et C3b-5c

pour l'introduction du référent, qui est par la suite maintenu principalement par des pronoms. Les CR classées dans ce cluster semblent correspondre à des CR qui réfèrent à des humains, comme dans l'exemple (28). Nous pouvons également interpréter la présence de suites de pronoms comme un indice de l'accessibilité élevée des référents. Ce qui peut indiquer que ces CR concernent le *personnage principal*, qui reste activé tout le long du texte. Il serait intéressant d'analyser le lien entre distance, nombre de référents actifs et la présence de ces séquences. Les CR du cluster C2-3c, comme nous l'avons vu, introduisent le référent par le biais de SN indéfinis ou démonstratifs. Nos premières observations suggèrent que ce type de composition est utilisé pour présenter des *idées ou des concepts*, donc des référents abstraits. Dans ces CR le premier maillon est associé à des expressions de basse accessibilité, tels les SN indéfinis, ou un

SN démonstratif (accessibilité moyenne) qui encapsule des arguments précédents avec une anaphore résumante qui introduit un nouveau référent (*cette question*). Le maintien du référent est fait par le biais d'expressions référentielles de moyenne ou faible accessibilité, tels des SN démonstratifs ou des pronoms.

À partir de ces résultats et de l'observation des exemples, nous avons fait l'hypothèse que les SN définis sont utilisés aussi bien pour introduire que pour maintenir les référents non humains, alors que les noms propres sont utilisés pour introduire et maintenir les référents humains. Ces hypothèses sont guidées par les résultats obtenus dans les études sur les CR, comme ceux de [Schneidecker \(1997, 2005\)](#). Les SN indéfinis sont plutôt utilisés, dans le corpus AnnoDis, pour introduire des référents abstraits, les pronoms pour maintenir les référents humains non ambigus, les SN démonstratifs pour maintenir ou encapsuler les référents abstraits. Comme l'ont montré plusieurs auteurs, le genre textuel peut avoir une influence sur la façon de gérer la construction de la continuité référentielle et sur les choix des expressions référentielles. Nous faisons l'hypothèse que, dans le corpus AnnoDis, nous allons retrouver les tendances mises en avant dans la littérature : nous nous attendons à ce que les CR des articles encyclopédiques présentent des types de composition différents par rapport à celles des articles scientifiques, notamment en linguistique.

Les sections suivantes sont dédiées à l'évaluation de ces tendances : nous testons l'impact de deux facteurs de variation sur les clusters obtenus, notamment la nature du référent et les genres textuels.

7.3 Lien entre les clusters AnnoDis et les facteurs de variation

Cette section présente les résultats obtenus en analysant le lien entre les clusters et les deux facteurs de variation retenus. Nous discutons les résultats obtenus avec la classification en trois et en cinq clusters.

7.3.1 Impact de la nature du référent sur la composition des chaînes

Afin de voir s'il existe une liaison entre la nature du référent et les différents types de séquences, nous avons effectué un test du χ^2 entre les trois clusters et la nature du référent. Le test a révélé l'existence d'une liaison significative ($df = 2$, $p\text{-value} < 2,2e-16$). Le test de Cramer, qui permet d'analyser la force de la relation, est égal à 0,45 ce qui indique une relation forte. L'analyse des résidus de Pearson met en évidence des relations d'attraction (liaison positive) et des relations de répulsion (liaison négative) significatives entre la nature des référents et chaque cluster. La [figure 7.6](#) permet de visualiser ces relations et leur degré de significativité : en bleu les relations d'attraction et en rouge de répulsion. La taille des cercles et les différentes nuances de couleur indiquent la force de la relation. Plus la relation est significative, plus la taille des cercles est grande et leur couleur foncée.

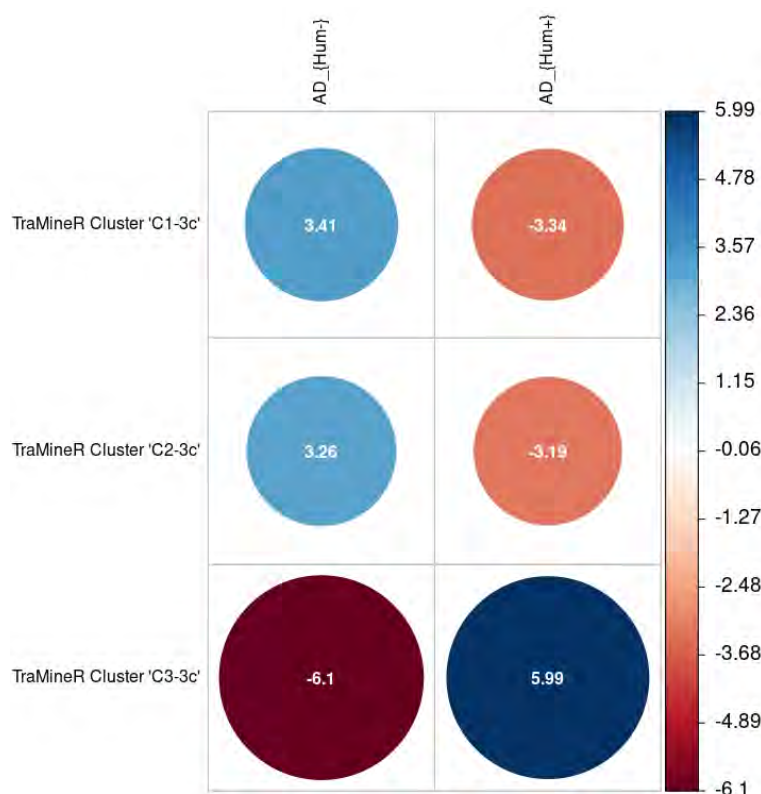


FIGURE 7.6 – Attraction et répulsion entre les 3 clusters AnnoDis et la nature du référent

Il existe une attraction significative entre le C1-3c et le C2-3c et les CR à référent non humain. Cela suggère que pour référer à des non humains les locuteurs ont tendance à utiliser des SN définis, indéfinis et démonstratifs, aussi bien pour introduire les référents que pour les maintenir activés tout au long du discours. À l’opposé, le C3-3c présente une attraction avec les référents humains. Cela suggère que la référence aux humains est faite principalement par des noms propres et leur maintien s’effectue principalement par des pronoms, des possessifs ou d’autres noms propres. La liaison significative est confirmée lorsqu’on prend en considération la combinaison en 5 clusters, avec encore une fois une relation forte ($df = 2$, $p\text{-value} < 2,2e-16$, $V\text{-Cramer} = 0,48$). Nous retrouvons des relations d’attraction et de répulsion similaires à celles que nous venons de présenter : C1b-5c ($SN_{\text{def}}\text{-hétérogène}$) et C2-5c (SN_{ind} et SN_{dem}) présentent une attraction avec les référents non humains, indiquant que pour ce type de référents, l’introduction est faite de préférence par un SN défini ou indéfini et le maintien assuré par des formes variées. Le C3a-5c ($SN_{\text{def}}\text{-Pro}$) et 3b-5c ($Np++$) présentent une attraction avec les référents humains. Le type de composition présent dans C3a-5c rappelle la composition du topique continu proposée par Givón (1983).

7.3.2 Impact du genre textuel sur la composition des chaînes

Comme nous l’avons vu dans le chapitre 4, le corpus AnnoDis est constitué de trois sous-corpus qui correspondent à deux genres de textes, relevant de trois domaines différents de type non narratif (expositif) : articles encyclopédiques, articles scientifiques

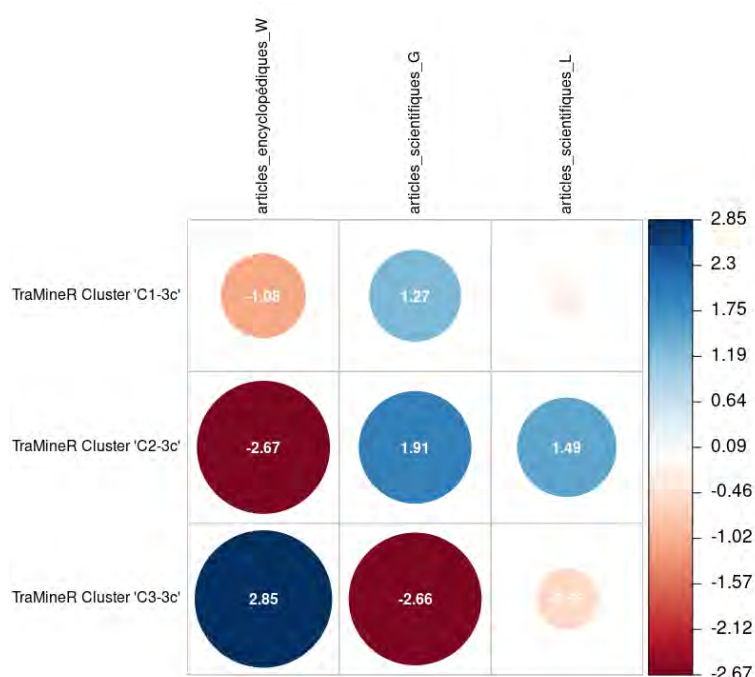


FIGURE 7.7 – Attraction et répulsion entre les 3 clusters AnnoDis et le genre textuel

en géopolitique et articles scientifique en linguistique. Cela nous permet d'effectuer une étude préliminaire, à petite échelle, de l'impact des genres textuels sur la composition des chaînes de référence.

Nous avons réalisé un test du χ^2 entre les clusters et la variable du genre textuel. Le test a révélé une liaison significative entre les classes obtenues dans la configuration en 3 clusters et le genre textuel ($df = 4$, $p\text{-value} = 2,757e-06$). Le test de Cramer indique que la relation est faible ($V\text{-Cramer} = 0,16$). L'analyse des résidus de Pearson nous permet d'analyser en détail cette liaison. Nous avons trouvé une attraction entre le C3-3c (Np et Pro) et les textes encyclopédiques et une *répulsion* entre ce même genre et la classe C2-3c (SN_{ind} et SN_{dem}). Nous avons aussi trouvé une *répulsion* entre les rapports de géopolitique et le cluster C3-3c. La liaison significative est confirmée lorsqu'on prend en considération la combinaison en 5 clusters, avec, encore une fois, une relation faible ($df = 8$, $p\text{-value} = 1,945e-05$, $V\text{-Cramer} = 0,17$). Nous retrouvons des relations d'attraction et de répulsion similaires à celles que nous venons de présenter : C3b-5c (Np++) présente une attraction avec les articles encyclopédiques et une répulsion avec les rapports de géopolitique. C2-5c présente une répulsion avec les articles encyclopédiques.

Si nous considérons les deux facteurs de variation ensemble, nous pouvons remarquer que le C3-3c a une attraction aussi bien pour les référents humains que pour les articles encyclopédiques, alors qu'il présente une répulsion avec les articles scientifiques en géopolitique. Ce résultat non seulement confirme les résultats obtenus dans la littérature, mais confirme notre hypothèse qui était celle que ces clusters permettent de distinguer les types de composition différents selon la nature des référents et du genre textuel. Les résultats indiquent donc que les CR qui réfèrent à des humains présentent

des types de composition différents selon le *genre textuel* (sous-corpus). Les CR composées de Np et pronoms prédominent uniquement dans les textes encyclopédiques. Cela suggère que les textes encyclopédiques dans AnnoDis favorisent la référence aux humains en tant d’individus, alors que les textes de géopolitique se focalisent davantage sur des référents collectifs, allant dans la direction des observations proposées par Longo et Todorascu (2014) sur les textes non narratifs légaux et administratifs.

7.4 Les patrons des chaînes AnnoDis

Suite à l’analyse des séquences, nous avons mené une analyse plus fine des enchaînements les plus fréquents qui illustrent la composition linéaire des chaînes. Pour cette étape aussi, nous avons utilisé la boîte à outil TraMineR (Gabadinho, Ritschard, Studer, & Muller, 2010) pour extraire la liste des séquences au format SPS (cf. 5.3). Comme attendu, le nombre de séquences proposées par le système est très élevé : le système renvoie 393 séquences différentes sur les 581 chaînes analysées.

Rappelons qu’afin de proposer des patrons plus généraux, nous avons regroupé ensemble les mentions relevant de la même nature et nous avons indiqué le nombre d’occurrences dans le format suivant : NAT{n}. Ainsi, le patron subsumant les chaînes constituées de 3, 4, ou 5 syntagmes nominaux définis correspondrait au patron suivant : SN_{def}{3:5}, où le premier *n* indique le nombre minimal que l’on trouve dans la séquence et le deuxième le nombre maximal.

Chaque patron est associé à un identifiant dans le format suivant : P_n, *P* pour *Patron*, *n* pour le numéro. Nous discutons des patrons présentés dans Federzoni et al. (2023). Dans le tableau 7.1 nous présentons les patrons les plus fréquents dans la ressource AnnoDis avec leur fréquence globale et leur fréquence en fonction du caractère humain (AD_{Hum+}) et non-humain (AD_{Hum-}) du référent. Les patrons que nous avons

Patron	AD _{Hum-}		AD _{Hum+}		AnnoDis	
	#	%	#	%	#	%
P1 : SN _{def} (>SN _{def} {1: 3}) >Pro{1:6}	24	8,4	42	14,2	66	11,4
P2 : SN _{def} >SN _{def} {1:6}	24	8,4	22	7,4	46	7,9
P3 : SN _{def} (>SN _{def} {1:2})>Pro{1:4} >SN _{def} {1:4} (>Pro{1:2})	32	11,2	19	6,4	51	8,8
P4 : SN _{def} (>SN _{def} {1:2}) >SN _{dem} {1:4} (>Pro SN _{def} {1:5})	30	10,5	3	1	33	5,7
P5 : Np (>Np{1:2}) >Pro{1:4} >Np{1:2} >Pro{1:4} (>Np{1})	0	0	10	3,4	10	1,8
P6 : Np (>Np) >Pro{1:4}	0	0	15	5,1	15	2,6
P7 : Np >Np{1:6}	3	1,1	2	0,7	5	0,9
Sous-total	113	39,7	113	38,2	226	38,9
Total CR	285	100	296	100	581	100

TABLEAU 7.1 – Les patrons les plus fréquents dans la ressource AnnoDis. Les pourcentages sont calculés par rapport au nombre total des chaînes à référent non humain (AD_{Hum-}), à référent humain (AD_{Hum+}) et au nombre total de chaînes dans la ressource

présentés couvrent 38,9% (226/581 CR) des chaînes annotées dans la ressource. Ils

ne couvrent donc pas la totalité des différentes stratégies mises en place pour l'introduction et le maintien du référent. Néanmoins, ils permettent d'effectuer une première distinction de différents types de CR.

Afin d'évaluer si les différences d'occurrence de chaque patron entre deux groupes (par exemple, $\text{AnnoDis}_{\{\text{Hum}+\}}$ et $\text{AnnoDis}_{\{\text{Hum}-\}}$) sont statistiquement significatives, nous avons appliqué un test binomial exact. Ce test permet de comparer les proportions d'apparition d'un patron donné dans deux échantillons de tailles différentes, tout en prenant en compte les effectifs relativement faibles pour certains cas. Concrètement, pour chaque patron, nous avons comparé le nombre d'occurrences dans chaque groupe à ce que l'on pourrait attendre si les deux groupes utilisaient ce patron de manière égale, en fonction de leur taille respective. Une valeur p inférieure à 0,05 indique que la différence observée est peu probablement due au hasard, et donc statistiquement significative. Lorsque la différence est significative et que l'un des deux groupes présente une fréquence plus élevée pour un patron donné, on peut en conclure que ce patron est préférentiellement utilisé dans ce groupe. À l'inverse, une différence non significative indique qu'aucune préférence claire ne peut être affirmée pour ce patron entre les deux groupes. Il est important de noter que l'analyse menée ici porte exclusivement sur les séquences identifiées comme correspondant aux patrons, et non sur l'ensemble des séquences présentes dans le corpus. Par conséquent, les préférences observées peuvent partiellement différer de celles issues de l'analyse des clusters, qui, elle, s'appuie sur la totalité des séquences annotées.

La [figure 7.8](#) illustre la répartition des patrons en fonction du caractère humain ou non humain du référent de la chaîne. Pour chaque patron, nous indiquons si la différence est significative (*, **, ***, selon la p -value obtenue) ou pas (« ns »)⁴. Nous pouvons observer que le patron le plus fréquent, P1, couvre deux cas de figure différents. Dans un premier cas, le référent est introduit par un SN_{def} et repris immédiatement par un pronom (30). Le maintien est ensuite assuré via une succession de pronoms. Ce patron rappelle la configuration d'un référent saillant ou du *topique continu* (Givón, 1983). Rappelons que Givon considère qu'un topique est continu si les pronoms sont utilisés sur une distance de 10 phrases consécutives. Pour confirmer cela, il est important de prendre en compte également les caractéristiques globales des chaînes, comme la longueur. La prise en compte de la persistance pourrait également permettre de confirmer notre interprétation.

- (30) Le colonel Picquart est lui aussi réhabilité officiellement et réintégré dans l'armée au grade de général de brigade. Il est même ministre de la Guerre de 1906 à 1909 dans le premier gouvernement Clemenceau. Il meurt en 1914 d'un accident de cheval. [AE_w; 9]

Dans le deuxième cas de figure, le référent est introduit par un SN_{def} et la reprise immédiate est faite par un autre SN_{def} . Dans ces cas, nous pouvons faire plusieurs hypothèses : ce type de composition peut se réaliser lorsque le premier maillon est dans un titre ; ou lorsque les deux premières mentions du référent sont éloignées ou en compétition avec d'autres référents (31). Dans (31), nous pouvons observer que le référent *Le LSD* est initialement secondaire et en compétition avec d'autres référents, ce qui peut

4. Le tableau détaillé avec les p -value est donné en annexe F.

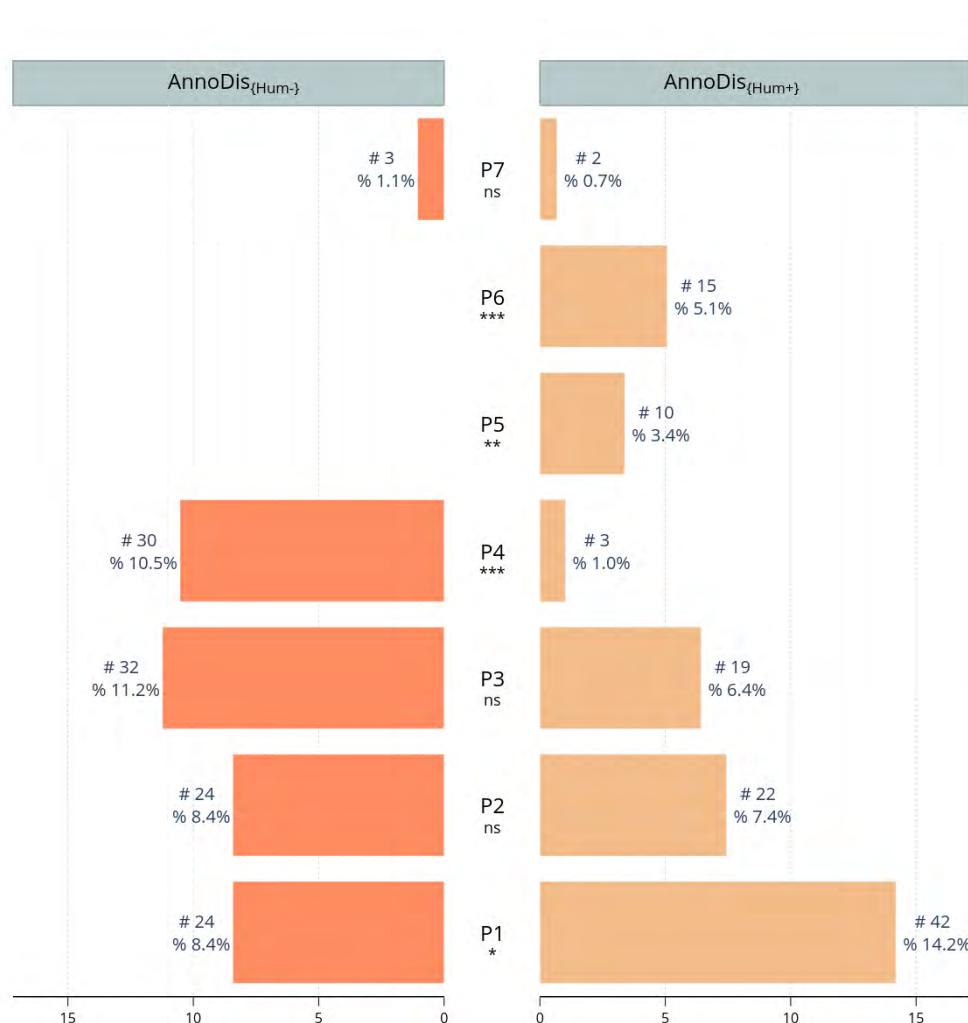


FIGURE 7.8 – Répartition des sept patrons AnnoDis en fonction du caractère *humain* ou *non humain* du référent de la chaîne.

expliquer l'utilisation d'expressions de faible accessibilité. Ensuite, nous observons qu'il passe dans une position saillante, avec un changement de fonction syntaxique, ce qui permet d'y référer par la suite pas des expressions de haute accessibilité, comme les pronoms. Un autre facteur qui pourrait entrer en jeu, c'est la nature des référents.

- (31) **Le LSD** (ou « acide ») fut découvert en 1943 par Albert Hofmann dans le laboratoire suisse Sandoz mais sera déclaré illégal aux États-Unis le 6 octobre 1966, ainsi que par l'ONU comme stupéfiant dans une convention de 1971. Jusqu'à cette interdiction sur le sol américain, la firme Sandoz met **Le LSD** à disposition des chercheurs sous la forme d'une préparation appelée delysid. **Le LSD** apparaît d'abord comme prometteur dans le traitement de certaines maladies psychiatriques. Puis, **il** est popularisé comme étant un traitement dit miraculeux par les médias à partir du milieu des années 1950. Dans les années 1960, **il** devient un ingrédient du courant hippie. [AEw;

En observant la [figure 7.8](#), nous pouvons remarquer que P1 est utilisé de préférence

avec des référents humains. À partir de l’observation de quelques exemples, nous avons pu identifier des cas où les SN défini fournit une information sur la profession du référent, comme dans (30); nous avons pu également observer que P1 couvre des CR dont le référent est un humain collectif, comme par exemple *les SS* ([AE_w; 

P2 formalise des chaînes homogènes, où l’introduction et le maintien du référent s’effectuent uniquement par le biais de SN_{def}. Il ne semble pas exister à ce stade une préférence claire en fonction du type de référent, ce patron ayant une fréquence similaire dans AnnoDis_{Hum{-}} et AnnoDis_{Hum{+}}. (32) illustre une CR à référent humain (collectif) correspondant au P2, et (25) une CR à référent non humain (*le vignoble champenois*).

- (32) Les années 1904 et 1905 sont consacrées aux différentes phases judiciaires devant la Cour de cassation. **La cour** emploie trois moyens (causes) à la révision : démonstration de la falsification du télégramme de Panizzardi. démonstration du changement de date d’une pièce du procès de 1894 (avril 1895 changé en avril 1894). démonstration du fait que Dreyfus n’avait pas fait disparaître les minutes d’attribution de l’artillerie lourde aux armées. Concernant l’écriture du bordereau, **la cour** est particulièrement sévère à l’encontre de Bertillon qui a « raisonné mal sur des documents faux ». Le rapport démontre que l’écriture est bien d’Esterházy, ce que ce dernier a d’ailleurs avoué entre-temps. Enfin, **la Cour** démontre par une analyse complète et subtile du bordereau l’inanité de cette construction purement intellectuelle, et une commission de quatre généraux dirigée par un spécialiste de l’artillerie, le général Sebert, affirme « qu’il est fortement improbable qu’un officier d’artillerie ait pu écrire cette missive ». [AE_w; 

P3 semble correspondre à la composition des topiques discontinus, proposée par Givón (1983). L’introduction du référent est effectuée par un SN_{def}, la deuxième mention du référent est un pronom, qui peut être suivi d’autres pronoms, à leur tour suivis d’une succession de SN_{def}. Ce patron suggère que certaines transitions peuvent se répéter dans les CR. Ici, par exemple, nous retrouvons des répétitions de la succession SN_{def} > Pro. Concernant la nature du référent, ce patron ne présente pas de différence significative entre humain et non humain. Au contraire, P4, qui présente une succession de SN_{def}, SN_{dem} et pronoms, est davantage utilisé pour les référents non humains, comme le montre l’exemple (33).

- (33) **La distinction** des registres énonciatifs histoire et discours illustre [...] **Cette dissociation** entre perspective énonciative et perspective textuelle, [...] Il est permis de penser **qu’elle** était une étape nécessaire [...] [AS_L; 

Dans cet exemple, nous pouvons observer que le SN_{dem} utilisé pour maintenir le référent correspond à une reformulation, et plus précisément à un cas d’encapsulation, comme nous l’avons déjà observé dans l’exemple (27). Suivant les travaux sur la position des maillons (Corblin, 1987), les démonstratifs peuvent être utilisés aussi pour recatégoriser le référent, il serait alors intéressant d’affiner l’étude des CR présentant ce type de composition pour étudier plus précisément les stratégies mises en place dans le corpus : par exemple, dans (26) la reprise par un SN démonstratif reprend exactement

la même tête lexicale. Une analyse croisée avec les types de reprise ou le coefficient de stabilité pourrait permettre de dégager les tendances et affiner la distinction entre les types de composition, en identifiant des chaînes plus ou moins stables.

P5 et P6 caractérisent des CR à référent humain qui commencent avec un Np suivi d'autres Np ou de pronoms, suivant des configurations différentes. Dans la première configuration, l'introduction du référent est faite par un Np et le maintien est assuré par une alternance de Np et de pronoms.

- (34) Les positions politiques prises par **Einstein** sont marquées par ses opinions pacifistes, qu' **il** relativise parfois, par exemple en déconseillant l'objection de conscience à un jeune Européen lui ayant écrit pendant les années 1930, « pour la sauvegarde de son pays et de la civilisation ». En 1913, il est cosignataire d'une pétition pour la paix que trois autres savants allemands acceptent de signer. **Einstein** éprouve une forte antipathie vis à vis des institutions militaires : « Si un homme peut éprouver quelque plaisir à défiler en rang aux sons d'une musique, je méprise cet homme... **Il** ne mérite pas un cerveau humain puisqu'une moelle épinière le satisfait. » Einstein est lié à de nombreuses causes pacifistes, car il se montre ouvert aux propositions multiples de soutien qu'il reçoit, et accepte souvent de s'engager pour les causes qu'il juge juste. [...]

[AE_w; 

Ce type de composition (P5), formalise des CR où le référent peut être en compétition avec d'autres, d'où le besoin d'utiliser le Np pour lever l'ambiguïté. Mais, comme nous pouvons l'observer dans (34), cela n'est pas toujours le cas. L'utilisation du Np pour référer à des entités accessibles et saillantes peut dépendre des choix faits au moment de l'écriture. [Schnedecker \(2005\)](#) a montré que ce type de composition, et encore plus P7 (Np++) est fréquent dans les portraits journalistiques. En suivant son exemple, il serait intéressant d'analyser plus finement la forme du Np (patronyme, nom complet, prénom). Dans la deuxième configuration, l'introduction du référent est faite par un Np et le maintien est assuré par une suite de pronoms (un ou plusieurs). Les pronoms peuvent être précédés par un Np, comme illustré dans (28). P6 caractérise des CR principales : l'importance du référent étant posée par le Np (cf. 2.1.1) et sa saillance par la possibilité d'utiliser des formes de haute accessibilité.

7.5 Bilan

L'analyse des CR AnnoDis nous a permis de commencer à mettre au jour des stratégies d'introduction et de maintien du référent. Nous avons proposé des patrons qui formalisent les différents types d'enchaînements des maillons. Nous avons fait le choix de présenter uniquement les patrons les plus fréquents pour la ressource AnnoDis et de discuter leur distribution en fonction de la nature du référent. Pour résumer les principaux résultats, nous avons pu identifier des patrons qui caractérisent des chaînes principales, homogènes, alternant des SN définis et des pronoms, comme dans les cas de P1 ou P6. Ce type d'enchaînement peut indiquer des chaînes centrales dans la structure discursive, où le référent est introduit de manière saillante et maintenu par plusieurs

expressions de haute accessibilité. Ces chaînes sont caractéristiques des topiques continus au sens de [Givón \(1983\)](#). Elles peuvent fréquemment être associées à des référents humains.

Nous avons ensuite identifié des CR plus homogènes, composées uniquement de SN définis (comme P2) ou des Np (comme P7). Ces chaînes peuvent concerner aussi bien des entités humaines que non humaines. La présence exclusive d'expressions de faible accessibilité pourrait indiquer qu'il s'agit de CR qui cohabitent avec d'autres, avec un risque de compétition de référents. Mais nous avons vu que tel n'est pas forcément le cas, notamment pour P7 (Np++), comme également suggéré par les travaux de [Schneidecker \(2005\)](#).

Ensuite, nous avons pu distinguer des CR dont la composition correspond à celle des topiques discontinus ([Givón, 1983](#)), comme P3 ou P4, qui impliquent des alternances entre expressions de faible et de haute accessibilité. Les chaînes correspondant à P4 sont souvent mobilisées pour des référents non humains, et nous avons vu que l'utilisation des SN_{dem} peut indiquer des cas d'encapsulation. Ces derniers ne sont pas souvent analysés étant donné la difficulté d'annoter l'antécédent. Une analyse plus fine de ce type de composition pourrait permettre de mieux étudier ces cas de figure pour affiner la description des stratégies mises en place.

Enfin, nous avons des CR principales (P5, P6 et P7), où le Np donne une indication de l'importance du référent. Ensuite, la reprise par les pronoms indique que le référent reste accessible. La reprise par des Np, nous l'avons déjà dit, n'indique pas forcément que le référent est en danger : il se peut qu'il reste accessible, malgré l'utilisation d'expressions de faible accessibilité. P5 et P6 sont liées à des référents humains.

Pour confirmer ces observations, il serait pertinent de croiser les patrons avec des critères de description portant sur la dimension des chaînes de référence, telles que la longueur, la persistance ou la distance entre les maillons. Ces facteurs permettraient d'évaluer plus finement le degré de saillance ou de stabilité des référents, et contribueraient à distinguer des chaînes à portée locale de celles plus globales – bien que cette dernière dimension soit partiellement contrainte par le modèle d'annotation d'AnnoDis (les chaînes ne peuvent pas dépasser une section).

Si cette première approche, centrée sur les patrons, ne permet pas d'appréhender l'intégralité des chaînes de référence, elle met néanmoins en évidence des configurations récurrentes et constitue un complément aux analyses par clustering. Cette première exploration permet donc de poser les bases d'une typologie des chaînes de référence, à partir des stratégies récurrentes d'introduction et de maintien des référents. Afin de consolider et de nuancer ces observations, le chapitre suivant propose d'appliquer cette même démarche à un corpus de plus grande taille, composé de textes plus variés du point de vue des types et des genres discursifs : la ressource DEMRC. L'objectif est d'examiner la robustesse des configurations identifiées et de compléter la description des différents types d'enchaînements des maillons.

Chapitre 8

Explorer les enchaînements : application de l'analyse des séquences à la ressource DEMRC

Contents

8.1	Projection des patrons AnnoDis sur la ressource DEMRC	166
8.2	Exploration des séquences	170
8.3	Description des clusters obtenus	173
8.3.1	Cluster C1-4c DEMRC : Np et Pro	176
8.3.2	Cluster C2-4c DEMRC : $SN_{\emptyset}$ et SN_{def}	178
8.3.3	Cluster C3-4c DEMRC : $SN_{def}++$	182
8.3.4	Cluster C4-4c DEMRC : des séquences hétérogènes	186
8.3.5	Bilan	197
8.4	Lien entre les clusters et les facteurs de variation	197
8.4.1	Impact de la nature du référent sur la composition des chaînes	197
8.4.1.1	Référents concrets et abstraits	198
8.4.1.2	Référents humains et non humains	199
8.4.2	Impact du type de texte sur la composition des chaînes	200
8.4.3	Impact du genre textuel sur la composition des chaînes	201
8.5	Bilan	203
8.6	Les patrons des chaînes DEMRC	205
8.7	Bilan	212

À partir de l'analyse des séquences des CR d'un premier corpus, AnnoDis, nous avons montré que les résultats sont prometteurs : les clusters obtenus correspondent à des types de composition différents, facilement identifiables grâce aux visualisations des *index plots*. Les classes sont donc interprétables d'un point de vue linguistique et les divisions obtenues sont cohérentes et traduisent différentes stratégies d'introduction et de maintien du référent, confirmant la validité de notre méthode. Dans ce

chapitre, nous présentons donc l'application de l'analyse des séquences aux chaînes de la ressource DEMRC.

Étant donné la variété des textes présents dans cette ressource et le plus grand nombre de CR annotées, nous nous attendons :

- à identifier des types de composition variés : en plus des types de composition mis au jour lors de l'analyse des CR de AnnoDis, nous supposons que la prise en compte d'un plus grand nombre de CR et de textes plus diversifiés permettra de mettre au jour de nouveaux types de composition, qui n'étaient pas très fréquents, ou pas du tout présents, dans AnnoDis ;
- à ce que la variété des types de composition soit liée au type textuel et aux genres textuels présents dans la ressource ;
- à ce que les types de composition varient selon la nature du référent (humain, non humain, concret ou abstrait).

Rappelons que les chaînes à partir desquelles nous avons proposé les patrons AnnoDis sont des chaînes topicales, annotées en suivant un modèle d'annotation différent par rapport à DEMRC (cf. 3.3). Il se peut donc que les types de composition soient spécifiques à l'identification de chaînes topicales. Ainsi, étant donné le modèle d'annotation de DEMRC, où toutes les expressions référentielles sont annotées et peuvent constituer une CR, nous nous attendons à identifier des types de composition permettant de distinguer des chaînes dont le référent est topical des chaînes dont le référent est secondaire. Nous pourrions ainsi proposer des patrons propres aux CR secondaires ou d'arrière-plan. Comme pour AnnoDis, nous proposons des patrons qui formalisent les types de compositions identifiés.

L'application de l'analyse des séquences à la ressource DEMRC a été effectuée en suivant deux types d'analyse en parallèle, que nous allons présenter dans l'ordre :

- la projection des patrons AnnoDis sur les CR de DEMRC ;
- l'analyse des séquences et la proposition des nouveaux patrons.

8.1 Projection des patrons AnnoDis sur la ressource DEMRC

Avant d'appliquer l'analyse des séquences aux CR de DEMRC, nous avons projeté les patrons AnnoDis sur cette nouvelle ressource. Cette projection avait pour objectif de tester la productivité des patrons sur une autre ressource. Si les patrons s'avèrent assez génériques, il sera alors possible d'identifier automatiquement les chaînes correspondant à ces types de composition. Dans ce qui suit, nous allons présenter les résultats obtenus.

La projection des patrons nous a permis de confirmer nos attentes concernant la présence des mêmes types de composition dans la ressource DEMRC. En effet, nous avons identifié 844 CR, soit 19,9% des 4235 CR présentes dans la ressource. Nous remarquons que la couverture est plus faible par rapport à AnnoDis (226 CR identifiées, soit 38,9% des 581 CR). D'une part, cela peut s'expliquer par une différence dans la nature des chaînes annotées. D'autre part, la variété des textes DEMRC et le plus grand

nombre de CR annotées peuvent justifier cette faible couverture. Le [tableau 8.1](#) illustre la couverture de chaque patron pour AnnoDis et DEMRC.

Patron	AnnoDis		DEMRC	
	#	%	#	%
P1 : $SN_{def} (>SN_{def}\{1:3\}) >Pro\{1:6\}$	66	11,4	183	4,3
P2 : $SN_{def} >SN_{def}\{1:6\}$	46	7,9	382	9
P3 : $SN_{def} (>SN_{def}\{1:2\}) >Pro\{1:4\} >SN_{def}\{1:4\} (>Pro\{1:2\})$	51	8,8	156	3,9
P4 : $SN_{def} (>SN_{def}\{1:2\}) >SN_{dem}\{1:4\} (>Pro\{1:5\})$	33	5,7	59	1,4
P5 : $Np (>Np\{1:2\}) >Pro\{1:4\} >Np\{1:2\} >Pro\{1:4\} (>Np\{1\})$	10	1,7	1	0,02
P6 : $Np (>Np) >Pro\{1:4\}$	15	2,6	13	0,3
P7 : $Np >Np\{1:6\}$	5	0,7	50	1,2
Sous-total	226	38,9	844	19,9
Total CR	581		4235	

TABLEAU 8.1 – Couverture des patrons : comparaison entre AnnoDis et DEMRC. Le symbole # renseigne les effectifs identifiés et le symbole % indique le pourcentage par rapport au nombre total des CR présentes dans chaque ressource, à savoir 581 pour AnnoDis et 4235 pour DEMRC.

Afin d'évaluer si les différences d'occurrence de chaque patron entre deux groupes (par exemple, AnnoDis et DEMRC) sont statistiquement significatives, nous avons appliqué un test binomial exact, dont les résultats pour chaque combinaison sont donnés en annexe F. Nous avons d'abord comparé les distributions entre AnnoDis et $DEMRC_{\{Nar-\}}$ dont les résultats sont illustrés dans la [figure 8.1](#). Pour chaque patron, nous indiquons si la différence est significative (*, **, ***, selon la p-value obtenue) ou pas (« ns »). Ensuite, nous avons effectué une comparaison entre $DEMRC_{\{Nar+\}}$ et $DEMRC_{\{Nar-\}}$ (cf. [annexe F](#)).

Nous pouvons observer que le pourcentage des CR identifiées via les patrons est plus élevé dans AnnoDis pour l'ensemble des patrons, à l'exception du patron P2 (suite de SN_{def}) et P7 (suite de Np). Ces deux patrons correspondent à des chaînes composées uniquement d'expressions de faible accessibilité, ce qui pourrait signifier que le référent n'est pas saillant dans le discours, qu'il est en compétition avec d'autres référents, ou que son degré d'accessibilité est menacé par une forte distance entre les maillons. Ces caractéristiques sont en décalage avec celles des chaînes topicales, pour lesquelles nous nous attendons à ce que le référent soit saillant et rarement en compétition avec d'autres référents. Ces raisons pourraient expliquer pourquoi les deux patrons couvrent davantage de CR dans DEMRC.

Considérons maintenant la distribution selon les types de textes. Les patrons AnnoDis ayant été proposés à partir de l'analyse des chaînes topicales dans les textes non narratifs (cf. [tableau 7.1](#)), nous nous attendons à ce que ces patrons identifient davantage de CR relevant des textes non narratifs de la ressource DEMRC. Nous pourrions également faire l'hypothèse que les patrons formalisant des CR constituées principalement de noms propres et de pronoms (P5 et P6) soient davantage présents dans les

textes narratifs pour introduire et maintenir les différents personnages. L'analyse des données montre que cette tendance n'est pas statistiquement significative.

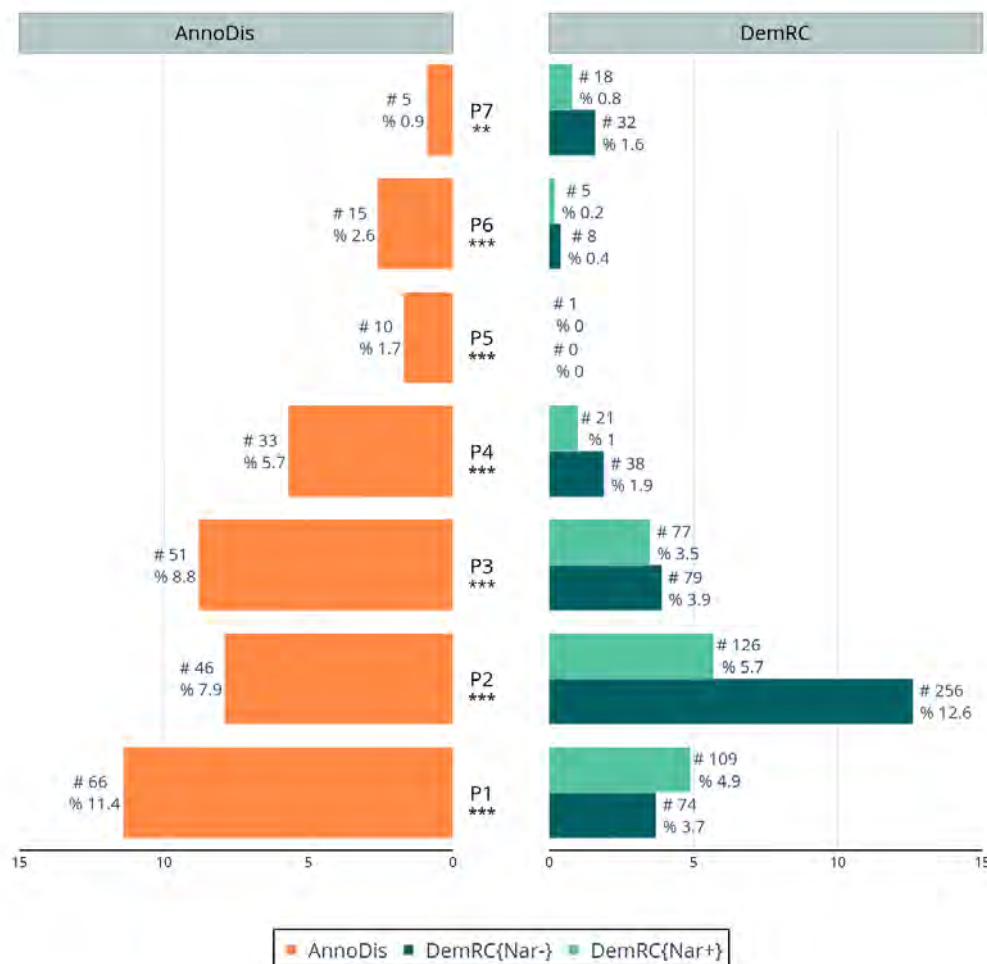


FIGURE 8.1 – Couverture des patrons : comparaison entre AnnoDis, $DEMRC_{\{Nar-\}}$ et $DEMRC_{\{Nar+\}}$

Afin d'explorer ces hypothèses, nous avons comparé la couverture des patrons dans les textes non narratifs ($DEMRC_{\{Nar-\}}$) et dans les textes narratifs ($DEMRC_{\{Nar+\}}$). La figure 8.1 illustre la couverture de chaque patron dans la ressource AnnoDis (à gauche) et dans la ressource DEMRC (à droite) selon le type de texte. Nous pouvons observer une tendance qui confirme une couverture plus élevée dans les textes non narratifs (le test binomial confirme une différence significative pour les P2, P4 et P7), à l'exception du P1 (suite de SN_{def} et Pro). La couverture plus importante de ce patron dans les textes narratifs est statistiquement significative et pourrait s'expliquer par la tendance à utiliser des descriptions de rôle (e.g. *le garçon*, *la serveuse*, etc.) pour introduire des personnages secondaires (Schneidecker, 1997). Une analyse qualitative plus fine serait nécessaire pour confirmer cette explication.

Un dernier aspect que nous avons analysé lors de la projection des patrons porte sur la nature du référent. Afin de pouvoir comparer les résultats entre les deux ressources, nous avons dû simplifier l'annotation de la nature du référent dans DEMRC. Nous

avons réduit les catégories à une annotation binaire comme celle de AnnoDis : ainsi, nous distinguons entre humain et non humain. Cela nous fait perdre la distinction entre les référents humains collectifs et individuels, mais nous analyserons ces catégories plus finement lorsque nous présentons les patrons DEMRC (cf. 8.6).

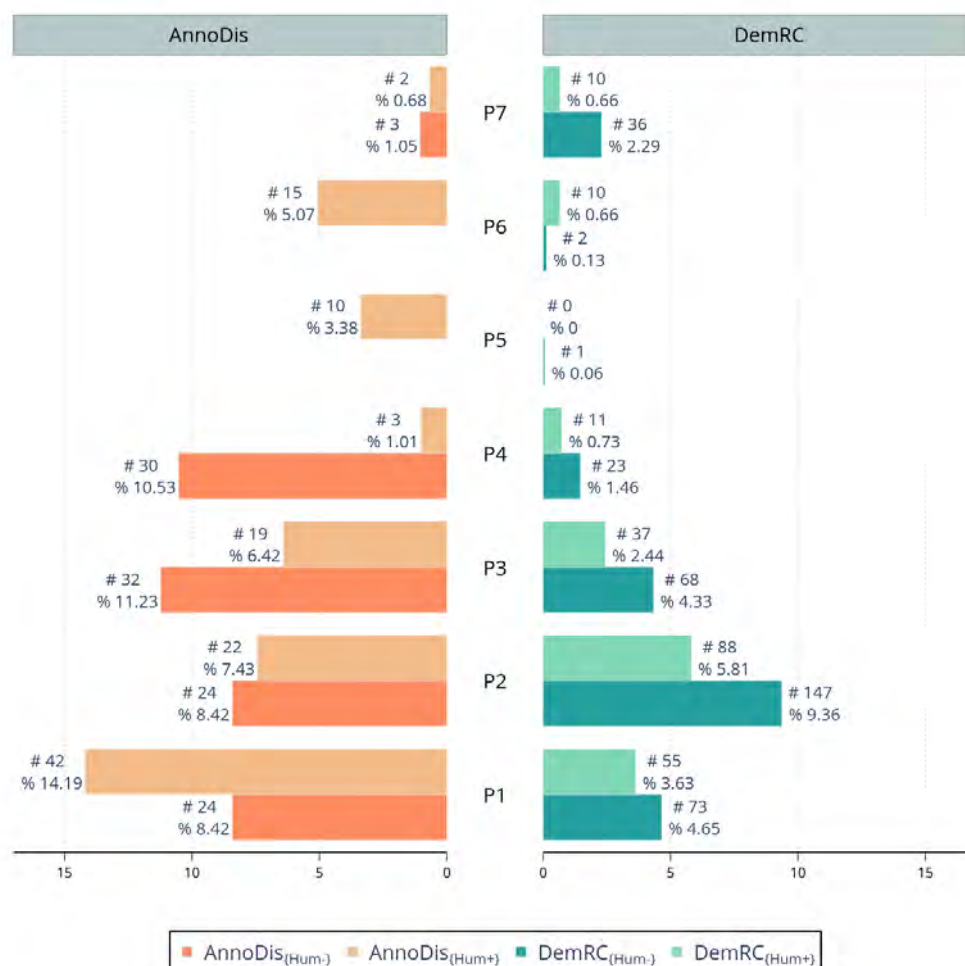


FIGURE 8.2 – Couverture des patrons : comparaison entre les chaînes à référent humain et les chaînes à référent non humain dans AnnoDis et dans DEMRC

Globalement, dans DEMRC, nous observons des tendances en accord avec les résultats présentés pour AnnoDis, sauf pour le P1 (figure 8.2). Ce patron semble être privilégié pour introduire et maintenir des référents non humains, contrairement à ce que nous avons vu pour AnnoDis. Cependant, le test binomial montre que la différence n'est pas significative (p -value = 0,18 (cf. annexe F)), c'est pourquoi nous parlons de *tendances*. Le patron P2 caractérise aussi bien les CR à référent humain que celles à référent non humain, avec une différence de couverture plus marquée dans DEMRC que dans AnnoDis. Pour les patrons P3 et P4, ils ont tendance, comme pour AnnoDis, à caractériser des chaînes à référent non humain. Pour P3, cette tendance est confirmée statistiquement ($p = 0,004$). P5 permet d'identifier une seule CR à référent non humain. Dans cette chaîne (ex. (35)), le Np correspond à un toponyme. Rappelons que

dans AnnoDis, les toponymes sont toujours considérés comme des humains puisque l’annotation a été effectuée semi-automatiquement, alors que dans DEMRC l’annotation manuelle a permis de distinguer les cas où le toponyme réfère à des humains (collectifs) ou pas, comme dans l’exemple (35) :

- (35) Il connaissait leur rêve, et un beau jour vint leur dire qu’on lui avait parlé d’un domaine, à Chavignolles, entre **Caen** et Falaise [...] Les instruments de jardin, les couchettes, les matelas, les tables, les chaises, un caléfacteur, la baignoire et trois fûts de Bourgogne iraient par la Seine, jusqu’au Havre, et de là seraient expédiés sur **Caen**, où Bouvard qui les attendrait les ferait parvenir à Chavignolles [...] Le soir, toutes les places pour **Caen** étaient retenues ; ne sachant que faire, il alla au théâtre des Arts, et il souriait à ses voisins, disant qu’il était retiré du négoce et nouvellement acquéreur d’un domaine aux alentours. Quand il débarqua le vendredi à **Caen**, ses ballots n’y étaient pas. [R_{BeP} ; 

Pour résumer, les patrons AnnoDis couvrent environ 20% des CR dans DEMRC, ce qui confirme leur capacité à identifier certains types d’enchâînements sur un corpus autre que le corpus source. Globalement, nous avons observé une couverture plus importante dans les textes non narratifs, ce à quoi nous nous attendions étant donné que les patrons ont été construits à partir de textes de ce type. Enfin, nous avons comparé la couverture des patrons selon la nature du référent, et observé les mêmes tendances que celles présentées pour AnnoDis. Toutefois, nous avons mis en évidence que P1 suit un comportement différent dans DEMRC : sa couverture est plus importante dans les textes narratifs et il semble être davantage mobilisé pour introduire et maintenir des référents non humains.

Cette première étape de l’analyse, nous a permis de commencer à identifier des tendances en fonction du type de texte et de la nature du référent, et ouvre la voie à une caractérisation plus fine des compositions. Pour affiner la description et tester nos hypothèses quant à l’identification d’autres types de composition, nous poursuivons avec l’application de l’analyse des séquences aux chaînes DEMRC.

8.2 Exploration des séquences

La première étape de l’analyse des séquences consiste en l’exploration des séquences (cf. 5.2.2). Cette première étape vise à décrire les caractéristiques globales de l’ensemble des séquences en entrée, comme la distribution des CR selon leur longueur et la distribution de la nature des maillons. Cette étape est importante pour pouvoir connaître les données que nous allons manipuler. Dans cette section, nous décrivons donc les caractéristiques de l’ensemble des séquences DEMRC.

Rappelons que les séquences que nous analysons sont constituées d’un minimum de trois états et un maximum de sept (cf. 6.2.1). Observons à présent la distribution — après coupure des CR — des différentes natures de maillons selon leur position dans la chaîne (figure 8.3).

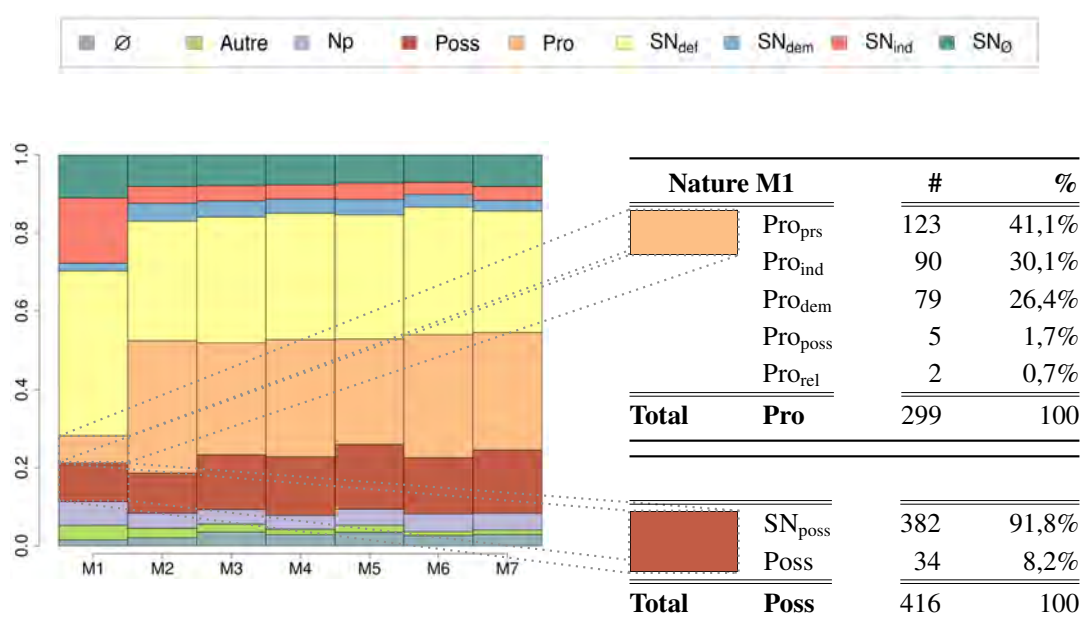


FIGURE 8.3 – Distribution de la nature des maillons dans DEMRC après la coupure des chaînes à sept maillons

TABLEAU 8.2 – Distribution de la nature détaillée du premier maillon dans DEMRC

Nous pouvons constater que, comme attendu et comme précédemment observé dans AnnoDis, la nature du premier maillon correspond principalement à des expressions de faible accessibilité, comme les SN_{def}, les SN_{ind} et les noms propres. Nous pouvons cependant noter également la présence d’expressions de haute accessibilité dès le premier maillon, ce qui n’est pas attendu. En analysant plus finement la nature pour ces cas inattendus, nous observons que la plupart des pronoms en première position sont des pronoms personnels¹ (36), suivis de pronoms indéfinis (37) et démonstratifs (38). La présence des pronoms représente 7% des CR et peut s’expliquer par le modèle d’annotation (cf. chapitre 3). Nous avons identifié plusieurs cas :

- des cas de fusion (Schneidecker, 2021) : les référents introduits individuellement dans un premier temps constituent, dans un deuxième temps, un ensemble. Ainsi, dans (36) la chaîne commence par le pronom personnel pluriel *ils*. Néanmoins, nous pouvons remarquer que les référents sont introduits individuellement au préalable par des Np (*Claude* et *Alexandre*). Ceux-ci étant disposés à gauche et à droite du syntagme verbal — cas de figure connu sous le nom de *split antecedent* — le *groupe* ne fait pas l’objet d’une annotation². Bien que le pronom occupe la position de premier maillon, les référents auxquels il renvoie sont introduits en amont par des expressions de faible accessibilité. Cette introduction préalable permet de bien ancrer le référent, conformément à ce qui est décrit dans la littérature (cf. 1.1.1). Parmi les pronoms qui occupent la position de premier maillon,

1. Nous avons repéré plusieurs cas de figure lorsque le premier maillon est un pronom personnel : nous reviendrons plus en détail sur les différents cas dans la section 8.3.4, où nous analysons les séquences en termes de stratégies pour l’introduction et le maintien du référent.

2. Les mentions des antécédents constituent des maillons séparés, donnant lieu à deux CR : une pour chaque référent pris isolément.

98 (soit 2,3% des CR) sont des pronoms pluriels, dont la plupart (63%) sont des pronoms personnels.

- (36) Claude l'appelait Alexandre, lui tapait sur les bras, lui demandait quand ils iraient à Charentonneau. Et ils parlaient d'une grande partie qu'ils avaient faite ensemble, en canot, sur la Marne. Le soir, ils avaient mangé un lapin. [...] Mais les autres n'en prirent pas, et Florent but son verre de punch ; il le sentit qui tombait dans son estomac vide, comme un filet de plomb fondu. [R_{VdP} ; ⚡]

- des cas où un référent en emploi générique est repris en emploi spécifique et vice-versa : dans ces cas, les référents ne coréférent pas et les expressions référentielles ne font pas partie de la même CR. (37) illustre ce cas de figure. La CR démarre par un pronom indéfini (*une* → *une couche*). L'utilisation du pronom est possible puisque le référent a été préalablement introduit (*de bonnes couches*), pour autant les deux mentions ne sont pas coréférentielles puisque la première désigne un référent générique et la deuxième un référent spécifique. Une analyse plus fine de ces cas pourrait confirmer cette observation. Dans le corpus, les pronoms indéfinis en premier maillon sont au nombre de 90, ce qui correspond à 2% des CR DEMRC (ce qui représente 30% des CR commençant par un pronom ; cf. [tableau 8.2](#)).

- (37) Bouvard s'occuperait de la ferme. Ils en délibérèrent, et cet arrangement fut décidé. Le premier point était d'avoir de bonnes couches. Pécuchet en fit construire une en briques. [...] Mais la couche fourmilla de larves ; malgré les réchauds de feuilles mortes, sous les châssis peints et sous les cloches barbouillées, il ne poussa que des végétations rachitiques. [...] Il sema les graines de plusieurs variétés dans des assiettes remplies de terreau, qu'il enfouit dans sa couche. [R_{BeP} ; ⚡]

- des cas où le pronom en première position est un démonstratif (79 occurrences, soit 1,9% des CR ; cf. [tableau 8.2](#)) dans une construction présentative, comme dans (38).

- (38) En effet, c'était le cadavre du passé qui tombait lentement en poudre dans ces salles où le présent ne mettait pas le pied, c'étaient les années endormies qui se berçaient comme dans des hamacs aux toiles grises des encoignures. [R_{CF} ; ⚡]

Concernant les possessifs en première position, rappelons que par souci de comparabilité avec le corpus AnnoDis, nous avons utilisé une catégorie unique *Poss* pour désigner aussi bien les syntagmes nominaux possessifs que les déterminants possessifs. Néanmoins, la caractérisation plus fine que nous avons effectuée, nous permet de proposer des interprétations plus précises lorsque cela est pertinent. Notamment ici, afin de comprendre la présence des possessifs en première position dans la chaîne : il

s'agit principalement de syntagmes nominaux possessifs (382 soit 91,8% des maillons possessifs en première position ([tableau 8.2](#))), ce qui leur confère un statut d'accessibilité moyenne. Cela peut indiquer l'introduction de référents secondaires, introduits par le biais de possessifs, car ils ont une relation de dépendance avec les référents principaux des textes ([Schneidecker, 2021](#)), comme dans les exemples (39a) et (39b).

- (39) a. **Mes frères** et mes sœurs commençaient d'en vouloir à la guerre, ils la trouvaient longue. [...] **Mes frères** plaisaient ma petite sœur. Les roues de sa bicyclette ont à peine quarante centimètres de diamètre : « **On** te laissera seule sur la route. [...] **Ils** proposent de réparer la mienne. **Ils** se lèvent dès l'aube pour connaître les nouvelles. [...] **Ils** eussent brûlé Paris pour partir plus vite. Ce qui terrifiait l'Europe était devenu **leur** unique espoir. [...] [R_{DaC} ; ⚡]
- b. Le jeune homme occupait **son logement** depuis deux mois environ et était dépressif. [...] L'occupant de **l'appartement du rez-de-chaussée**, à l'origine du sinistre, a été brûlé au visage et aux bras. [...] Réveillés en sursaut, les huit locataires, voisins de **l'appartement du rez-de-chaussée** où l'explosion s'est produite, ont ouvert les yeux sur un spectacle de désolation : leur appartement n'était plus qu'un amas de ce qui restait des murs et de leur mobilier. [AP_{ER} ; ⚡]

Revenons à la distribution de la nature des maillons suivants : en observant la distribution des maillons qui suivent le premier ([figure 8.3](#)), nous pouvons remarquer qu'ils correspondent majoritairement à des expressions de haute accessibilité, notamment des pronoms. Néanmoins, nous pouvons constater une fréquence assez élevée de SN_{def}. Nous avons observé cette même tendance pour la distribution de la nature des maillons dans AnnoDis (cf. [chapitre 6](#)). Cela pourrait être dû à différents facteurs : référents en compétition, nature du référent, genre textuel. Une analyse plus approfondie des stratégies à l'aide des clusters de séquences nous permettra de mieux décrire ces tendances.

8.3 Description des clusters obtenus

Dans cette section, nous présentons l'application de l'analyse des séquences à la ressource DEMRC. Nous discutons les résultats obtenus en deux temps : une classification à gros grain en quatre clusters, afin de décrire les tendances générales ; puis une classification à petit grain en onze clusters, afin d'analyser les distinctions plus fines pour chaque type de CR identifié avec la classification en quatre classes. Nous rappelons que la prise en compte de la classification en onze clusters permet de rendre compte de toutes les distinctions effectuées aux niveaux intermédiaires, sans devoir passer en revue chaque niveau. Rappelons également que le nombre de clusters analysés diffère entre les deux ressources. En raison du grand nombre de CR traitées dans DEMRC et de leur variété, la division en trois et cinq clusters retenue pour AnnoDis ne permettait pas d'identifier clairement les types de compositions. Les observations ayant conduit à la sélection de quatre et onze clusters sont présentées en détail au [chapitre 6](#) (cf. 6.2.3).

La classification des 4 235 séquences obtenue avec le regroupement en quatre clusters (4c) permet d'identifier les tendances suivantes (figure 8.4) :

- des séquences composées principalement de Np et pronoms (C1-4c, figure 8.4_a);
- des séquences composées principalement de $SN_{\emptyset}$ et de SN_{def} (C2-4c, figure 8.4_b);
- des séquences composées principalement de SN_{def} (C3-4c, figure 8.4_c);
- des séquences composées principalement de SN_{def} , SN_{ind} et pronoms (C4-4c, figure 8.4_d).

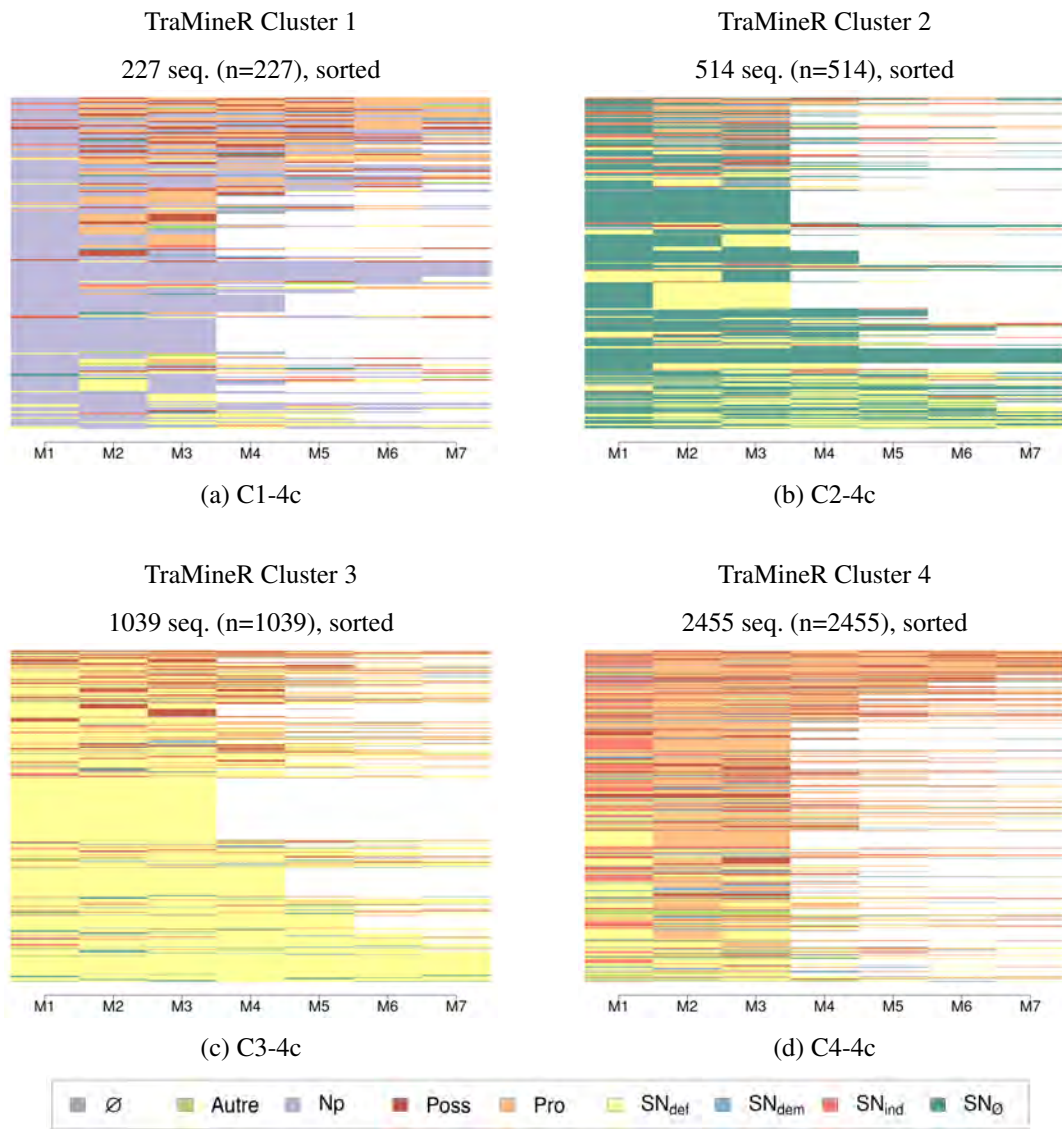


FIGURE 8.4 – Visualisation par des *index plots* des résultats de la classification des séquences DEMRC en quatre clusters

Nous remarquons que les quatre classes sont très déséquilibrées en nombre de séquences, les deux dernières regroupant à elles deux 82,5% des séquences.

L'analyse de l'entropie transversale (figure 8.5) — qui, nous le rappelons, permet de décrire comment l'homogénéité de la classe évolue selon les états (cf. chapitre 5) — montre que le C4-4c est non seulement le plus fourni, mais aussi le plus hétérogène,

avec une entropie entre 0,6 et 0,8. Il présente une très forte variation pour les premiers maillons, qui diminue légèrement dans les positions suivantes. Étant donné le grand nombre de CR et la forte hétérogénéité, il est difficile d'identifier des tendances pour ce cluster sans passer par une analyse à petit grain (classification en onze clusters). Globalement, toutes les classes, à l'exception du C4-4c, suivent la même tendance : la nature du premier maillon est stable, alors que les suivants sont plus diversifiés.

Dans les sections suivantes, nous discutons plus en détails la composition des clusters obtenus.

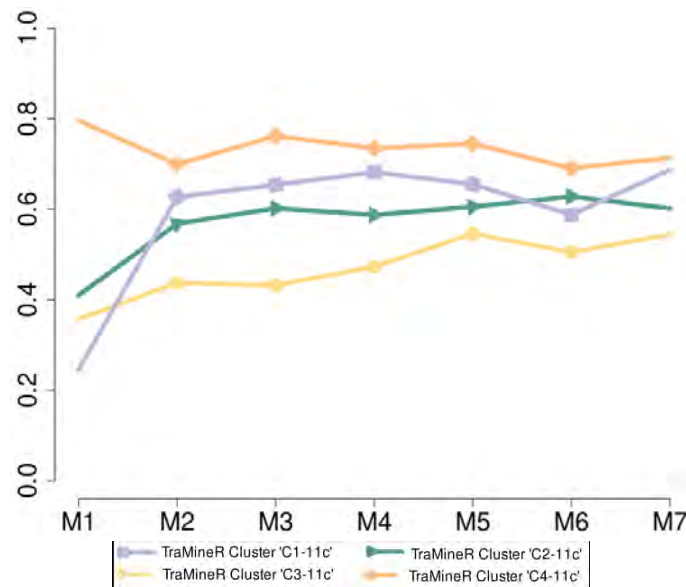


FIGURE 8.5 – Entropie transversale des séquences selon les quatre clusters obtenus

8.3.1 Cluster C1-4c DEMRC : Np et Pro

Le C1-4c (figure 8.4_a) regroupe 227 séquences, soit 5,36% des CR analysées. Ce cluster regroupe les CR dont le premier maillon est majoritairement un nom propre (88,1%). Dans la suite, on retrouve d'autres noms propres ou bien une alternance de pronoms, noms propres et possessifs. De façon marginale, certaines CR peuvent commencer par un maillon d'une autre nature : SN_{def} (4,8%), SN_{ind} (2,2%), et encore plus marginalement, possessif, pronom et $SN_{\emptyset}$. Bien que ce cluster ne fasse pas l'objet d'une division ultérieure, nous observons deux tendances : les cas où les CR sont composées exclusivement de Np et les CR où le Np alterne avec des pronoms, ou dans une moindre mesure, avec des SN_{def} .

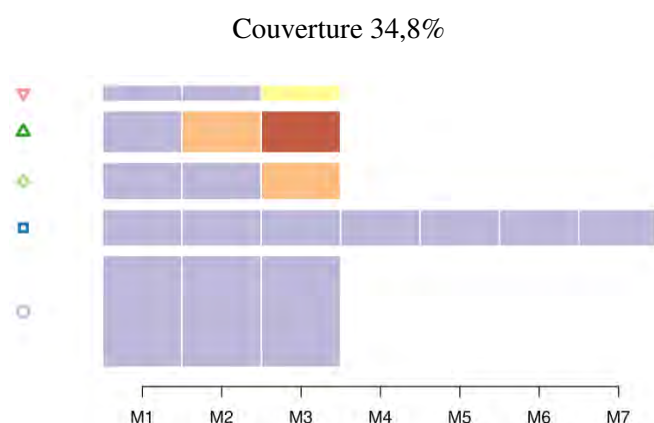


FIGURE 8.6 – Séquences représentatives du C1-4c

Ces deux tendances sont confirmées par l'analyse des séquences représentatives du cluster (cf. chapitre 5, 5.2.5). La figure 8.6 illustre les cinq séquences représentatives de ce cluster³. La largeur des barres traduit la fréquence des différents types de composition. Par exemple, les CR composées uniquement de trois Np sont plus fréquentes que les CR composées de deux Np et un SN_{def} . Les cinq types de séquences proposées, dont nous fournissons des exemples dans (40), couvrent 34,8% du cluster. Les exemples sont accompagnés des symboles présents dans la figure 8.6 pour signaler la correspondance avec leur type de composition.

- (40) a. ▼ Lucette Monti a reçu le consentement mutuel de ce jeune couple, lui sans emploi domicilié à Bar-le-Duc et elle coiffeuse demeurant à Paris, avant qu'il ne s'en aille faire la fête. [...] [article] Fête des rois
En partenariat avec le CCAS de Bar-le-Duc, la commune invite les habitants de plus de 60 ans à l'animation proposée au hall des Brasseries dimanche 5 janvier. [AP_{ER} ; ⚡]
- b. ▲ Tu devrais peut-être écrire à Constance, elle est chez son père, à Hermenches. [B_{RdC} ; ⚡]

3. Rappelons que nous avons utilisé la fréquence comme paramètre pour le calcul des séquences représentatives car avec celui-ci, nous obtenons les meilleurs résultats tant en termes de couverture que de qualité (cf. chapitre 5). Nous avons également choisi de visualiser cinq séquences, mais nous aurions pu en visualiser moins ou plus.

- c.  Je pensais à **Daphnis**. Ici c'est Chloé qui avait reçu quelques leçons, et **Daphnis** n'osait lui demander de les **lui** apprendre. [$R_{\{DaC\}}$; 
- d.  Elle habitait avec ses parents la belle propriété de St-Jean, qui domine la jonction de l'Arve avec le Rhône, en aval de **Genève** [1 phrase] On sait qu'au début de son séjour dans notre pays, Voltaire habita en été Les Délices près **Genève**, en hiver tantôt Monrion, près Lausanne, tantôt cette ville-même, au faubourg du Petit-Chêne. À Lausanne comme à **Genève**, Voltaire avait trouvé une société très aimable et lettrée, toute disposée à jouer ses tragédies. [14 phrases] Quelquefois Samuel revenait pour lui donner la réplique, et alors la jeune femme jouissait de tout son cœur des répités que lui laissaient soit une santé déjà très compromise, soit l'arrivée successive des enfants que l'on mettait en nourrice dans les environs de **Genève**. [70 phrases] Rosalie et ses frères et sœur habitaient tantôt **Genève** avec leur père, tantôt St-Jean avec M. et Mme Pictet.. [25 phrases] Rosalie eut peu d'amies à **Genève**. [1 phrase] Cela tient, nous dit-elle elle-même quelque part, à ce que dans ce temps-là les portes de **Genève** se fermaient en hiver à quatre ou cinq heures [...] [$B_{\{RdC\}}$; 
- e.  Le Centre d'art et de plaisanterie de **Montbéliard** est venu donné un coup de main pour l'animation avec sa « Brigade d'intervention théâtrale », composée d'une quinzaine de personnes. [319 phrases] Originalité de l'opération : partout, un représentant de la préfecture était sur place, les sous-préfets dans les arrondissements de Pontarlier et de **Montbéliard**, le chef de cabinet, M. Bertrand Gaume et le procureur de la république M. Jean-Pierre Nahon à Besançon. [68 phrases] « Ce restaurant doit garder un flux rapide de clientèle, mais nous ne voulons pas mépriser les personnes qui viennent nous voir » explique Philippe Fayard, le « Monsieur Mac Do » de l'Aire urbaine, qui dispose déjà des restaurants de **Montbéliard**, d'Audincourt et aujourd'hui d'Andelnans. Ce fast food nouveau aura en tout coûté la somme de quatre millions de francs. [$AP_{\{ER\}}$; 

Interprétation du cluster En termes de stratégies d'introduction et de maintien du référent, nous constatons une utilisation majoritaire d'expressions de faible accessibilité. Les exemples des séquences représentatives distinguent néanmoins :

- les cas où l'introduction et le maintien sont assurés par le biais d'expressions de faible accessibilité, tels les Np et les SN_{def} (illustrés par (40a), (40d) et (40e)); Plusieurs raisons peuvent expliquer ce recours à des expressions de faible accessibilité : distance entre les maillons, compétition de référents, nature particulière du référent (cf. chapitre 1). Ainsi, dans les exemples (40d) et (40e), les expressions référentielles sont distantes entre elles. Dans (40a) une expression de faible accessibilité, tel un SN_{def} (*la commune*), est utilisée pour reprendre le référent, bien que les deux mentions soient très proches. Cela suggère que dans (40a), ce n'est pas la distance entre les maillons qui a influencé le choix de la nature des expressions référentielles. Ce choix pourrait être influencé par la nature des référents. Nous constatons que, dans les exemples proposés, les référents des CR composées d'expressions de faible accessibilité sont des toponymes, ce qui en plus de la distance, pourrait expliquer l'enchaînement des Np.

- les cas où le référent est introduit par des Np et maintenu par des expressions de haute accessibilité, tels les pronoms et les possessifs (illustrés par (40b) et (40c)). Dans ces cas, la possibilité de reprendre le référent via des formes de haute accessibilité pourrait être liée à une distance moindre et à l'absence d'ambiguïté référentielle. Ce cas de figure pourrait correspondre au topique continu (Givón, 1983). Concernant la distance entre les maillons, les exemples (40b) et (40c), nous pouvons constater que les maillons sont très rapprochés. Cependant, les deux CR sont courtes (3 maillons) : le référent n'est actif que localement. Dans (40b) et (40c), nous pouvons également observer une introduction du référent par un Np en position d'objet indirect, qui devient ensuite sujet. Dans les deux cas que nous venons de présenter, les expressions référentielles concernent des personnes.

L'analyse des facteurs de variation (cf. 8.4) nous permettra de mesurer le poids de certains paramètres, notamment concernant la nature des référents. En revanche, pour confirmer nos interprétations quant à la distance entre les maillons et distinguer les chaînes éphémères et celles à topique continu, il faudrait mettre les séquences en correspondance avec les caractéristiques globales, telles la longueur, la distance entre les maillons et la persistance de ceux-ci. Cette mise en correspondance ne sera pas abordée dans le cadre de cette thèse. Nous en discutons dans les conclusions.

8.3.2 Cluster C2-4c DEMRC : $SN_{\emptyset}$ et SN_{def}

Le cluster C2-4c (figure 8.4_b) regroupe 514 séquences (soit 12,13% des CR analysées) composées principalement de $SN_{\emptyset}$ et de SN_{def} . Plus précisément, concernant la caractérisation linéaire, la majorité des CR commence par un $SN_{\emptyset}$ (378 CR, soit 73,5% des chaînes de ce cluster). Parmi celles-ci, nous pouvons observer deux tendances principales : des CR constituées exclusivement de $SN_{\emptyset}$ (106 CR, soit 20,6% de ce cluster) et des CR dont les maillons suivants se caractérisent par la succession de SN_{def} et l'alternance de $SN_{\emptyset}$, SN_{def} et plus rarement de pronoms. Environ un quart des CR démarrent par un SN_{def} (16,3%). Pour les maillons suivants, nous observons des tendances similaires à celles déjà évoquées, à savoir une alternance de $SN_{\emptyset}$ et de SN_{def} , la présence de pronoms dans une moindre mesure, et très rarement de SN_{dem} ou de possessifs.

Les tendances en détail Les tendances que nous venons de mentionner sont davantage visibles lorsque nous observons les distinctions plus fines (figure 8.7). En effet, lors de la classification en onze clusters, le cluster C2-4c est divisé en deux : cette subdivision permet de distinguer des CR dont la composition est homogène (C2a-11c) de celles présentant une composition plus hétérogène (C2b-11c). Cette tendance est confirmée par l'entropie transversale (cf. figure 8.7 b et d) : le premier cluster présente un score entre 0,2 et 0,4, alors que le deuxième présente une entropie plus élevée, entre 0,4 et 0,8.

Le premier cluster, C2a-11c (figure 8.7_a), regroupe 163 CR (soit 31,7% des CR du cluster initial). La plus grande partie des CR de ce cluster commence par un $SN_{\emptyset}$

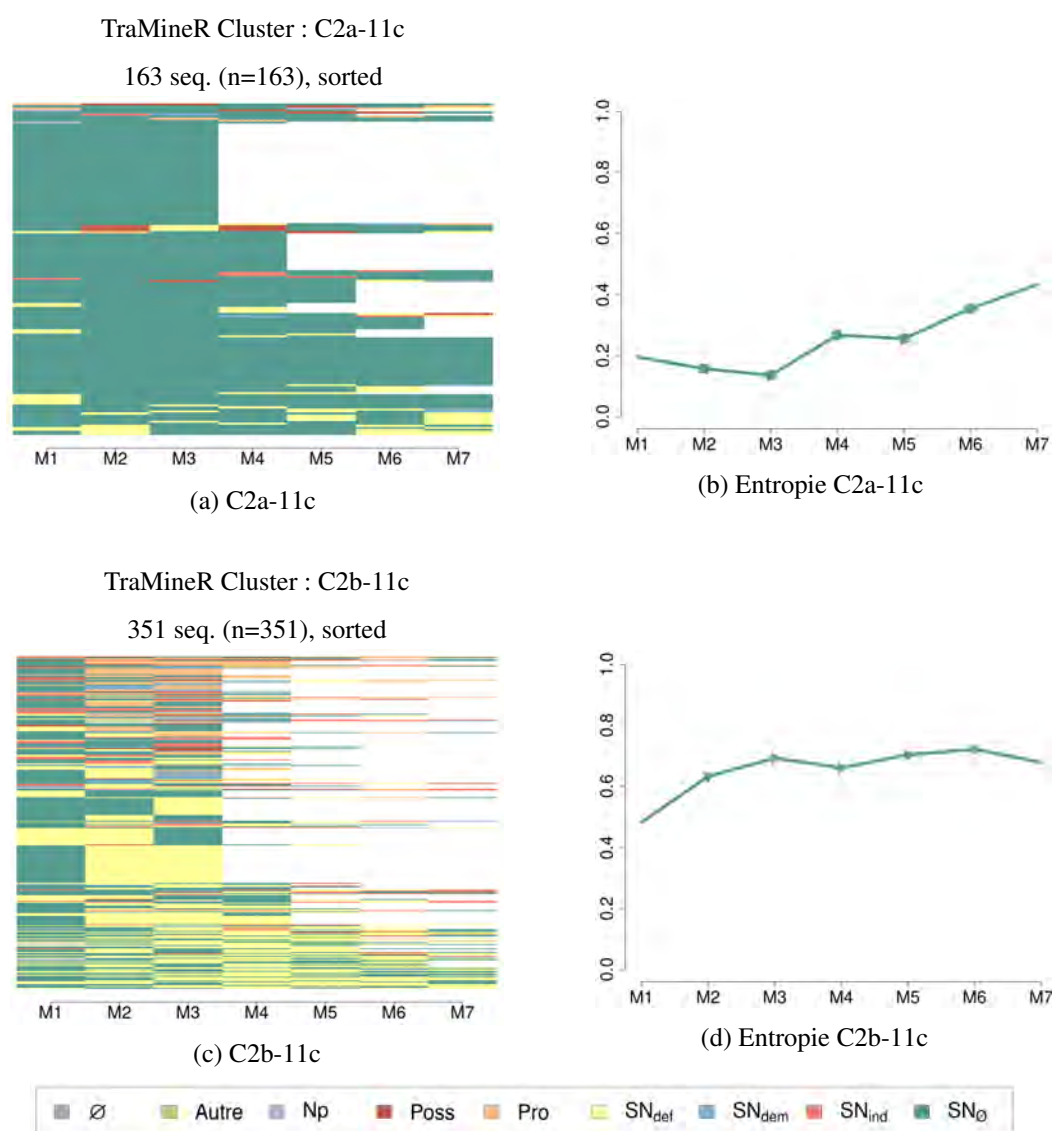


FIGURE 8.7 – Clusters résultant de la division du C2-4c avec leur entropie

(89,6%) et est composée exclusivement de SN_0 (65% des CR), avec quelques exceptions caractérisées principalement par la présence de SN_{def} . Le deuxième cluster, C2b-11c (figure 8.7c), regroupe 351 CR (soit 68,3% des CR du C2-4c) dont la structure est très proche de celle du C2-4c. Un grand nombre de CR commence par un SN_0 (66,1%) ou un SN_{def} (20,5%). Plus marginalement, nous retrouvons des SN_{ind} , des Np, des possessifs, des pronoms et des SN_{dem} , comme c'était le cas dans le C2-4c. Les maillons suivants présentent une alternance de SN_0 et SN_{def} et des pronoms (visibles dans la partie haute du cluster).

L'analyse des séquences représentatives (figure 8.8) met en évidence les deux types de composition différents. Pour le C2a-11c, les séquences représentatives sont toutes constituées exclusivement de SN_0 , avec des longueurs différentes (de 3 à 7 états). Ces cinq séquences couvrent 65% du cluster. Les cinq CR ayant toutes le même type de

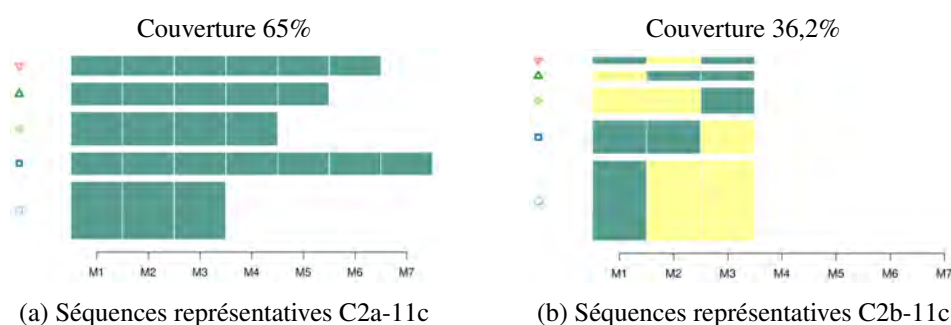


FIGURE 8.8 – Clusters et séquences représentatives résultant de la division du C2-4c

composition, nous proposons un exemple seulement, en sélectionnant la CR la plus longue (les autres exemples sont donnés en annexe E.1).

- (41) a.  Les registres seront cotés par première et dernière, et paraphés sur chaque feuille, par le président du tribunal de **première instance**, ou par le juge qui le remplacera. [...] l'un des doubles sera déposé aux archives de la commune, l'autre au greffe du tribunal de **première instance** [...] Les extraits délivrés conformes aux registres, et légalisés par le président du tribunal de **première instance** [...] Dans tous les cas où la mention d'un acte relatif à l'état civil devra avoir lieu en marge, d'un autre acte déjà inscrit, elle sera faite à la requête des parties intéressées, par l'officier de l'état civil, sur les registres courans ou sur ceux qui auront été déposés aux archives de la commune, et par le greffier du tribunal de **première instance** sur les registres déposés au greffe à l'effet de quoi l'officier de l'état civil en donnera avis dans les trois jours au commissaire du Gouvernement près ledit tribunal [...] Toute contravention aux articles précédens de la part des fonctionnaires y dénommés sera poursuivie devant le tribunal de **première instance**, et punie d'une amende qui ne pourra excéder cent francs. [...] Le commissaire du Gouvernement près le tribunal de **première instance** sera tenu de vérifier l'état des registres lors du dépôt qui en sera fait au greffe il dressera un procès verbal sommaire de la vérification dénoncera les contraventions ou délits commis par les officiers de l'état civil et requerra contre eux la condamnation aux amendes. [...] Dans tous les cas où un tribunal de **première instance** connaîtra des actes relatifs à l'état civil, les parties intéressées pourront se pourvoir contre le jugement. [...]

[TJ_{CCF}; 

Les séquences représentatives du C2b-11c (couvrant 36,2% du cluster) présentent une composition alternant des SN_{def} et des SN_∅. Les séquences représentatives de ce deuxième cluster ont toutes une longueur égale à trois états. Nous en présentons deux exemples, les autres sont disponibles en annexe E.2.

- (42) a.  – Il l'appelait : scélérat, bête féroce, homme sans **moralité**. Et si ce langage avait pour objet de rétablir dans l'esprit de l'enfant le sens de la justice, il faut avouer qu'il manquait son but ; car la logique enfantine

risquait fort de conclure : « Si un grand homme comme celui-là n'avait pas de **moralité**, c'est donc que **la moralité** n'est pas grand-chose, et que la première affaire, c'est d'être un grand homme. [R_{JC} ; ⚡]

- b. ⚡ **Article 18** Dommage causé à la marchandise 1 [...] Le transport aérien, au sens du paragraphe 1 **du présent article**, comprend la période pendant laquelle la marchandise se trouve sous la garde du transporteur. [...] Dans le cas de transport intermodal effectué en partie par air et en partie par tout autre moyen de transport, les dispositions de la présente convention ne s'appliquent, sous réserve du paragraphe 4 de **l'article 18**, qu'au transport aérien et si celui-ci répond aux conditions de l'article 1. [TJ_{CTA} ; ⚡]

Interprétation des clusters C2a-11c et C2b-11c Ce qui est intéressant dans ces exemples est la nature syntaxique des expressions référentielles : dans les cinq exemples des séquences représentatives (exemple (54), en annexe E.1), ces expressions sont toutes des compléments de nom, indiquant qu'il ne s'agit pas du référent principal. Il en va de même pour les CR des exemples des séquences représentatives du C2b-11c (42a) et (42b) (et ceux en annexe E.1). Les deux clusters semblent donc regrouper des CR d'arrière plan, qui correspondent à des référents secondaires.

Cet aspect est directement lié aux choix d'annotation du corpus Democrat. Les compléments de noms sont annotés comme des expressions pouvant donner lieu à des CR, comme c'est le cas dans les exemples que nous venons d'analyser (cf. chapitre 3). Ce type d'annotation est également celle que l'on obtient avec des systèmes de résolution automatique, qui identifient toutes les expressions référentielles. Nous retrouvons ici des CR qui ne sont généralement pas décrites dans les études existantes. Par ailleurs, ces CR s'éloignent de la définition de chaîne en tant que structure discursive contribuant à la cohésion du discours (cf. chapitre 1). Les CR regroupées dans ces clusters seraient alors des CR qui n'ont pas de pouvoir structurant.

Un autre aspect intéressant dans les exemples concerne la nature lexicale des expressions référentielles. Nous pouvons observer que l'introduction et le maintien se font par le biais de reprises directes à l'identique. Les cinq exemples en question relèvent tous de textes non narratifs et notamment de textes juridiques, ce qui pourrait expliquer la stabilité lexicale des maillons : les textes juridiques sont caractérisés par l'utilisation de termes spécifiques du domaine, ce qui pourrait expliquer les reprises à l'identique sans apport d'informations supplémentaires (modificateurs ou autres compléments). Par exemple, dans (41a) *la première instance* est un complément du nom *tribunal de première instance*, et le terme n'est pas modifié.

Dans les exemples du C2b-11c, nous observons des cas de figure variés lorsque le maillon correspond à des SN_∅. Notamment, nous observons le même cas de figure que les exemples précédents, à savoir des CR où les SN_∅ est un complément de nom. Les exemples (55) correspondant aux séquences représentatives du C2b-11c relèvent tous de textes non narratifs, à l'exception de (55d), ce qui laisse penser que le type de texte a une influence sur la présence de ce type de composition. De plus, nous remarquons qu'il s'agit principalement d'articles présents dans les conventions ((55e), (55c) et (55a)), l'analyse des facteurs de variation permettra de confirmer s'il existe un

lien entre type textuel, genre textuel et ce type de composition de CR. Si tel est le cas, nous avons identifié un type de composition qui serait propre à des textes juridiques.

8.3.3 Cluster C3-4c DEMRC : SN_{def}++

Le troisième cluster, le C3-4c (figure 8.4_c), regroupe 1039 séquences (soit 24,5% du total des chaînes analysées) constituées principalement de SN_{def}. En termes de caractérisation linéaire, la plupart des chaînes de ce cluster commencent par un SN_{def} (844 CR, soit 81,2% des CR dans ce cluster) et sont constituées exclusivement de SN_{def} (382 CR, soit 36,8% des CR de ce cluster). Les autres CR regroupées dans ce cluster présentent une alternance de SN_{def}, de possessifs et de pronoms, plus rarement de SN_{dem}. De manière plus marginale, des CR démarrent par un autre type de maillon : possessif (7,7%), SN_{ind} (4,5%), et plus marginalement encore, SN_∅, Np, sujet zéro, pronom, SN_{dem}. Cependant, la tendance pour les maillons suivants reste la même. On retrouve principalement des SN_{def}. Ces observations semblent correspondre à deux tendances différentes, avec des chaînes plutôt homogènes et des chaînes dont la composition est plus hétérogène. Étant donné le grand nombre de CR présentes dans ce cluster, il est préférable de s'intéresser aux divisions plus fines.

Les tendances en détail Lors de la classification en onze clusters, les CR qui étaient regroupées dans le C3-4c sont séparées en deux clusters différents (figure 8.10) : le C3a-11c regroupe 250 CR et le C3b-11c en regroupe 789. Bien que la structure de ces deux clusters reste proche de celle que nous avons observée avec l'analyse à gros grain, cette division permet de séparer les séquences homogènes (C3a-11c, figure 8.10_a) et celles plus hétérogènes (C3b-11c, figure 8.10_c). L'observation de l'entropie transversale confirme cette distinction.

Dans le C3a-11c 92% des séquences commencent avec un SN_{def}. Concernant la suite des maillons, dans ce premier cluster, on a principalement affaire à d'autres SN_{def}. La nature du premier maillon est plus variée dans le C3b-11c, bien que la présence des SN_{def} reste majoritaire (77,8% des CR du cluster).

Les tendances que nous venons de décrire sont confirmées par l'analyse des séquences représentatives (figure 8.10). Les cinq séquences représentatives du C3a-11c sont toutes constituées d'au moins cinq maillons et couvrent 46,4% du cluster. Nous observons également que les plus fréquentes sont constituées exclusivement de SN_{def}. Les deux moins fréquentes, commencent par SN_{def}, le deuxième maillon correspond respectivement à un SN_∅ et un pronom, et les maillons suivants sont tous des SN_{def}. Nous proposons ici trois exemples, qui reflètent ces tendances, dont un seul exemple de CR composée uniquement de SN_{def} ((43_c)), les autres sont donnés en annexe E.3.

- (43) a. ▼ [...] Rappelant les résultats positifs obtenus dans le contexte de la convention pour la prévention de la pollution marine par les opérations d'immersion effectuées par les navires et aéronefs, signée à Oslo le 15 février 1972, telle qu'amendée par les protocoles du 2 mars 1983 et du 5 décembre 1989 [...]
- q) On entend par « convention d'Oslo » : la convention pour la prévention de la pollution marine par les opérations d'immersion effectuées par

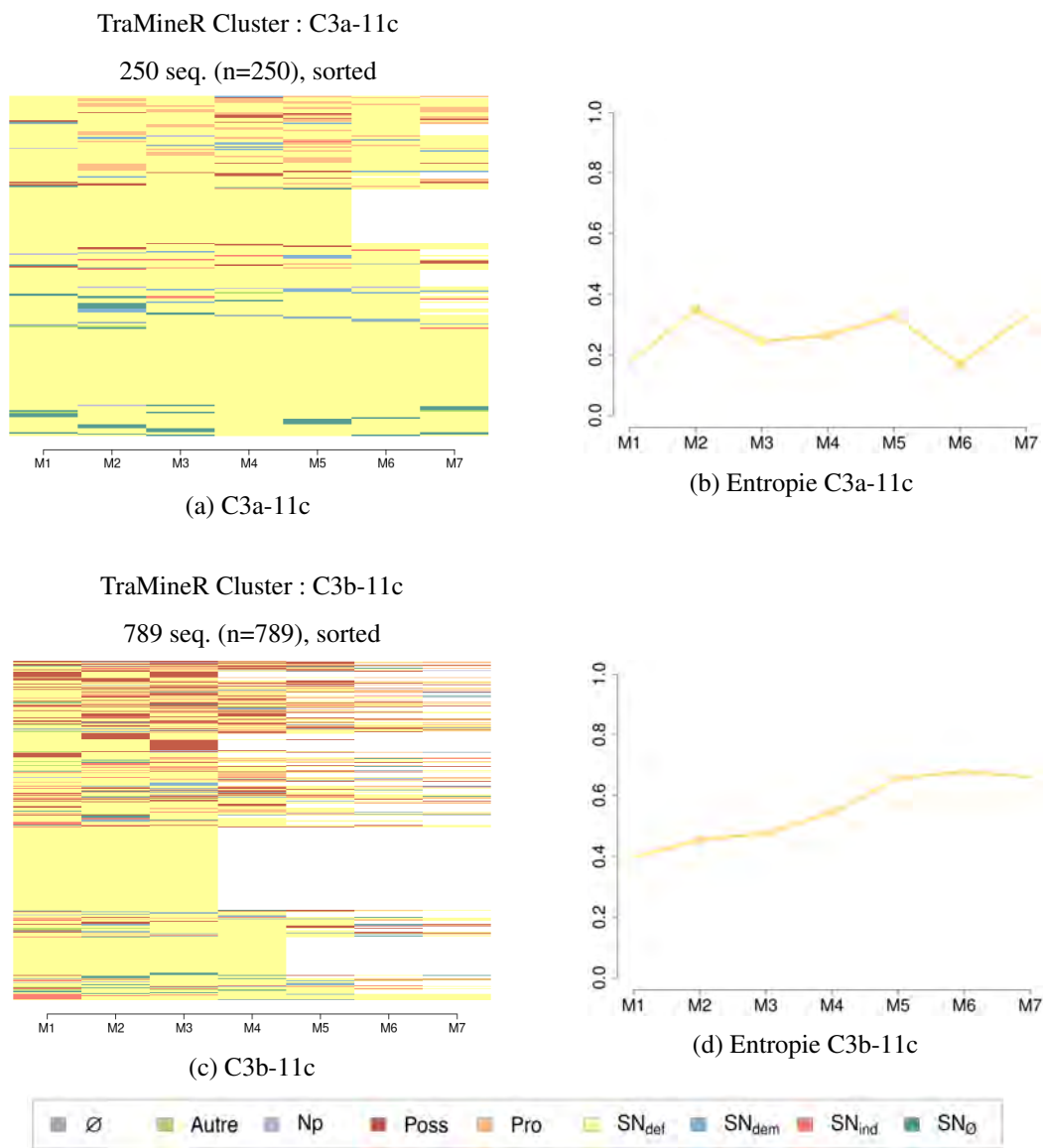


FIGURE 8.9 – Clusters résultant de la division du C3-4c avec leur entropie

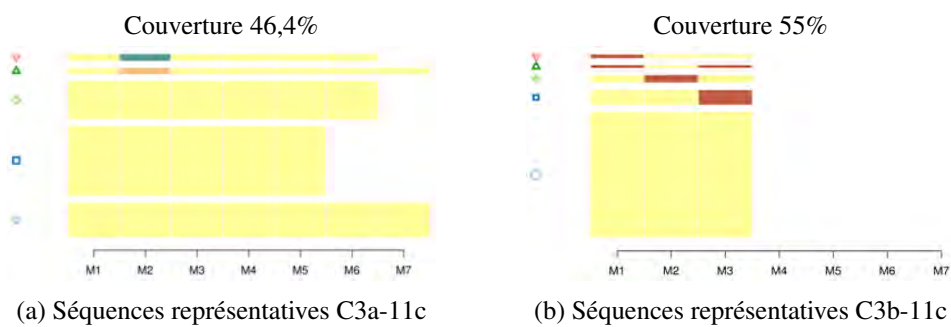


FIGURE 8.10 – Séquences représentatives des clusters résultant de la division du C3-4c

les navires et aéronefs, signée à Oslo le 15 février 1972, telle qu'amendée par les protocoles du 2 mars 1983 et du 5 décembre 1989. [...]

La convention est ouverte à la signature à Paris, du 22 septembre 1992 au 30 juin 1993 par :

a) les parties contractantes à la convention d'Oslo ou à la convention de Paris ; [...]

La convention entrera en vigueur le trentième jour suivant la date à laquelle toutes les parties contractantes à la convention d'Oslo et toutes les parties contractantes à la convention de Paris auront déposé leur instrument de ratification, d'acceptation, d'approbation ou d'adhésion. [...]

Sans préjudice des dispositions du paragraphe 1, les décisions, les recommandations et les autres accords adoptés en application de la convention d'Oslo ou de la convention de Paris continuent d'être applicables et conservent le même caractère juridique [...]

[TJ_{CMM} ; ⚡]

b. ▲ Il avait déjà passé quinze jours dans l'état d'extase qui saisit toutes les jeunes imaginations à l'aspect de la reine des ruines, quand, un soir, il entra au théâtre d'Argentina, devant lequel se pressait une grande foule. [...] Elle s'avança par coquetterie sur le devant du théâtre, et salua le public avec une grâce infinie. [...] Sarrasine voulait s'élancer sur le théâtre et s'emparer de cette femme. [...] Bientôt, il fut obligé de quitter le théâtre [...] En sortant du théâtre, un inconnu l'arrêta par le bras. [...] Il reconnut les chanteurs et les cantatrices du théâtre, mêlés à des femmes charmantes, tous prêts à commencer une orgie d'artistes qui n'attendait plus que lui. [...]

[N_{S} ; ⚡]

c. ◆ Un livre sur la fièvre jaune avait porté au loin le nom de Richard Bayley et fortement intéressé le monde scientifique d'alors. [...] Elle y était au mois d'août 1801, quand la fièvre jaune éclata à New-York. C'était surtout parmi les pauvres émigrants irlandais que le fléau sévissait. [...] Il y prit la fièvre : et, après six jours de cruelles souffrances, expira entre les bras de sa fille, en implorant la miséricorde de son Sauveur. [...] Mais, au départ, on avait négligé certaines formalités, et, la fièvre jaune sévissant à New-York, l'équipage et les passagers du brick américain allaient être traités comme des pestiférés. [...] On venait d'apprendre, à Livourne, que la fièvre jaune sévissait à New-York, et comme nous n'avions pas de certificat de la Santé, il n'y avait rien à répondre.

[B_{ES} ; ⚡]

Concernant le C3b-11c, les séquences représentatives sont constituées uniquement de trois maillons et couvrent 55% du cluster. Là encore, les plus fréquentes sont composées exclusivement de SN_{def}. Les autres types de composition sont caractérisés par la présence de possessifs qui alternent avec les SN_{def}. Nous soulignons que les CR les plus longues et qui présentent plus de variété ne sont pas prises en compte dans les cinq séquences représentatives.

(44) a. ▼ Le jeune homme occupait son logement depuis deux mois environ et était dépressif. [...] L'occupant de l'appartement du rez-de-chaussée., à l'origine du sinistre, a été brûlé au visage et aux bras. [...] Réveillés en sursaut, les huit locataires, voisins de l'appartement du rez-de-chaussée où l'explosion s'est produite, ont ouvert les yeux sur un spectacle de

- désolation : leur appartement n'était plus qu'un amas de ce qui restait des murs et de leur mobilier. [AP_{ER} ; ⚡]
- b. ▲ Sans doute avez-vous oublié que vous avez choisi **mes meubles**? [...] Dès qu'on s'en éloignait, il faisait nuit, et on se cognait **aux meubles** [...] Je revins tous les soirs ; je ne pensai même pas à la prier de me montrer sa chambre, encore moins à lui demander comment Jacques trouvait **nos meubles**. [R_{DaC} ; ⚡]
- c. ◆ Le père est toujours maître d'abréger la durée de **la détention par lui ordonnée ou requise**. Si après **sa** sortie l'enfant tombe dans de nouveaux écarts, **la détention** pourra être de nouveau ordonnée de la manière prescrite aux articles précédents. [TJ_{CCF} ; ⚡]
- d. ■ Les habitants du secteur lui ont fait part de leurs inquiétudes face aux risques que peut faire courir le remblai de la vallée de **la Meuse**, les travaux du tronçon de la Woëvre devant bientôt débiter, les déblais et remblais devront être terminés en 2004. [...] Se rassembler, s'ouvrir, s'unir avec d'autres villes et d'autres régions pour que **la Meuse** se développe et enraye **sa** désertification, tel est le message que l'on peut retenir. [AP_{ER} ; ⚡]
- e. ● La Commission exerce les fonctions suivantes, en accordant la priorité aux thons et aux espèces apparentées : [...]
- d) déterminer si, au vu des meilleures informations scientifiques disponibles, un stock de poissons donné visé par la présente convention est totalement exploité ou surexploité et, sur cette base, si une augmentation de la capacité de pêche ou du niveau de **l'effort de pêche** est susceptible de mettre en péril la conservation de ce stock ; [...]
- h) adopter les mesures appropriées en vue d'empêcher ou de faire cesser la surexploitation et la surcapacité de pêche et faire en sorte que **l'effort de pêche** n'atteigne pas un niveau incompatible avec l'utilisation durable des stocks de poissons visés par la présente convention ; [...]
- l) définir, en tant que de besoin, des critères et prendre des décisions, portant sur la répartition du volume admissible de captures, ou la capacité de pêche totale admissible, y compris la capacité de charge, ou le niveau de **l'effort de pêche**, en tenant compte de tous les facteurs pertinents ; [...]
- [TJ_{CTT} ; ⚡]

Interprétation des clusters Les deux clusters regroupent des CR caractérisées principalement par la présence d'expressions de faible accessibilité, aussi bien pour introduire que pour maintenir le référent. Comme nous l'avons dit plus haut, il s'agit principalement de SN_{def}. Dans (43a), (43c) et (44c) nous pouvons observer que les expressions qui occupent la première position dans la CR sont des SN_{def} avec leurs expansions : la présence des expansions dans les maillons reflète les choix d'annotation adoptés. Cela nous permet d'observer que lorsque le référent est introduit pour la première fois, l'introduction passe par un transfert de toutes les informations nécessaires pour la bonne interprétation, alors que lorsque le SN_{def} est utilisé pour la reprise du référent, la mention peut être moins complète (*au théâtre d'Argentina* → *le théâtre*, dans (43b)). Nous retrouvons ici les considérations de Corblin (1987) : les SN_{def} occupent la première position lorsqu'ils portent l'information nécessaire et suffisante pour garantir une interprétation correcte, indépendamment du contexte. Alors que lorsque le

SN_{def} se trouve en deuxième position, sa fonction serait de réactiver le référent déjà mentionné. Dans les exemples, nous observons une forte stabilité lexicale lorsque les reprises sont faites par le biais de SN_{def} . Une analyse croisée entre les clusters et le coefficient de stabilité pourrait constituer une piste intéressante pour évaluer la stabilité des CR présentant ce type de composition.

Les exemples (44a) et (44b) illustrent les cas de figure où l'introduction du référent passe par un possessif. Nous constatons qu'il s'agit ici des SN_{poss} . Le déterminant possessif, quant à lui, est considéré coréférent avec son possesseur (reflet des choix d'annotation), c'est pourquoi nous le retrouvons plutôt dans des positions de deuxième maillon ou de maillon-n. Dans ces cas, les mentions sont assez proches, comme l'illustrent les exemples (44c) et (44d).

L'utilisation d'expressions de faible accessibilité pour le maintien du référent peut s'expliquer par différents facteurs : distance entre les maillons, la saillance du référent (Salles, 2015a), compétition des référents (Corblin, 1987 ; Kleiber, 1994) ou nature même du référent (Schnedecker & Landragin, 2014). Selon les facteurs qui entrent en jeu, nous pouvons distinguer des CR dont le statut référentiel est en danger à cause de la présence de référents en compétition ou de la distance entre les expressions référentielles qui peut impacter l'activation du référent. Lorsque le référent n'est pas saillant, on aurait affaire à des CR qui concernent des référents secondaires. Pour conclure, nous pouvons dire que les CR regroupées dans ces deux clusters sont des chaînes homogènes et semblent être stables d'un point de vue lexical.

8.3.4 Cluster C4-4c DEMRC : des séquences hétérogènes

Le C4-4c regroupe le plus grand nombre de séquences : 2 455 séquences (soit 58% du total des séquences analysées). Comme nous pouvons l'observer dans la figure 8.4d, ce cluster est hétérogène. Plus d'un tiers (34,4%) des CR de ce cluster commencent par un SN_{def} , un quart (25,7%) par un SN_{ind} . Les CR regroupées dans ce cluster commencent également avec des possessifs (13,2%) et des pronoms (11,65%). L'analyse de l'entropie transversale de ce cluster (figure 8.5), confirme l'hétérogénéité de cette classe, notamment pour le premier maillon. La variation est légèrement plus faible pour les maillons suivants, mais elle reste élevée.

Lorsque l'on s'intéresse aux modalités d'enchaînement des maillons, le nombre très élevé de CR regroupées dans ce cluster et l'hétérogénéité de la classe ne permettent pas de distinguer clairement les tendances associées à différents types de composition. Il est alors préférable de prendre en considération une division plus fine de cette classe. Par ailleurs, nous avons vu dans la figure 6.5 que ce cluster est divisé à plusieurs reprises : les regroupements effectués donnent lieu à six clusters différents. Ainsi, la prise en compte de la configuration en onze clusters permet de rendre compte de ces regroupements, facilitant l'identification des types de composition.

Dans ce qui suit, nous présentons donc les six clusters résultant de la division du C4-4c, en suivant les regroupements observés dans la figure 6.5. Pour ce faire, nous nous concentrons sur les divisions finales obtenues avec une classification en onze clusters, c'est-à-dire que nous ne commentons pas les distinctions observées à des niveaux antérieurs. Pour mieux comprendre, nous illustrons dans la figure 8.11 les divisions qui

sont prises en compte avec la classification en onze clusters. Les subdivisions intermédiaires des niveaux antérieurs sont grisées. Par exemple, le cluster C4-4c est d'abord divisé en deux classes, chacune subissant à son tour des subdivisions à plusieurs reprises. Ainsi, C4a-5c donne lieu à deux clusters : l'un se subdivise en deux classes (C4a-9c et C4b-9c), l'autre ne subit pas d'autres divisions. Les lignes pointillées indiquent ce cas de figure (figure 8.11).

Des introductions différentes Nous commençons par le C4a-11c (figure 8.12_a) et le C4b-11c (figure 8.12_c). La principale différence entre ces deux clusters concerne la nature des premiers maillons : les chaînes regroupées dans le premier cluster C4a-11c commencent par des SN_{ind} (78,2% des CR du cluster) alors que dans le C4b-11c le premier maillon des CR est principalement un SN_{def} (79,2% des CR). Concernant la suite, les reprises sont effectuées par des pronoms, des SN définis et des démonstratifs dans les deux cas. Avec une différence : les démonstratifs semblent être plus nombreux dans le deuxième cluster que dans le premier. Nous remarquons la présence de SN indéfinis dans des positions autres que la première pour le C4a-11c alors qu'ils sont moins présents dans les C4b-11c (surtout à partir du M3). Dans les deux clusters, de nombreuses CR sont constituées de trois maillons : 43,1% dans le C4a-11c et 58,4% dans le C4b-11c.

Les séquences représentatives des deux clusters permettent de mettre en évidence deux stratégies d'introduction différentes (figure 8.12). Les cinq séquences représentatives de chaque cluster ont une longueur de trois maillons et couvrent respectivement 43,3% du C4a-11c et 54,4% du C4b-11c.

Les séquences les plus fréquentes (avant-dernière et dernière dans la figure 8.13) diffèrent par la nature du premier maillon, mais la nature des maillons suivants est la même. Des différences sont visibles pour les trois autres séquences représentatives : dans C4a-11c nous observons la présence d'un pronom en premier maillon et un SN_{ind} en deuxième position, ce qui peut paraître étonnant. Les exemples (45) illustrent les cinq séquences représentatives du cluster C4a-11c, qui reflètent les différentes modalités.

Dans C4b-11c, nous observons l'alternance de SN_{def} (en premier maillon) et dans la suite de SN_{dem} et de pronoms. La présence des SN_{dem} dans les cinq séquences représentatives semble confirmer notre observation portant sur la fréquence plus importante de ces formes dans le C4b-11c par rapport au C4a-11b.

Interprétation des clusters C4a-11c et C4b-11c En termes de stratégies d'introduction et de maintien des référents, ces deux clusters mettent en avant l'utilisation d'expressions de faible accessibilité pour l'introduction des référents : des SN_{ind} et des SN_{def}. Le maintien est assuré principalement par des expressions de haute accessibilité, comme les pronoms ou les SN_{dem}. Le choix entre SN_{ind} et SN_{def} pour l'introduction du référent pourrait dépendre du degré de familiarité ou d'importance du référent. Schne-decker et Landragin (2014) soulignent que lorsque le référent est introduit par le biais d'un SN_{ind}, celui-ci n'est pas très important et ne constitue pas, de leur point de vue, un « focus d'attention ». Nous pouvons observer que les quatre exemples de (45) où l'introduction est faite par le biais de SN_{ind} sont des CR courtes, où le référent reste

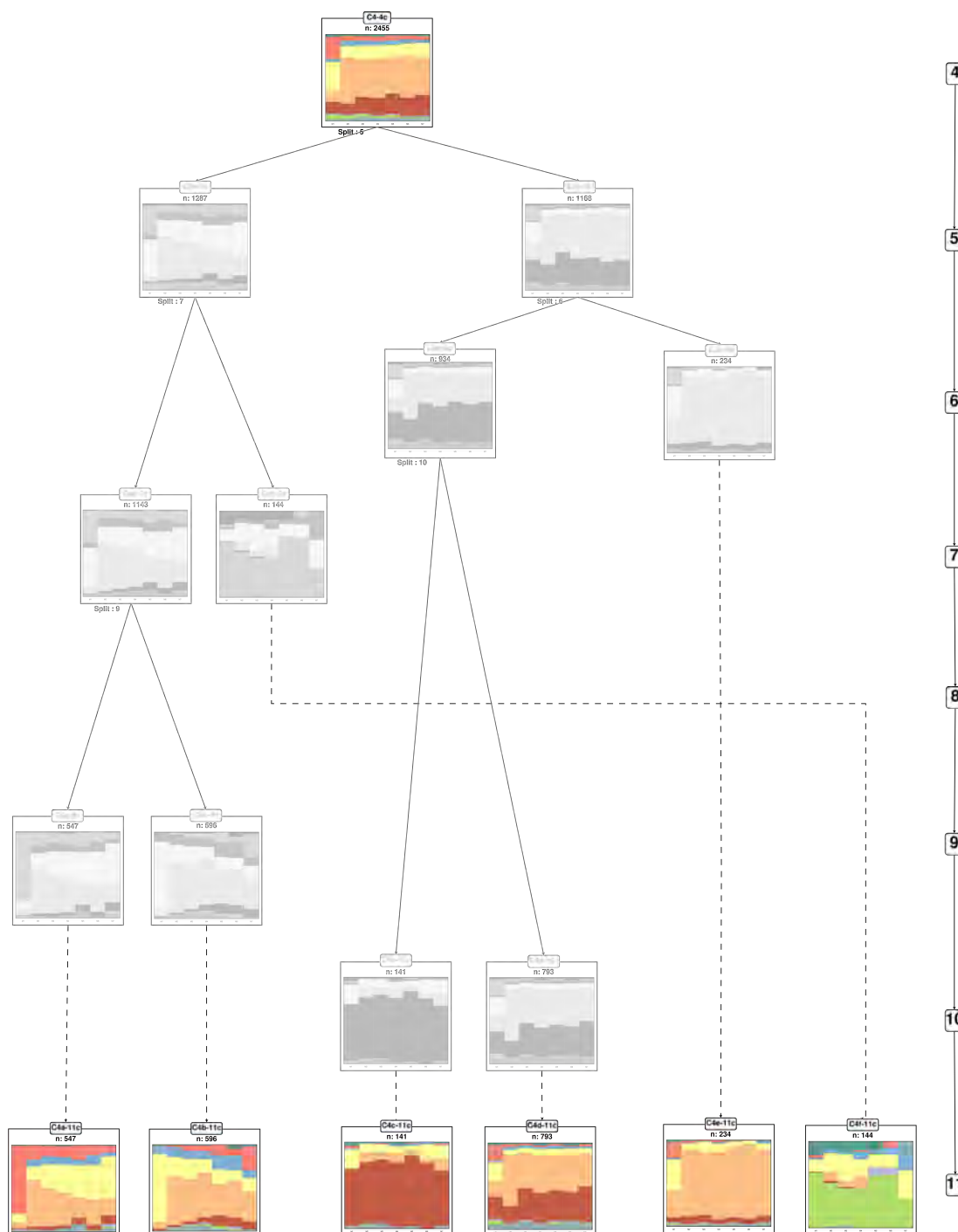


FIGURE 8.11 – Subdivisions du C4-4c selon les différents niveaux de classification. Les lignes pointillées indiquent que le cluster ne subit pas de subdivisions ultérieures

actif localement. Nous avons alors affaire à des chaînes éphémères ou locales, ce qui correspond à l'idée que le référent en question n'est pas le plus important dans le texte. Néanmoins, il se peut que ce caractère éphémère soit également lié au genre textuel : les (45a), (45c) et (45d) sont tirés de textes juridiques, et notamment du Code civil des

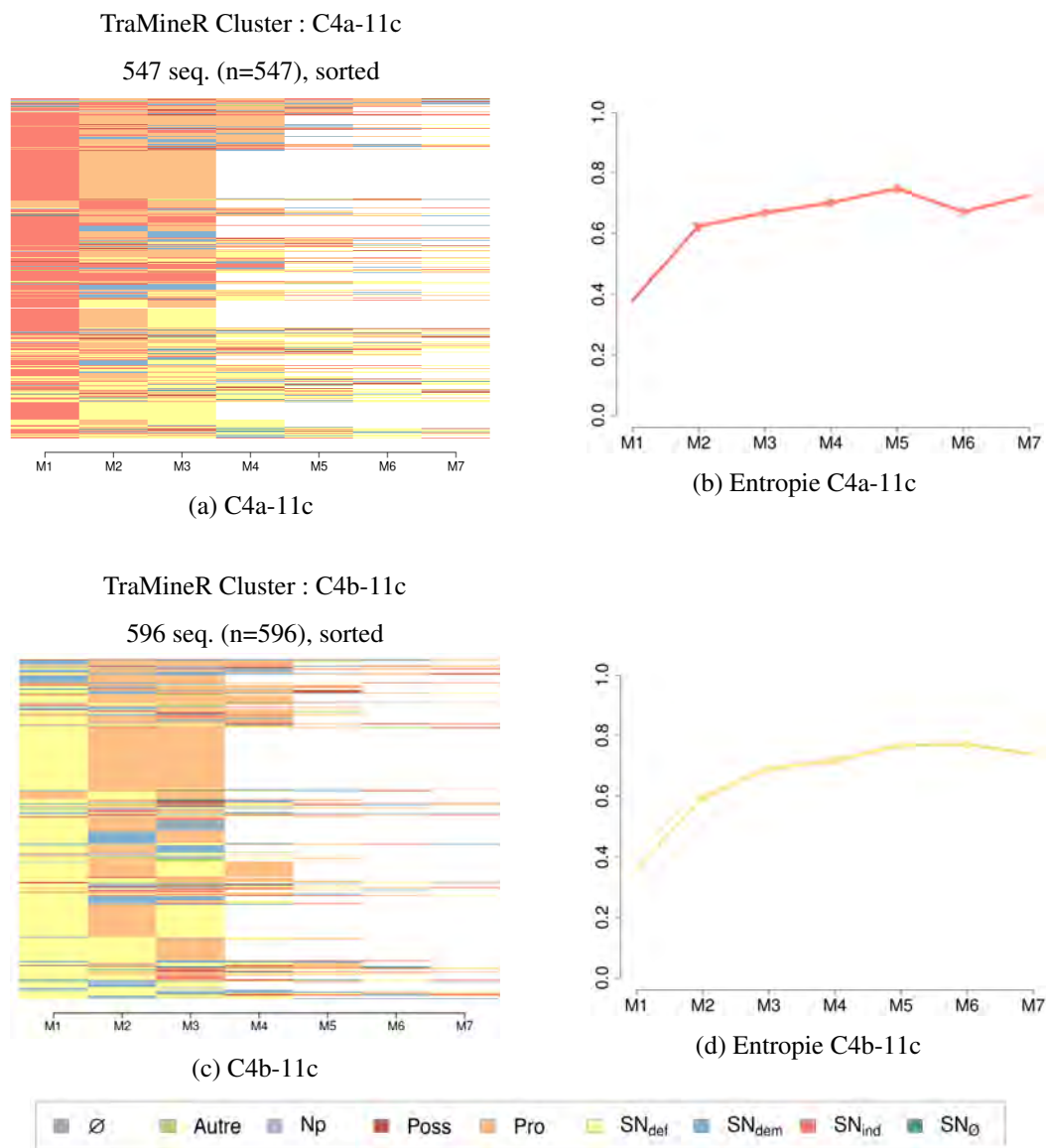


FIGURE 8.12 – Clusters résultant de la division du C4-4c avec leur entropie

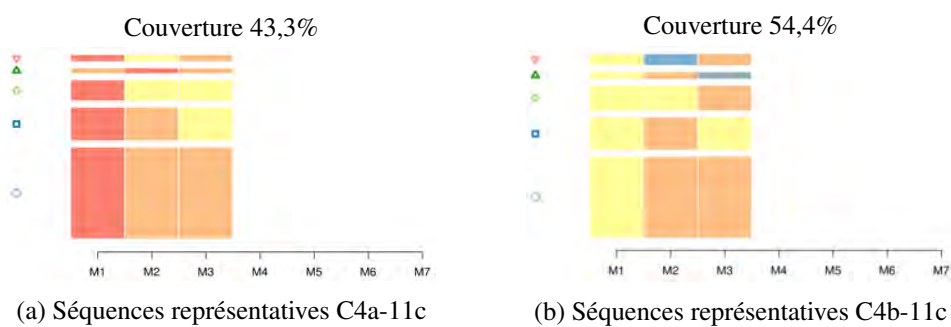


FIGURE 8.13 – Séquences représentatives des clusters résultant de la division du C4-4c

français et du Code de la procédure pénale. La structuration en articles de ces documents pourrait expliquer le caractère éphémère des CR. Plus étonnante est la présence de SN_{ind} dans des positions autres que la première. D'après Corblin (1987), en raison de leur nature, les SN_{ind} ne peuvent pas occuper une autre position que celle de premier maillon (cf. 1.1.1). L'exemple (45b) illustre une chaîne où le SN_{ind} est en deuxième position, le premier maillon étant un pronom, ce qui peut paraître également surprenant. Toutefois, il s'agit ici d'une construction présentative (comme nous l'avons vu précédemment dans la section 8.2), et la présence du pronom en première position et du SN_{ind} en deuxième peut être une conséquence du modèle d'annotation.

- (45) a. ▽ Pour chacun de ces agents, une demande d'habilitation est adressée au procureur général près la cour d'appel de Paris par le magistrat délégué aux missions judiciaires de la douane. [...] Ce dossier comprend notamment :
- 1° Les demandes d'habilitation et les documents qui y sont joints ;
[...]
- b. ▲ Mais quand le bleu pâle de sa prunelle, ce bleu liquide comme de l'eau, se découvrait, encore languissant, encore fatigué, encore malade des récentes caresses, c'était comme une flamme rapide qui me brûlait, exaspérant mes ardeurs.
- c. ◆ De l'Incapacité, des Exclusions et Destitutions de la Tutelle. [...] La condamnation à une peine afflictive ou infamante emporte de plein droit l'exclusion de la tutelle [...] Toute délibération du conseil de famille qui prononcera l'exclusion ou la destitution du tuteur, sera motivée, et ne pourra être, prise qu'après avoir entendu ou appelé le tuteur.
- d. ■ Article R15-33-15
Le procureur de la République près le tribunal de grande instance de Paris, après avoir recueilli, le cas échéant, les observations des juges d'instruction et des présidents des chambres correctionnelles, établit, tous les deux ans, une proposition de notation des agents des douanes habilités à exercer des missions de police judiciaire qu'il transmet au procureur général près la cour d'appel de Paris. [...]
Article R15-33-16
Les propositions de notation et les notations prévues à l'article R. 15-33-15 sont établies sur un imprimé conforme au modèle fixé par le ministre de la justice.
- e. ● La place de chaque pierre avait un sens pour lui ; il les connaissait toutes

Concernant le maintien du référent, nous observons que, contrairement aux autres clusters, le maintien est fait par des expressions de haute accessibilité (tels les pronoms et les sujets zéro), ce qui peut indiquer un statut différent du référent : le référent ici reste activé, les expressions référentielles étant proches entre elles. Dans (46a), nous constatons également la présence d'un SN_{dem} en deuxième position. D'après Corblin (1987), les SN_{dem} peuvent avoir différentes fonctions : recatégoriser le référent ((47a)) ou apporter une nouvelle information sur celui-ci ((47b)). Les démonstratifs peuvent

également être utilisés comme reprise directe d'un SN_{ind} ((47a)). La nature lexicale des SN permet de classer la fonction des SN_{dem}.

- (46) a. ▼ La division interne du genre Simia est en effet basée sur l'antique distinction entre les animaux sans queue (les « singes des Anciens », Simiae veterum), ceux à queue courte (les « babouins », Papiones) et ceux à longue queue (les « guenons », Cercopithecii). En 1777, le naturaliste allemand Johann Christian Erxleben reprend cette première classification mais y inclut les travaux de Buffon en séparant les singes du Nouveau Monde. [AE_{w} ; ⚡]
- (47) a. Cette succion me fit un mal atroce, on eût dit qu'on m'arrachait de fines lamelles de chair ayant leurs racines dans le fin fond du cœur. Pour échapper à ce supplice, j'essayai de repousser la petite Reine. [...] Et, tout en pleurs, autant de la pitié qu'elle m'inspirait que du mal qu'elle continuait à me faire, je lui présentai le lait qui avait gardé toute sa tiédeur à l'abri de mon oreiller :
— Bois -le, ma Reine, puisque ta mère a emporté le tien pour le donner au petit garçon. [R_{DVM} ; ⚡]
- b. Le gouverneur et son secrétaire ayant pris place aux bancs armoriés qu'on avait préparés pour eux, une main habile joua les premiers accords d'un chant de Noël [...] Une voix douce et vibrante modulait alors les premiers mots de ce chant toujours si beau « Minuit, chrétiens. » En entendant cette voix -là, le gouverneur tressaillit et son voisin, M. Duflot, posa la main sur son cœur, comme pour en arrêter les battements.[...] Quand le chant pieux fut terminé, M. Duflot se pencha à l'oreille de Roger et lui demanda : « Qui donc chante ainsi ? » [...] Monsieur le curé, demanda, tout à coup : « M. Duflot, cette jeune fille, qui a si bien chanté « Minuit Chrétiens », a-t-elle ses parents à Némoville ? [R_{N} ; ⚡]

Pronoms et possessifs

Une autre distinction donne lieu aux clusters C4c-11c (figure 8.14_a), qui regroupe 141 CR (soit 3,3% des CR analysées) et le C4d-11c (figure 8.14_c) qui en regroupe 793 (soit 18,7% des CR analysées).

Le premier cluster regroupe des chaînes dont le premier maillon est principalement un possessif, qui représentent 65,2% des chaînes de ce cluster. Les chaînes restantes commencent principalement par un SN_{def} (22,7%). Les autres catégories sont moins fréquentes. Les chaînes sont ensuite composées principalement de possessifs.

Le deuxième cluster, quant à lui, regroupe des CR dont le premier maillon est principalement un possessif (25,3%), un SN défini (22,3%), un SN indéfini (18,3%) ou un pronom (17,6%). Le maintien des référents est fait par une alternance de pronoms et de possessifs. L'utilisation d'expressions de haute accessibilité suggère que le référent reste activé et dans le focus d'attention. Nous pourrions interpréter cela comme des CR

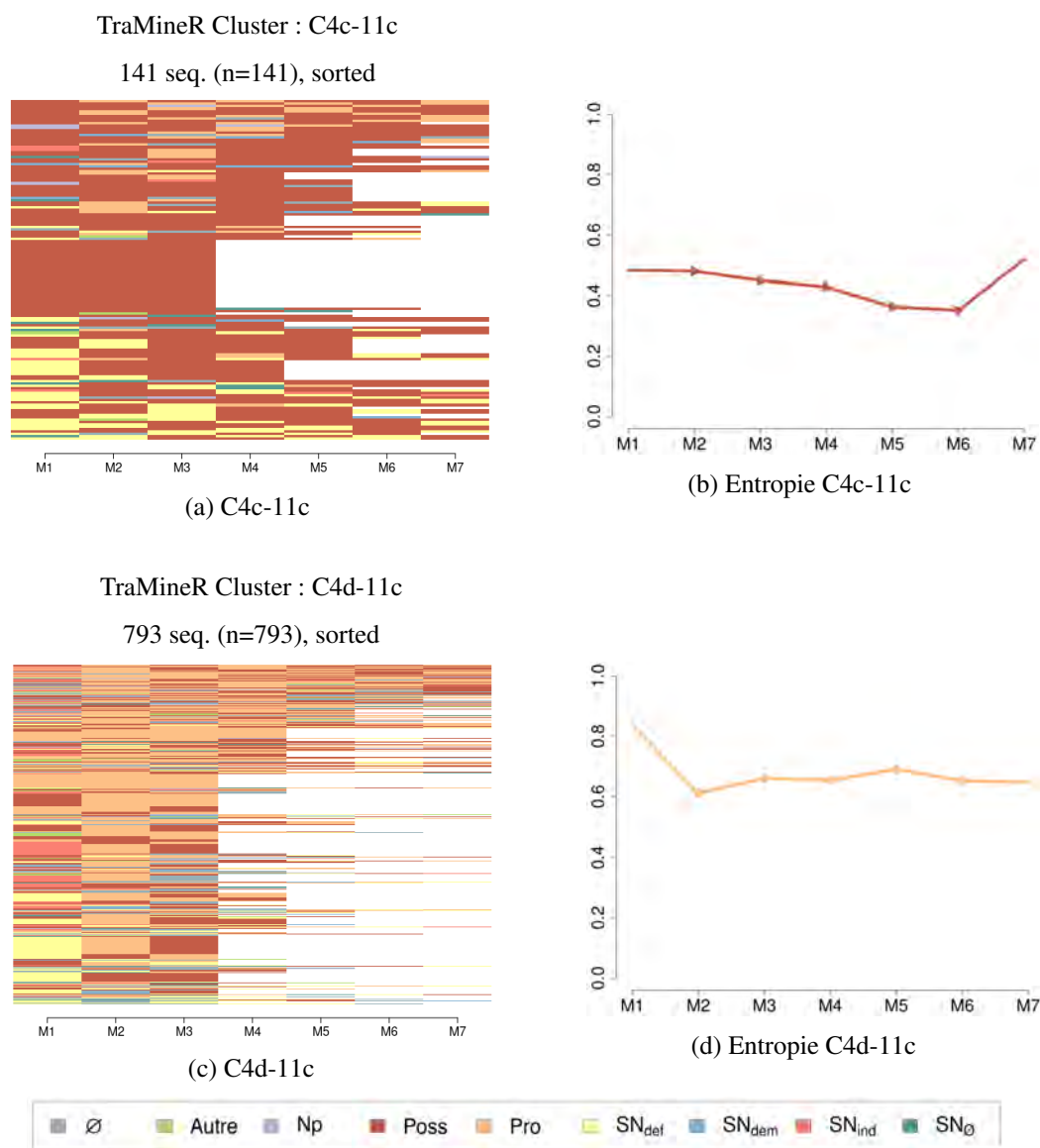


FIGURE 8.14 – Clusters résultant de la division du C4-4c avec leur entropie

dont le référent est un topique continu. Cependant, nous pouvons observer que la plupart des CR dans ce cluster sont constituées de 3 maillons. Cela pourrait alors signifier que ces CR portent sur un référent qui est un topique local. Le référent reste activé sur une courte portion de texte (il peut s'agir, par exemple, du topique d'un paragraphe). Concernant les démonstratifs, il se peut qu'ils indiquent un changement de point de vue, ou bien qu'ils apportent des informations supplémentaires sur le référent.

Le premier cluster est plus homogène que le deuxième : le score d'entropie est moyen, ce qui indique que la classe est plus homogène que le deuxième, dont l'entropie pour le premier maillon est proche de 1. Cela signifie qu'il existe beaucoup de variation dans la nature du premier maillon.

L'analyse des séquences représentatives permet de consolider nos observations :

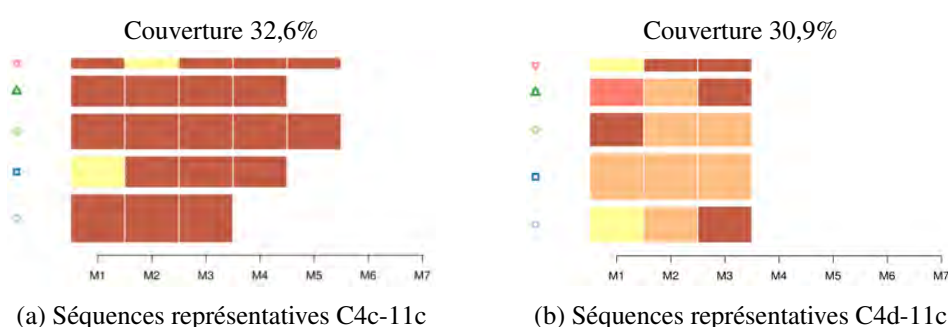


FIGURE 8.15 – Séquences représentatives des clusters résultant de la division du C4-4c

nous pouvons observer que le C4c-11c présente des séquences composées exclusivement de possessifs, une séquence dont le premier maillon est un SN_{def} et une séquence dans laquelle le deuxième maillon est un SN_{def} . Cela montre que les séquences de ce cluster présentent une composition plutôt homogène. Dans le C4d-11c, nous pouvons observer que les séquences représentatives sont plus variées, avec la présence de SN_{def} , SN_{ind} , de pronoms et de possessifs.

Interprétation des clusters C4c-11c et C4d-11c La présence de possessifs peut indiquer un référent secondaire, qui présente une relation de dépendance avec une autre référent du texte, comme illustré dans (48a). Comme nous l’avons anticipé dans la [section 8.2](#), la grande majorité des possessifs en première position sont des SN_{poss} (95,7% des possessifs, soit 88 sur 92). L’exemple (48a) illustre une chaîne dont le premier maillon est un SN_{poss} . Nous remarquons que dans ces CR, les maillons suivants sont également des SN_{poss} .

- (48) a. ▽ La pâleur de ses joues, le rose moins vif de ses lèvres, ses longs cils baissés et découpant leur frange brune sur cette blancheur lui donnaient une expression de chasteté mélancolique et de souffrance pensive d’une puissance de séduction inexprimable ; [...] La nuit s’avançait, et, sentant approcher le moment de la séparation éternelle, je ne pus me refuser cette triste et suprême douceur de déposer un baiser sur les lèvres mortes de celle qui avait eu tout mon amour. [...] le dedans de mes pauvres mains en est tout meurtri. [...] Et elle effleura mon front du bout de ses lèvres [...] De temps à autre elle s’interrompait pour me baiser la main, puis elle recommençait à presser de ses lèvres les lèvres de la plaie pour en faire sortir encore quelques gouttes rouges. $[N_{\{MA\}} ; \text{🌀}]$
- b. □ La Commission s’attache à adopter des mesures relatives à l’assistance technique, au transfert de technologie, à la formation et à d’autres formes de coopération, afin d’aider les pays en développement membres de la Commission à se conformer à leurs obligations découlant de la présente convention, ainsi que pour améliorer leur capacité à développer la pêche relevant de leur juridiction nationale respective et pour participer de manière durable à la pêche en haute mer. $[TJ_{\{CTT\}} ; \text{🌀}]$

En revanche, dans (48b), où le référent est introduit par un SN_{def}, les reprises suivantes via les possessifs ne concernent que le déterminant (à l'image également de (49a)), ce qui coréfère à *les pays en développement membres de la Commission*.

Le cluster C4d-11c présente plus de variété, comme le montrent les exemples des séquences représentatives.

- (49) a. ▽ Le jour où **les sirènes** lâchent **leurs** aigrettes de vapeur et **leurs** cris d'accouchées sur les échos des docks, ils attendent en échange de ce qu'ils abandonnent une liberté inconcevable des forces nues, et restées nues depuis Adam. [P_{A} ; ⚡]
- b. ▲ Et sans laisser au prêtre le temps de répliquer, il lui jeta sur la tête **un sac de toile goudronné**, **qu'** il ficela étroitement. [...] On fit quelques pas — une cinquantaine, peut-être — puis, le prêtre sentit qu'on enlevait **son bandeau**; il vit qu'il se trouvait dans une chambre vivement éclairée à l'électricité, au milieu de laquelle, sur un lit d'une extrême propreté, reposait un mourant. [R_{N} ; ⚡]
- c. ◆ Non contente d'avoir ainsi rempli ses tablettes, Rosalie le même jour écrivit à son père une lettre qui eut l'heur de le satisfaire pleinement, car voici ce qu'il y répondit : « Je suis toujours plus content de **tes lettres**, j'**en** serais même très fier si **elles** étaient mieux peintes. [B_{RdC} ; ⚡]
- d. ■ Elle **le** sera certainement. [...] Dans ma précipitation de marcher sans soutien d'aucune sorte, il y avait surtout la hâte de savoir jusqu'à quel point j'étais infirme, car je n'avais pas oublié les paroles du médecin de l'hôpital, ni la sécheresse de sa voix qui était comme un blâme à l'adresse de mes parents : « Boiteuse, elle **le** sera certainement ». Boiteuse, je **l'** étais, certainement. [R_{DVM} ; ⚡]
- e. ● Affaibli physiquement, crédité de 2% d'opinions favorables dans les sondages, Eltsine est paradoxalement en position de force face à **la majorité communiste de la Douma**, **qui** n'a pas réussi samedi à faire voter **son** projet de destitution du président. [AP_{ER} ; ⚡]

Dans les exemples (49), nous pouvons observer que les référents sont secondaires. Par ailleurs, les CR des séquences représentatives ont toutes une longueur égale à trois maillons. Cela semble suggérer que nous avons affaire à des CR secondaires et éphémères, dont le référent n'est actif que localement et dépend d'un autre référent.

Pronoms et autres Une dernière subdivision du C2-4c permet de distinguer deux classes : la première regroupe (cluster C7-11c) 234 CR dont le premier maillon est principalement un SN_{def} (36,6%), un pronom (30,3%) ou un SN_{ind} (18,8%). Les maillons suivants sont principalement des pronoms. Néanmoins, nous observons des alternances avec des possessifs, des démonstratifs ou de SN_{def}. La deuxième, C9-11c, regroupe 144 CR dont la nature des maillons appartient à la catégorie "autre" (63,9%) : principalement SN_{num} ou num, comme le montrent les exemples (58) en annexe E.5. On retrouve aussi des SN_∅ (16%) et des SN_{def} (14%), tant en premier maillon que dans les maillons suivants. Nous observons aussi la présence de SN_∅, de pronoms et de SN_{dem}.

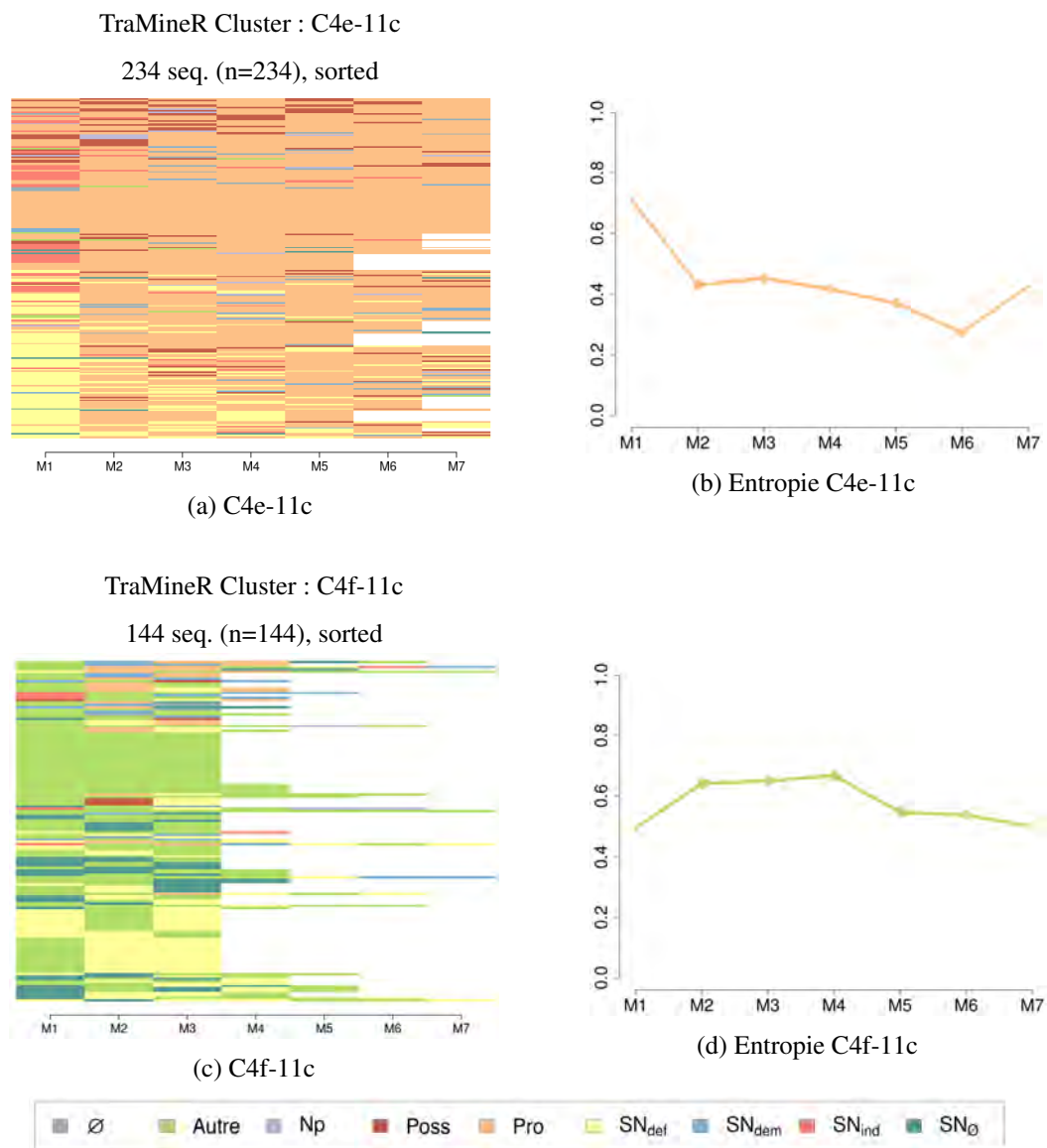


FIGURE 8.16 – Clusters et entropie des clusters résultants de la subdivision du C4-4c

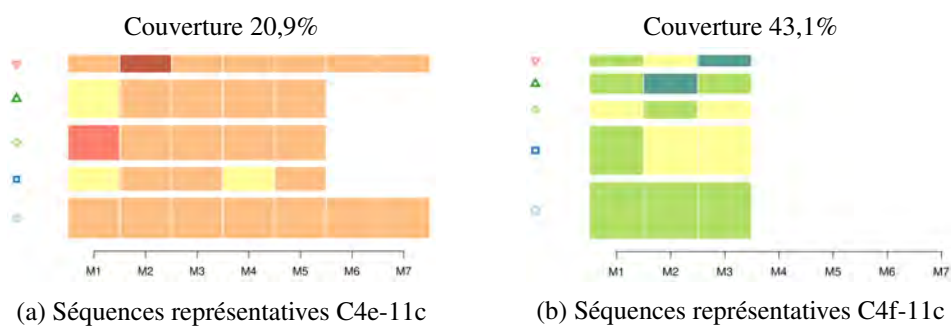


FIGURE 8.17 – Clusters et séquences représentatives résultant de la division du C4-4c

Interprétation des clusters C4e-11c et C4f-11c La présence de pronoms en premier maillon est étonnante si l'on considère que les pronoms sont des expressions de haute accessibilité. Comme nous l'avons déjà expliqué, cette anomalie peut s'expliquer par les choix d'annotation du corpus Democrat. Nous avons montré les cas de fusion (cf. 8.2). L'analyse des séquences représentatives nous a permis de relever deux autres cas de figure :

- les cas où le référent correspond au narrateur comme dans (50a);
- les cas où le référent est mentionné pour la première fois dans un discours direct, comme dans (50b).

(50) a. ▼ MADAME BAPTISTE

Quand j'entrai dans la salle des voyageurs de la gare de Loubain, mon premier regard fut pour l'horloge. J'avais à attendre deux heures dix minutes l'express de Paris. Je me sentis las soudain comme après dix lieues à pieds; puis je regardai autour de moi comme si j'allais découvrir sur les murs un moyen de tuer le temps; puis je ressortis et m'arrêtai devant la porte de la gare, l'esprit travaillé par le désir d'inventer quelque chose à faire. [...]

[N_{MF}; ⚡]

- b. Quand il avait fini : – Allons, prends- la donc ! [...] Il arrivait que son père lui reprochât de faire le difficile, et qu'il s'adjudgeât la dernière pomme de terre. Mais Christophe se méfiait maintenant; et il la réservait sur son assiette pour Ernst, le petit frère, toujours vorace, qui la guettait du coin de l'œil depuis le commencement du dîner, et qui finissait par lui demander : – Tu ne la manges pas ? Donne- la-moi, dis, Christophe.

[R_{JC}; ⚡]

Généralement, dans les dialogues, la présence de pronoms est assez importante. Dans ce cas, les pronoms auraient un rôle de déictiques dans le discours. Nous faisons l'hypothèse que ces cas sont fréquents, surtout dans les textes narratifs. Cependant, pour confirmer cette hypothèse, la prise en compte de la structure textuelle et l'identification des dialogues dans les textes seraient nécessaires, ce qui dépasse le cadre de cette étude.

Concernant le cluster C4f-11c, l'analyse des exemples des séquences représentatives a mis en évidence qu'il s'agit principalement de renvois à des articles dans les textes juridiques, comme le montre (51) (les autres exemples sont disponibles en annexe (E.5)).

(51) ▼ Titre IX : Dispositions finales

Article XXVII : Signature

1 [...]

En ce qui concerne les organisations régionales d'intégration économique visées au paragraphe 1 du présent article, aucun État membre d'une telle organisation ne peut signer la présente convention à moins qu'il ne représente un territoire situé en dehors du champ d'application territorial du traité établissant l'organisation et sous réserve que la participation de cet État membre soit limitée exclusivement à la représentation des intérêts de ce territoire. [...]

a) en signant, pendant la période visée à l'article XXVII, **paragraphe 1**, de la présente convention, un instrument rédigé à cette fin conformément à une résolution que la Commission doit adopter en vertu de la convention de 1949 [...]

[TJ_(CTT) ; 

Par ailleurs, toutes les séquences représentatives du C4f-11c sont constituées de trois maillons. Nous pouvons rapprocher ce type de composition des chaînes secondaires ou d'arrière-plan, potentiellement propres à des textes juridiques. L'analyse des facteurs de variations nous permettra de confirmer s'il existe un lien entre ce cluster et le genre textuel.

8.3.5 Bilan

Pour résumer, nous avons vu que l'analyse des séquences a permis de distinguer onze types de chaînes de référence, révélant des stratégies d'introduction et de maintien variées. Nous avons pu distinguer principalement des chaînes homogènes et des chaînes plus hétérogènes (C3a-11c, C3b-11c, C4a-11c et C4b-11c, par exemple). Nous avons également pu mettre en évidence la présence de CR secondaires ou d'arrière-plan (C2a-11c), généralement peu étudiées dans la littérature. Ces CR seraient des CR peu structurantes. Nous avons également fait l'hypothèse que ce type de composition serait propre à des textes juridiques. Nous avons également identifié des chaînes éphémères ou locales (C4a-11c, C4b-11c, C4d-11c et C4f-11c).

Nous avons évoqué à plusieurs reprises des hypothèses quant à l'impact du genre textuel ou de la nature des référents sur la composition des CR. Dans la section suivante, nous allons analyser les facteurs de variation afin de compléter notre description des clusters.

8.4 Lien entre les clusters et les facteurs de variation

Dans cette section, nous discutons du lien entre les facteurs de variation et les clusters obtenus. Les facteurs de variation retenus pour l'analyse DEMRC sont la nature du référent, le type et le genre textuel. Nous discutons des résultats obtenus avec la classification en quatre et en onze clusters.

8.4.1 Impact de la nature du référent sur la composition des chaînes

Concernant la nature du référent, nous avons tout d'abord analysé l'aspect concret vs abstrait et ensuite le caractère humain vs non humain, avec les distinctions plus fines entre humain collectif et humain individuel (cf. 6.1).

Le test du χ^2 montre qu'il existe une liaison significative entre les clusters obtenus et la nature du référent, aussi bien entre les référents concrets et abstraits qu'entre les référents humains et non humains.

8.4.1.1 Référents concrets et abstraits

Le test est significatif aussi bien pour la classification en quatre clusters que pour celle en onze clusters. Le test de Cramer qui, nous le rappelons, permet d'évaluer la force de la relation, est de 0,22 pour la classification en quatre clusters et de 0,31 pour la classification en onze, ce qui indique une association modérée.

Pour l'analyse des résidus de Pearson, nous nous concentrons sur la classification en onze clusters. Nous ne présentons que les associations significatives.

Dans la [figure 8.18](#)⁴, nous pouvons observer que le C1-11c, cluster des CR composées principalement de Np, présente une attraction avec les référents concrets. Cela peut s'expliquer par le fait que ce type de composition est utilisé principalement pour des référents humains, ou, comme nous l'avons vu, pour des toponymes. Les clus-



FIGURE 8.18 – Corrplot : résidus de Pearson standardisés

ters C2a-11c et C2b-11c, composés principalement de SN_∅, présentent une attraction lorsque les référents sont abstraits. Cela pourrait conforter notre hypothèse sur les CR d'arrière plan, où les référents abstraits occupent la fonction syntaxique de complément de nom, fonction moins saillante que celle de sujet. Nous retrouvons ici des résultats qui vont dans la même direction que ceux de [Oberlé et al. \(2018\)](#) pour les articles IMRAD.

Nous observons également une attraction entre le C3a-11c et les référents abstraits. Les CR composées exclusivement de SN_{def} et d'au moins 5 maillons auraient donc tendance à être utilisées pour référer à des entités abstraites. Le C4a-11c présente une répulsion pour les référents abstraits alors que le C4b-11c présente une attraction avec ce type de référents. Cela pourrait indiquer une tendance à introduire les référents de nature différente par des expressions de nature différente, SN_{ind} pour les référents concrets et SN_{def} pour les abstraits.

Les clusters C4d-11c et C4e-11c entretiennent une relation d'attraction avec les référents concrets et le C4f-11c une attraction avec les référents abstraits. Nous avons vu que ce dernier cluster regroupe les CR ayant une composition présentant la catégorie "autre". Nous avons également vu que la catégorie autre couvre des SN_{num} ou des numéros. Nous avons vu que la force et la relation est modérée, ce qui indiquerait l'influence d'autres facteurs sur les différents types de composition des CR.

4. Pour des raisons de lisibilité du graphique, nous avons choisi une orientation horizontale ; cela ne change rien à la manière de lire le graphique.

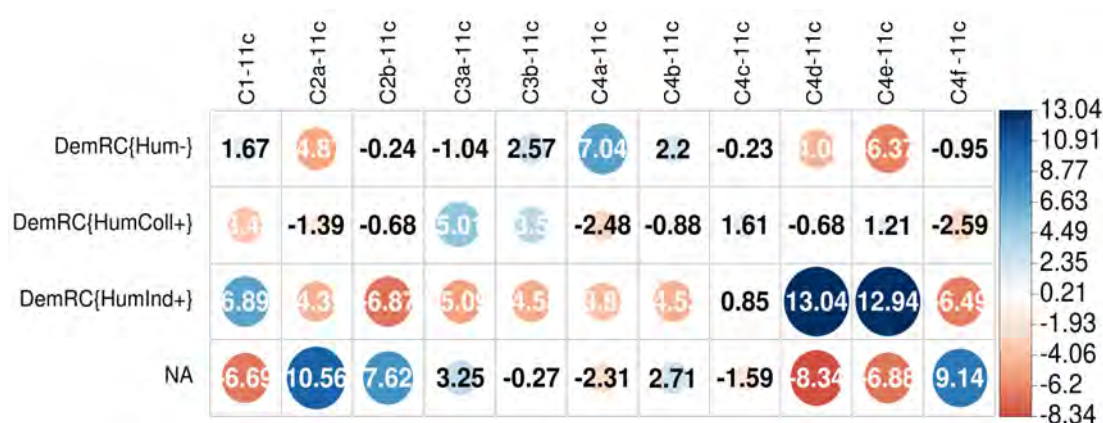


FIGURE 8.19 – Corrplot 11c et type de référent : humain vs non humain

8.4.1.2 Référents humains et non humains

Nous avons pris en compte la distinction entre référents humains individuels, humains collectifs et non humains.

Le test du χ^2 pour la classification en quatre clusters montre qu'il existe une liaison entre les quatre clusters et le caractère humain et non humain ($\chi^2 = 320,27$; $df = 9$; p -value $< 2,2e-16$). Cependant, nous avons obtenu un score de Cramer de 0,16, ce qui indique une faible force de cette relation. Lorsque l'on considère la classification plus fine en onze clusters, nous obtenons un score significatif ($\chi^2 = 811,2$; $df = 30$; p -value $< 2,2e-16$), confirmant qu'il existe une liaison entre le type de référent et le type de composition des CR. Le score du test de Cramer est de 0,25, ce qui indique une relation modérée.

L'analyse des résidus de Pearson (figure 8.19) montre qu'il existe une attraction significative entre :

- les référents humains individuels et les clusters C1-11c (Np++), C4d-11c (pro-poss) et C4e-11c (pro++).

Pour les référents humains individuels, il existe donc une préférence pour les noms propres aussi bien pour l'introduction que le maintien du référent (C1-11c). De même, l'utilisation d'expressions de haute accessibilité, tels les possessifs et les pronoms, est préférée lorsque les référents sont des humains individuels. Cela pourrait également venir conforter l'hypothèse que ce type de composition signale des référents importants (utilisation des Np) et qui restent saillants et accessibles.

- les référents humains collectifs et les clusters C3a-11c (SN_{def} ++) et C3b-11c (SN_{def} -pro);

Cette attraction indique que les référents humains collectifs sont de préférence introduits par le biais de SN_{def} : cette tendance à préférer les SN_{def} aux SN_{ind} pourrait s'expliquer par le fait que les référents collectifs, telles les organisations, ont un certain degré de familiarité. Par la suite, le maintien du référent est assuré soit par d'autres SN_{def}, tendance que nous avons commencé à observer dans AnnoDis, soit par le biais de pronoms.

- les référents non humains et les clusters C3b-11c (SN_{def} -hétérogènes), C4a-11c (SN_{ind} -pro) et C4b-11c (SN_{def} -pro);

Ces observations montrent que les référents non humains sont de préférence introduits pas des SN_{def} et des SN_{ind} . Nous avons déjà observé cette tendance lors de l'analyse des CR AnnoDis. Le maintien est assuré soit par des pronoms, soit par d'autres expressions tels les pronoms ou d'autres SN_{def} . Nous pouvons également observer que les CR des référents humains peuvent présenter une composition plus hétérogène (C3b-11c) par rapport aux référents humains.

Les référents NA (correspondant aux référents abstraits) confirment les observations que nous avons proposées en 8.4.1.1.

8.4.2 Impact du type de texte sur la composition des chaînes

Le test du χ^2 montre qu'il existe une liaison significative entre le type de texte et les clusters obtenus. Cette liaison existe aussi bien pour la classification en quatre clusters qu'en onze clusters : respectivement $\chi^2 = 451,45$; $df = 3$; $p\text{-value} < 2,2e-16$ et $\chi^2 = 599,11$; $df = 10$; $p\text{-value} < 2,2e-16$. Le test de Cramer nous permet de situer la force de cette liaison comme modérée (0,33 et 0,38 respectivement).

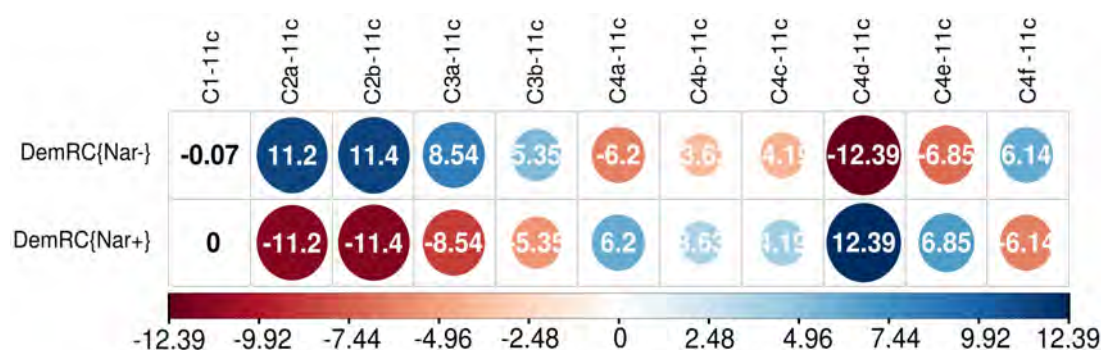


FIGURE 8.20 – Corrplot 11c et type de texte : $DemRC_{\{Nar-\}}$ vs $DemRC_{\{Nar+\}}$

En observant les résidus de Pearson (figure 8.20), nous pouvons remarquer que le type de texte nous permet de distinguer des groupes de types de chaînes utilisés : le premier groupe (C2a-11c, C2b-11c, C3a-11c et C3b-11c) présente une attraction avec les textes non narratifs. Les C2a-11c et C2b-11c sont caractérisés par la présence de $SN_{\emptyset}$. Cela pourrait montrer que dans les textes narratifs, on introduit difficilement des référents en utilisant des SN sans déterminant, hormis dans les titres de chapitres, s'il y en a. Par ailleurs, il y a sans doute un impact des genres textuels inclus dans les textes non narratifs : notamment avec les articles de presse et les textes juridiques (cf. 8.3.2). C3a-11c et C3b-11c regroupent des CR composées principalement de SN_{def} . Ce type de composition était très fréquent dans la ressource AnnoDis, dont les textes sont non narratifs. L'autre groupe, plus varié, regroupe tous les clusters obtenus à partir de la division du C4-4c. Ceux-ci présentent une attraction pour les textes narratifs - à l'exception du C4f-11c, qui présente une répulsion avec les textes narratif - où nous observons différentes stratégies : introduction avec des SN_{ind} ou des SN_{def} et maintien via

des pronoms ; introduction et maintien via des pronoms ou des possessifs. Cette attraction ne nous apparaît pas étonnante, notamment du fait de la présence de pronoms. Par ailleurs, nous avons vu que les C4d-11c et C4e-11c, composés principalement de pronoms, présentent une attraction avec les référents humains individuels. Cela pourrait indiquer que, dans les textes narratifs, les référents instanciés sont principalement des humains, ce qui est souvent le cas. L'utilisation des expressions de haute accessibilité indique que le référent est accessible, et potentiellement central.

8.4.3 Impact du genre textuel sur la composition des chaînes

Le test du χ^2 montre qu'il existe une liaison significative entre les clusters obtenus et le genre textuel. La force de la relation est modérée.

L'analyse des résidus de Pearson montre une attraction entre le C1 et les articles de presse, les articles encyclopédiques et la biographie (figure 8.21). La présence des CR composées principalement par des Np dans les articles de presse (ou biographie) va dans le sens des observations faites par [Schneidecker \(2005\)](#), qui a analysé les CR dans des portraits journalistiques. Nous observons également une attraction entre le C2-4c et les textes juridiques, ce qui vient confirmer nos observations sur la présence de CR propres à ces textes, dont les référents sont introduits et maintenus par des SN₀. Ce cluster a une répulsion avec les articles de presse, la biographie, les nouvelles et les romans. Le C3-4c, composé principalement de SN définis, a une attraction avec les articles de presse et les textes juridiques : des compositions similaires pour ces deux genres ont été observées par [Longo et Todirascu \(2014\)](#). Il existe une répulsion entre ce cluster et les biographies, les nouvelles, les pamphlets et les romans, alors que le C4-4c présente une attraction avec ces mêmes genres de textes. Nous discuterons ces derniers avec les résultats pour les onze clusters, puisque le C4-4c est très hétérogène comme nous l'avons vu.

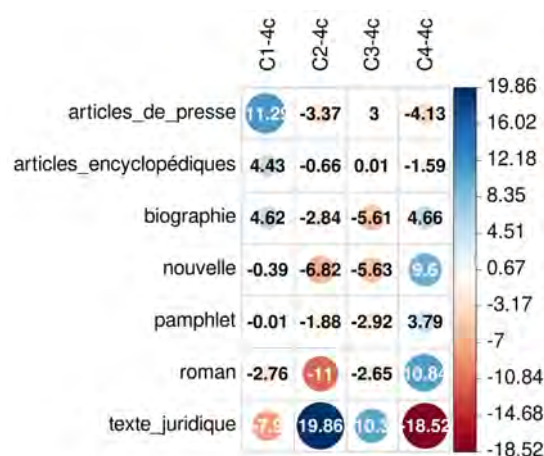


FIGURE 8.21 – Corrplot 4c et genres textuels

Les onze clusters Concernant les résultats pour la classification en onze clusters, nous avons effectué plusieurs tests, car certains effectifs attendus ne rejoignent pas le seuil de 5 nécessaire pour valider les résultats du test de χ^2 . Nous avons effectué notamment un test du χ^2 avec permutation, un test exact de Fisher, prévu pour remplacer

le test du χ^2 lorsque le seuil des effectifs attendus n'est pas rejoint. Nous avons également effectué le test en groupant les genres textuels pour lesquels la condition des effectifs attendus n'était pas satisfaite. Les tests effectués montrent tous qu'il existe une liaison significative entre les clusters et les genres textuels.

L'analyse des onze clusters nous permet d'observer des attractions et des répulsions plus précises pour les clusters divisés. Nous observons notamment que le cluster C2a-11c et le C2b-11c, confirment une attraction avec les textes juridiques, mais également avec les articles encyclopédiques. Cela pourrait s'expliquer par la présence de titres de sections pour les articles encyclopédiques, permettant l'utilisation de SN sans déterminant pour introduire les référents. Une analyse orientée sur les titres serait nécessaire pour confirmer cette hypothèse. Cependant, l'état actuel des annotations DEMRC ne permet pas de s'intéresser à cet aspect, les titres n'ayant pas été signalés explicitement en tant que tels. Le C3b-11c présente une relation d'attraction avec les articles de presse, alors que le C3a-11b, il tend vers une répulsion, sans que cette relation soit significative. Sans surprise, les deux clusters ont une attraction avec les textes juridiques. Nous observons également une attraction avec les textes encyclopédiques.

Le C4a-11c (SN_{ind}-Pro) a une attraction avec les romans et les pamphlets et le C4b-11c (SN_{def}-Pro) avec les biographies, les romans et les pamphlets. Ces relations montrent que, dans les biographies, l'introduction des référents se fait majoritairement par le SN défini. Cela pourrait s'expliquer par le fait que les biographies racontent l'histoire d'une personne dont l'identité est connue, ou présumée comme telle, par le lecteur dès le début du texte, les SN définis servant alors à décrire son occupation ou certaines de ses caractéristiques.

Le C4c-11c (Poss++) présente une attraction avec les nouvelles et les pamphlets, le C4d-11c (Poos-Pro), quant à lui, présente une attraction avec les biographies, les romans, les pamphlets, et les nouvelles. C4e-11c (Pro++) présente une attraction avec ces même genres de texte. Cela pourrait s'expliquer par le fait que ces genres instantient des personnages principaux, ce qui peut justifier l'utilisation d'expressions de haute accessibilité. Enfin, le C4f-11c présente une attraction avec les textes juridiques et les textes encyclopédiques.

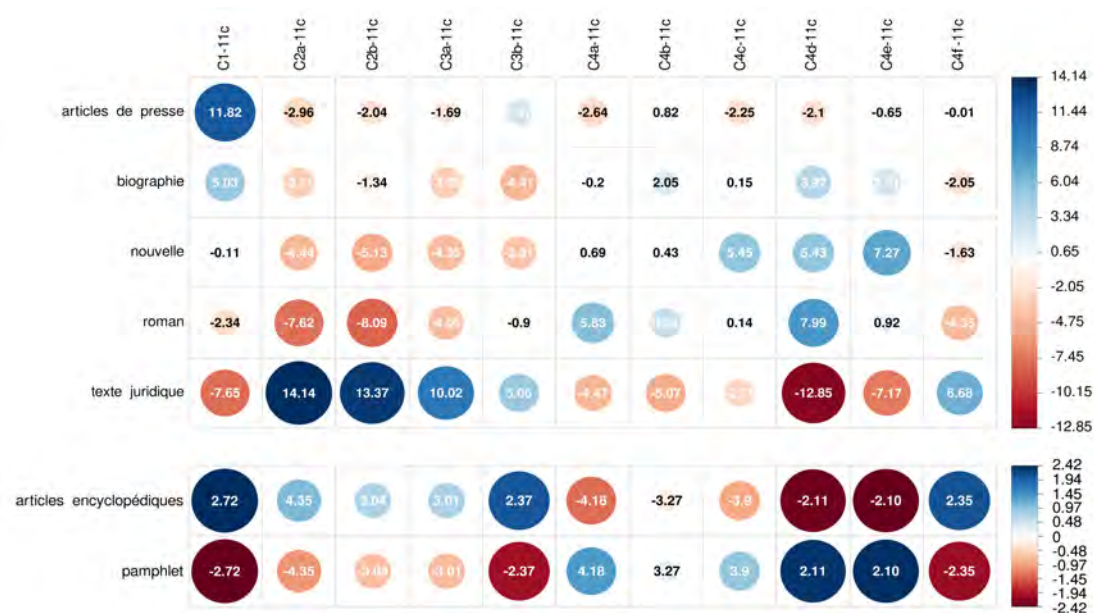


FIGURE 8.22 – Corrplot 11c et genres textuels

8.5 Bilan

Les résultats présentés dans ce chapitre étendent l'analyse des séquences, précédemment appliquée au corpus AnnoDis ([chapitre 7](#)), à la ressource DEMRC. Cette application avait un double objectif : d'une part, tester la productivité des patrons identifiés dans AnnoDis sur un corpus différent ; d'autre part, approfondir la compréhension des stratégies d'introduction et de maintien des référents compte tenu de la plus grande diversité de types et de genres textuels.

Pour atteindre notre premier objectif, dans un premier temps, nous avons projeté les patrons AnnoDis sur DEMRC. Cette projection a montré que malgré une couverture globale plus faible (19% des CR dans DEMRC contre 38,9% dans AnnoDis), plusieurs patrons se retrouvent dans les deux ressources. Les différences observées s'expliquent par la diversité des modèles d'annotation et par la diversité des textes présents dans DEMRC. Nous avons également mis en évidence que certains patrons (notamment ceux liés aux expressions de faible accessibilité) sont plus fréquents dans les textes non narratifs, tandis que des patrons impliquant des noms propres et des pronoms apparaissent davantage dans les textes narratifs.

Concernant notre deuxième objectif, nous avons procédé à l'analyse des séquences sur la ressource DEMRC. Afin de limiter l'impact des variations extrêmes de longueur sur les mesures de similarité, nous avons effectué une coupure à sept maillons (cf. [6.2.1](#)). Comme pour AnnoDis, l'analyse repose sur la nature des maillons. La classification à gros grain des séquences a permis d'identifier quatre grands clusters (4c) ; ceux-ci ont ensuite été affinés par une classification à petit grain (11c). Dans ce qui suit, nous résumons les principales caractéristiques des clusters obtenus :

- Cluster C1-4c (Np et Pro) : regroupe des CR dont le premier maillon est presque

systématiquement un nom propre (Np), suivi par des alternances avec des pronoms ou, moins fréquemment, par des SN_{def}. Bien que ce cluster ne soit pas divisé lors de l'analyse à petit grain deux tendances se dégagent :

- des CR homogènes constituées de Np ;
- des CR homogènes mais où le Np alterne avec des pronoms, indiquant une activation locale forte du référent.
- Cluster C2-4c (SN_Ø et SN_{def}) : ce cluster se caractérise par une prédominance de SN_Ø en position initiale (73,5% des cas) et une alternance avec des SN_{def} dans les maillons suivants. Lors de l'analyse à petit grain, cette classe est subdivisée en deux (C2a-11c et C2b-11c) :
 - C2a-11c : des CR homogènes, presque exclusivement composées de SN_Ø (65% des CR du sous-cluster) ;
 - C2b-11c : des CR présentant une alternance entre SN_{def} et SN_Ø.
- Cluster C3-4c (SN_{def} ++) : regroupe des CR principalement constituées de SN_{def}. La classification en onze révèle deux tendances :
 - C3a-11c : les CR sont généralement plus longues et homogènes. Elles commencent presque toujours par un SN_{def} et le maintien se fait par d'autres SN_{def} ;
 - C3b-11c : les CR sont plus courtes (trois ou quatre maillons) et présentent une variabilité accrue dès le premier maillon, avec l'apparition marginale de possessifs et d'autres catégories.
- Cluster C4-4c (CR hétérogènes) : regroupe un nombre important de chaînes et se distingue par une grande hétérogénéité, caractérisée par une multiple alternance entre SN_{def}, SN_{ind} et pronoms. Cette diversité impose de recourir à une analyse à petit grain pour identifier des tendances plus fines.
 - C4a-11c et C4b-11c : ces deux sous-groupes se distinguent par la nature du premier maillon. Dans C4a-11c, le premier maillon est un SN_{ind}, tandis que dans C4b-11c, il s'agit d'un SN_{def}.
 - C4c-11c : ces CR sont homogènes et composées presque exclusivement de possessifs.
 - C4d-11c : regroupe des CR présentant une plus grande hétérogénéité, avec une variation marquée de la nature du premier maillon. Le maintien se fait principalement par une alternance de pronoms avec d'autres catégories, telles que SN_{dem}, possessifs, SN_{def} ou SN_{ind} ;
 - C4e-11c : ces CR, globalement longues (au moins cinq maillons), sont homogènes et composées presque exclusivement de pronoms.

Nous avons ensuite étudié l'impact de différents facteurs sur la composition des chaînes. L'analyse statistique montre que la composition des chaînes varie significativement en fonction de la nature du référent (humain versus non humain, concret versus abstrait) et du genre textuel (articles encyclopédiques, rapports de géopolitique, etc.).

Discussion

Les résultats obtenus dans les deux ressources mettent en évidence des similarités et des différences. D'une part, les patrons identifiés sur AnnoDis se retrouvent,

en partie, dans DEMRC, confirmant la validité et la robustesse de l'approche adoptée. D'autre part, la diversité textuelle et la granularité de l'annotation dans DEMRC permettent de dégager de nouveaux types de composition. Cette richesse se manifeste particulièrement dans les clusters très hétérogènes, comme C4-4c.

L'analyse des séquences des deux corpus démontre que la composition des CR est influencée par la nature du référent, par le type de texte et par le genre textuel. Alors que les CR d'AnnoDis tendent à présenter des patrons relativement clairs et homogènes, ceux de DEMRC révèlent une diversité accrue, tant au niveau de la composition que de la stratégie de maintien du référent. L'analyse à différents niveaux de granularité – une première classification en quatre clusters, suivie d'une subdivision en onze sous-clusters – permet de saisir finement ces variations internes, en identifiant des tendances diversifiées, telles que la reprise directe avec des expressions de faible accessibilité versus l'alternance avec des expressions de haute accessibilité.

Par exemple, la présence de pronoms constitue un indicateur d'accessibilité élevée : elle suggère que la distance entre les maillons est faible et qu'aucun autre référent en compétition ne vient perturber la continuité référentielle, minimisant ainsi le risque d'ambiguïté. Cependant, ce lien ne s'observe pas systématiquement pour tous les types de référents. En effet, nous avons constaté que les chaînes composées exclusivement de noms propres (Np) peuvent correspondre à des référents non humains, notamment les toponymes. Contrairement aux référents humains, où la distance entre les mentions joue un rôle déterminant dans le choix des formes référentielles, cette variable semble peu pertinente pour les toponymes. Dans ces cas, le Np est utilisé pour l'introduction et pour le maintien du référent, indépendamment de la distance entre les mentions. Toutefois, lorsque deux occurrences d'un même référent sont très proches dans le texte, la répétition du Np tend à être évitée au profit d'expressions de moindre accessibilité, comme les SN_{def}. Les CR issues des clusters C2a-11c et C2b-11c constituent un autre exemple de type de composition spécifique selon la nature du référent et le genre textuel. L'introduction du référent est faite par des SN_∅, mais pour le maintien, l'analyse fine de ces CR met en évidence deux tendances : d'un côté, des chaînes homogènes où les SN_∅ sont utilisés systématiquement, aussi bien pour l'introduction que pour le maintien du référent ; de l'autre côté, des CR où l'introduction alterne entre SN_∅ et SN_{def}, le maintien étant assuré par une alternance entre ces deux catégories. Ces compositions sont particulièrement fréquentes lorsque les référents désignent des entités abstraites.

Cette analyse des séquences, fondée sur la nature des maillons, peut être enrichie par l'intégration de la fonction syntaxique des expressions référentielles, du type de reprise, ainsi que des caractéristiques globales des chaînes référentielles, telles que leur longueur et la distance entre les maillons. Cela permettrait d'élaborer une typologie robuste des CR.

8.6 Les patrons des chaînes DEMRC

Dans cette section, nous présentons les patrons des chaînes DEMRC les plus fréquents. Le cas échéant, nous indiquons si le patrons existait déjà pour la ressource

AnnoDis. Pour conclure, nous ouvrons la discussion avec des hypothèses quant aux types de chaînes que ces patrons peuvent représenter.

Les patrons les plus fréquents

Rappelons que pour proposer des patrons des CR nous exploitons le format SPS disponible dans TraMineR (cf. 5.3). Comme attendu, le nombre de séquences proposées par le système est très élevé : 2 070 séquences sont proposées. Après un premier regroupement automatique consistant à fusionner toutes les séquences dont la nature grammaticale est identique, nous obtenons 1 209 séquences. Malgré cette première étape, le nombre de séquences reste trop élevé pour proposer des patrons illustrant des types des CR. Afin de généraliser les patrons, nous avons procédé à un regroupement ultérieur, pour lequel nous avons fixé un seuil au-dessous duquel nous ne prenons pas en compte les séquences proposées. Ce seuil est établi arbitrairement à 10 occurrences sur l'ensemble du corpus. Une fois ce seuil appliqué, nous avons procédé à un regroupement manuel visant à proposer des patrons plus génériques. Cette opération repose sur la fusion des séquences partageant une même structure de base, même si la nature précise des maillons peut varier. Lorsque ces variations ne traduisent pas un degré d'accessibilité différent, nous avons regroupé les séquences dans un même patron — en signalant les différentes possibilités à l'aide d'une notation inspirée des expressions régulières, par exemple `ProlPoss1 :3`, qui indique que l'on peut retrouver soit une succession de pronoms (min. 1, max. 3) soit une succession de possessifs (min. 1, max. 3). En revanche, lorsque la variation de la nature des maillons traduit un degré d'accessibilité différent, nous avons fait le choix de garder deux patrons distincts. La variation du degré d'accessibilité peut être révélatrice de différentes stratégies, ce qui justifie de les maintenir distinctes.

Le [tableau 8.3](#) présente les patrons que nous avons identifiés, classés par ordre de fréquence sur l'ensemble des chaînes analysées. Les pourcentages indiqués sont calculés par rapport au nombre total de CR. Pour des raisons de lisibilité, ce tableau montre uniquement la distribution globale des patrons. Leur répartition en fonction des différents facteurs de variation est illustrée à l'aide de visualisations. Un tableau détaillé récapitulant l'ensemble des patrons et la distribution en fonction de tous les facteurs de variation analysés est disponible en [annexe G](#).

Les patrons identifiés couvrent 57,5% (soit 2 434 CR sur 4235) de l'ensemble des CR de la ressource DemRC. Globalement, les patrons les plus fréquents correspondent aux tendances observées pour les clusters : des introductions différentes (e.g. P3 avec des SN_{def} et P4 avec des SN_{ind}), des CR homogènes, notamment pour les patrons qui ne présentent aucune variation sur la nature de maillons, comme c'est le cas pour P1, P6, P10, P15, P20, P21 et P30. Nous retrouvons également des patrons qui reflètent une composition plus hétérogène, alternant des expressions de faible et de haute accessibilité (e.g. P9).

Le nombre de patrons DEMRC étant élevé (#36), nous ne les présentons pas tous en détail. Nous nous focalisons sur les patrons les plus fréquents, en faisant des rapprochements entre les patrons qui présentent des caractéristiques communes. Ensuite,

Patrons	Occurrences	
	#	%
P1 : $SN_{def} > SN_{def}\{2:6\}$	382	9,02
P2 : $SN_{def} (> SN_{def}\{1:5\}) > Poss\{1:5\} Pro\{1:6\}$	258	6,09
P3 : $SN_{def} (> SN_{def}\{1:3\}) > Poss Pro\{1:4\} > SN_{def}\{1:4\} (> Pro\{1:4\} Poss\{1:2\})$	256	6,04
P4 : $SN_{ind} (> SN_{ind}\{1:1\}) > Pro Poss\{1:6\} (> Poss\{1:3\} \emptyset\{1:2\})$	166	3,92
P5 : $SN_{def} (> SN_{def}\{1:2\}) > Poss \emptyset\{1:2\} Pro\{1:5\} > Poss\{1:5\} Pro\{1:3\} \emptyset\{1:2\}$	124	2,93
P6 : $SN_{\emptyset} > SN_{\emptyset}\{2:6\}$	106	2,50
P7 : $SN_{\emptyset} > (> SN_{\emptyset}\{1:4\}) > SN_{def}\{1:6\}$	92	2,17
P8 : $Poss (> Poss\{1:5\}) > Pro\{1:6\} (> Poss\{1:4\})$	84	1,98
P9 : $SN_{ind} (> SN_{ind}\{1:1\}) > Pro\{1:5\} > SN_{def}\{1:4\} SN_{dem}\{1:2\} (> Pro\{1:4\})$	80	1,89
P10 : $Pro > Pro\{2:6\}$	76	1,79
P11 : $SN_{ind} (> SN_{ind}\{1:1\}) > SN_{def}\{1:6\} (> Pro\{1:3\})$	69	1,63
P12 : $Pro (> Pro\{1:5\}) > Poss\{1:3\} (> Pro\{1:5\})$	54	1,28
P13 : $Pro (> Pro\{1:3\} > SN_{ind}\{1:2\} SN_{def}\{1:5\} (> Pro\{1:5\})$	54	1,28
P14 : $SN_{def} (> SN_{def}\{1:4\}) > SN_{\emptyset}\{1:4\} > SN_{def}\{1:5\} (> SN_{\emptyset}\{1:2\})$	51	1,20
P15 : $Np > Np\{2:6\}$	50	1,18
P16 : $Poss (> Poss\{1:5\}) > SN_{def}\{1:6\} (> Poss\{1:3\})$	49	1,16
P17 : $SN_{def} (> SN_{def}\{1:5\}) > SN_{\emptyset}\{1:6\}$	46	1,09
P18 : $SN_{def} (> SN_{def}\{1:5\}) > SN_{dem}\{1:3\} (> Pro\{1:5\})$	46	1,09
P19 : $SN_{\emptyset} > (> SN_{\emptyset}\{1:3\}) > SN_{def}\{1:5\} > SN_{\emptyset}\{1:4\} (> SN_{def}\{1:3\})$	38	0,90
P20 : $Poss > Poss\{2:6\}$	37	0,87
P21 : $Autre > Autre\{2:6\}$	28	0,66
P22 : $Np (> Np\{1:5\}) > SN_{def}\{1:5\} (> Np\{1:4\})$	28	0,66
P23 : $SN_{def} (> SN_{def}\{1:5\}) > SN_{ind}\{1:2\} (> SN_{def}\{1:5\})$	26	0,61
P24 : $Np (> Np\{1:5\}) > Pro\{1:4\} (> Np\{1:5\})$	26	0,61
P25 : $SN_{ind} > Pro\{1:3\} SN_{dem}\{1:2\} > SN_{ind}\{1:2\} SN_{def}\{1:5\}$	23	0,54
P26 : $SN_{dem} (> SN_{dem}\{1:2\}) > Pro\{1:6\}$	23	0,54
P27 : $SN_{def} > Pro\{1:4\} > Poss\{1:3\} > Pro\{1:2\}$	21	0,50
P28 : $SN_{def} (> SN_{def}\{1:4\}) > SN_{dem}\{1:2\} > SN_{def}\{1:5\}$	21	0,50
P29 : $Autre (> Autre\{1:1\}) > SN_{def}\{1:5\}$	21	0,50
P30 : $SN_{ind} > SN_{ind}\{2:6\}$	19	0,45
P31 : $SN_{def} > SN_{def}\{1:1\} > Pro\{1:5\} > SN_{dem}\{1:1\}$	15	0,35
P32 : $SN_{ind} (> SN_{ind}\{1:1\}) > SN_{dem}\{1:1\} > Pro\{1:3\}$	15	0,35
P33 : $SN_{def} (> SN_{def}\{1:1\}) > Pro\{1:2\} > SN_{def}\{1:3\} > Pro\{1:3\} > SN_{def}\{1:3\}$	14	0,33
P34 : $Autre (> Autre\{1:1\}) > Pro\{1:6\}$	14	0,33
P35 : $SN_{def} (> SN_{def}\{1:2\}) > Pro\{1:2\} > Poss\{1:3\} > SN_{def}\{1:2\}$	11	0,26
P36 : $SN_{\emptyset} (> SN_{\emptyset}\{1:1\}) > Pro\{1:3\}$	11	0,26
Sous-total	2434	57,47
Total	4235	100,00

TABLEAU 8.3 – Les patrons les plus fréquents dans DEMRC

nous nous concentrons sur leur analyse en fonction des critères de variation. Comme nous l'avions déjà remarqué lors de la projection des patrons AnnoDis (cf. 8.1), nous observons que le patron le plus fréquent (P1) correspond à une succession de SN_{def} ,

correspondant à des chaînes homogènes que nous avons retrouvé principalement dans les clusters C3a-11c et C3b-11c. Nous avons vu que ce type de composition est utilisé principalement dans les textes non narratifs, notamment dans les textes juridiques et encyclopédiques, pour désigner, de préférence, des référents abstraits ou des humains collectifs. La succession d'expressions de faible accessibilité tout le long de la CR pourrait indiquer une distance importante entre les maillons, la présence de référents en compétition ou une construction typique d'un type ou d'un genre de texte particulier.

Les patrons P2 et P5, sont très proches, la différence consiste en la présence d'une transition supplémentaire dans P5. Ces deux patrons incluent deux stratégies : une introduction effectuée par le biais de SN_{def} et une reprise directe par des pronoms, ce qui traduit une haute accessibilité du référent, une distance entre les maillons potentiellement faible ou l'absence de référents en compétition. Nous retrouvons la même tendance pour le P4, à la différence près que l'introduction est faite par le biais de SN_{ind} . Ces deux patrons reflètent ainsi les tendances que nous avons observées pour les clusters C4a-11c (figure 8.12_a) et C4b-11c (figure 8.12_c). La deuxième stratégie identifiée par P1 consiste à utiliser un ou plusieurs SN_{def} après son introduction et avant d'enchaîner sur des expressions de haute accessibilité. La même stratégie est observée pour P4, mais contrairement à P1, nous observons que le passage d'une expression de faible accessibilité aux expressions de haute accessibilité se fait plus rapidement : un seul SN_{ind} entre l'introduction et les autres expressions est possible. Cette tendance dévie de ce qui est généralement décrit dans la littérature, où les SN_{ind} sont considérés comme des expressions ne pouvant occuper que la première position dans la chaîne (cf. 1.1.1). Nous pourrions alors nous attendre à ce que le premier SN_{ind} corresponde à un titre, mais cela n'est pas toujours le cas, comme nous le montrons dans (52). Une analyse plus approfondie serait nécessaire pour quantifier ces cas et mieux comprendre cette tendance.

- (52) Dieulouard
 Houcine, le finaliste de la « Star Academy », vient de rendre **une petite visite d'amitié et de courtoisie** aux gendarmes de Dieulouard.
Une visite **qui** n'est pas passée inaperçue au sein de la brigade et **qui** a permis aux enfants de militaires d'ajouter un autographe à leurs collections.
 [AP_[ES] ; ↻]

La présence de plusieurs SN_{ind} est d'autant plus étonnante lorsqu'on considère le P30. Cependant, on peut remarquer que ce patron est peu fréquent. Par ailleurs, en observant la distribution de ce patron en fonction du genre textuel, nous remarquons que ce type de composition se retrouve principalement dans des textes juridiques et traduit principalement un emploi générique, comme illustré dans (53). Dans cet exemple, nous pouvons également observer que la distance entre les maillons est bien plus importante que celle observée dans (52).

- (53) [...] Article R15-33-1 La commission prévue à l'article 28-1 dont l'avis conforme est requis pour la désignation des agents des douanes des catégories A et B habilités à effectuer **des enquêtes judiciaires** sur réquisition du procureur de

la République ou sur commission rogatoire du juge d'instruction est composée comme suit : 1 ° Le procureur général près la Cour de cassation ou son délégué choisi par lui parmi le premier avocat général ou les avocats généraux à la Cour de cassation, président ; 2 ° Deux magistrats du ministère public dont un au plus peut être magistrat honoraire ; 3 ° Le directeur général des douanes et droits indirects ou son représentant ; 4 ° Le fonctionnaire chargé de la sous-direction du personnel et du budget de la direction générale des douanes et droits indirects ou son représentant ; 5 ° Le fonctionnaire chargé de la sous-direction des affaires juridiques, contentieuses et de la lutte contre la fraude de la direction générale des douanes et droits indirects ou son représentant ; 6 ° Un fonctionnaire de l'administration des douanes et droits indirects ayant au moins le grade d'inspecteur principal ; 7 ° Le magistrat délégué aux missions judiciaires de la douane, mentionné au VII de l'article 28-1, ou son représentant.

[8 phrases]

Article R15-33-5 Les agents des douanes chargés d'effectuer **des enquêtes judiciaires** en application de l'article 28-1 sont désignés parmi ceux qui ont été reçus à l'examen technique, par arrêté conjoint du garde des sceaux, ministre de la justice, et du ministre chargé des douanes, sur proposition du directeur général des douanes et droits indirects et après avis conforme de la commission mentionnée à l'article R. 15-33-1.

[28 phrases]

A l'occasion d' **une enquête judiciaire** ou de l'exécution d'une commission rogatoire, ils ne peuvent solliciter ou recevoir des ordres ou instructions que de l'autorité judiciaire qui les a requis. [TJ_{CPP} ; ⌘]

Parmi les patrons, nous pouvons distinguer des motifs qui se répètent. Prenons pas exemple P3. Ce patron se caractérise par une alternance d'expressions de faible accessibilité (les SN_{def}) et des expressions de haute accessibilité, tels les pronoms, les possessifs ou les sujets zéro. Cette alternance SN_{def} > Pro pourrait continuer, comme le montre le P33, où nous observons qu'une transition supplémentaire Pro > SN_{def} est présente. Nous pouvons faire l'hypothèse que cet enchaînement pourrait encore continuer bien au-delà des sept premiers maillons.

P6 et P7 reflètent les types de compositions que nous avons retrouvés dans les clusters C2a-11c (100% des CR couvertes par P6) et C2b-11c (75% de CR du cluster couvert par P7) respectivement. P6 est spécifique au cluster C2a-11c : nous ne retrouvons pas ce patron dans les autres clusters. Rappelons que les chaînes présentes dans ce cluster semblent être des chaînes d'arrière-plan et peu structurantes. P6 pourrait ainsi permettre d'identifier des CR d'arrière-plan, non structurantes, et qui caractérisent notamment les textes non narratifs, en particulier les textes juridiques : 90% des occurrences de ce patron se trouvent dans les textes juridiques.

Passons maintenant à l'observation de la distribution des patrons selon différents facteurs :

- le type de texte, narratif ou non narratif ;

- le genre textuel ;
- la nature des référents, abstraits ou concrets ;
- l'animacité des référents, humain individuel, collectif ou non humain.

Les patrons et les types de textes La figure 8.23 montre la distribution des patrons selon le type textuel. Pour chaque patron, nous signalons si la différence entre les deux types de textes est significative (* à ***) ou non (ns). Nous avons utilisé un test binomial, utilisé pour déterminer si la proportion observée d'un certain résultat dans un échantillon diffère significativement d'une proportion attendue. Dans notre cas, l'hypothèse nulle est que la distribution des patrons est équivalente dans les deux types de textes. Si $p < 0,05$, alors cela signifie qu'il existe une préférence significative pour l'un des deux types de texte. Nous observons que P1, P6, P7, P14 présentent une

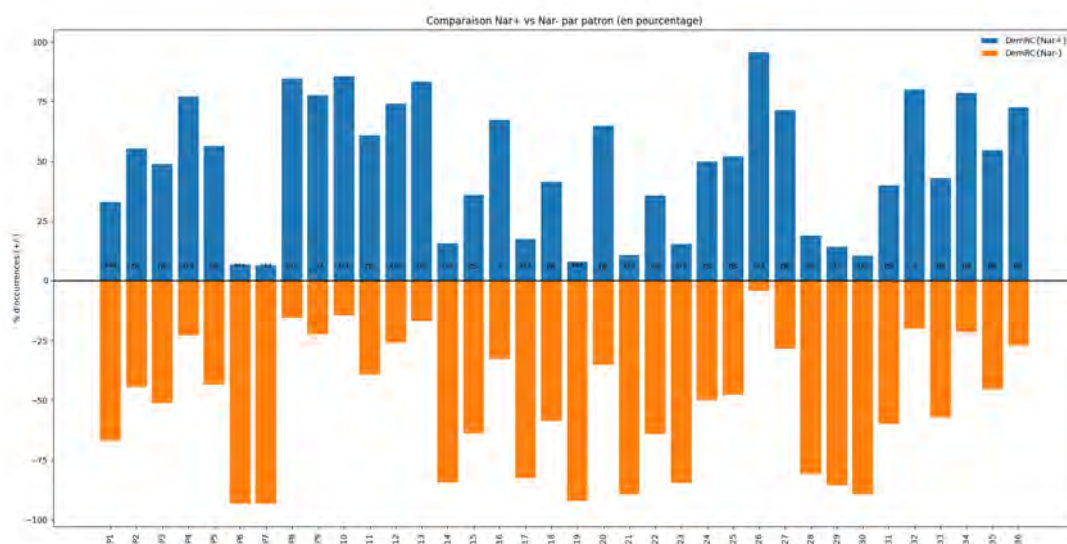


FIGURE 8.23 – Distribution des patrons selon le type de texte

préférence significative pour les textes non narratifs. Ainsi, nous pouvons conclure qu'une succession de SN_{def} ou $SN_{\emptyset}$, ou une alternance des deux sont plus utilisés dans les textes non narratifs.

Concernant le genre textuel, nous avons effectué un test du χ^2 d'ajustement, utilisé pour déterminer si la distribution observée d'une variable catégorielle à plusieurs modalités diffère significativement d'une distribution théorique attendue. Les résultats montrent que les différences entre les patrons et les différents genres textuels ne sont pas significatives.

Les patrons et la nature du référent Dans la figure 8.24 nous pouvons observer la distribution des patrons selon le degré de concrétude des référents. Pour analyser si la différence entre les référents abstraits et concrets est significative, nous avons utilisé le test binomial, comme pour le type de texte. Nous pouvons observer que les cinq premiers patrons présentent une différence significative, avec une tendance à être plus

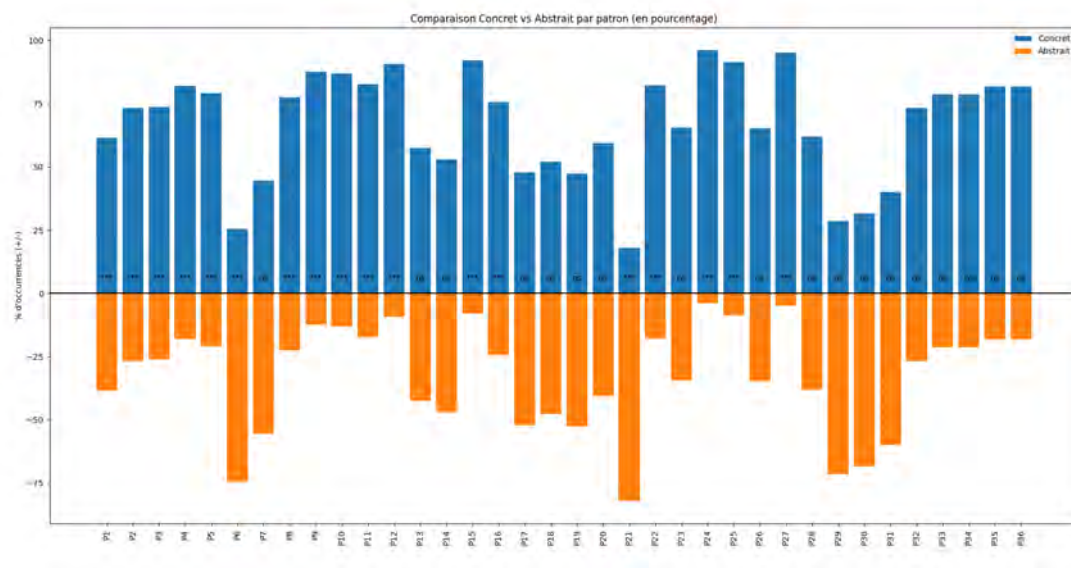


FIGURE 8.24 – Distribution des patrons selon le degré de concretude des référents

utilisés pour référer à des entités concrètes. Il en va de même pour P8, P9, P10, P11 et P12. Au contraire, P6 présente une préférence pour les entités abstraites.

Concernant la différence entre référent humain, non humain et collectif, il ne semble pas exister de différence significative entre les patrons et la nature du référent. Pour des raisons de lisibilité, dans la [figure 8.25](#), nous illustrons la distribution des 18 premiers patrons.

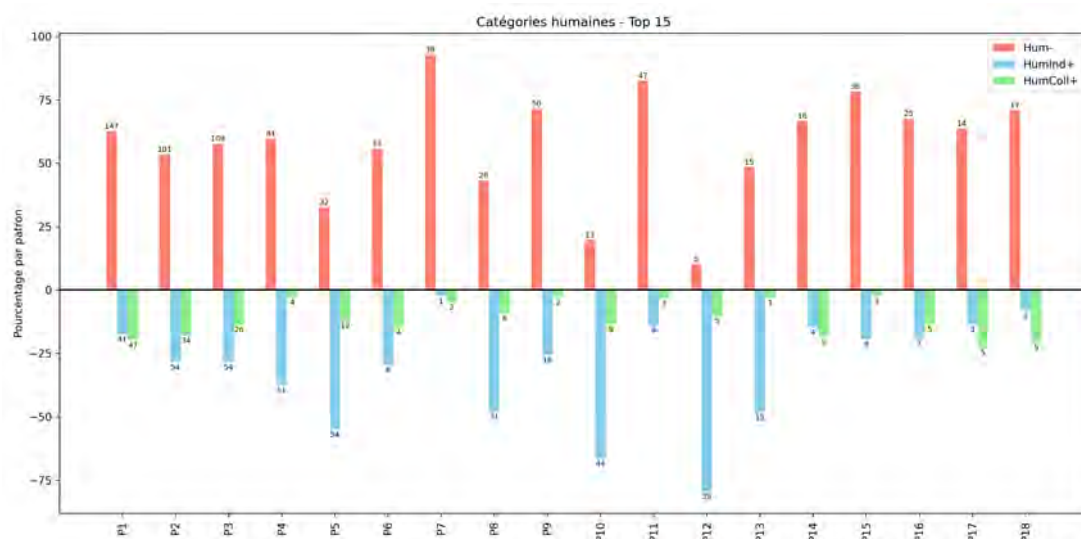


FIGURE 8.25 – Distribution des patrons selon l'animacité des référents

8.7 Bilan

L'analyse des CR dans la ressource DEMRC nous a permis de mettre en évidence des stratégies d'introduction et de maintien du référent qui complètent les analyses proposées à partir de AnnoDis. Les patrons sont distribués de manière différenciée en fonction de la nature du référent, du type de texte (narratif ou non narratif) et du genre textuel (biographie, texte juridique, article encyclopédique, etc.). Leur interprétation repose sur leur lien avec les clusters des séquences, qui révèle en même temps des structures homogènes et des formes plus hétérogènes. Le patron P1 ($SN_{\text{def}} > SN_{\text{def}} \{2:6\}$) est le plus fréquent du corpus, couvrant à lui seul plus de 9 % des chaînes observées. Il correspond à des CR homogènes, exclusivement composées de SN définis. Ce patron se retrouve principalement dans les clusters C3a-11c et C3b-11c, et il est caractéristique des textes non narratifs, en particulier des textes juridiques et encyclopédiques. Ce même patron avait déjà été identifié dans la ressource AnnoDis. Le même type de composition dans des textes non narratifs, notamment des articles de presse et des textes juridiques, a été observé par [Longo et Todirascu \(2014\)](#). La même composition a été également observée dans des données orales, notamment par [Antoine et al. \(2016\)](#). Cela montre que des CR homogènes sont présentes dans les textes écrits, mais également à l'oral.

Les patrons P2 et P5 combinent un SN défini initial à des formes plus accessibles de type pronom ou possessif. Ces patrons traduisent une stratégie de passage rapide vers une référence saillante, avec un référent déjà bien ancré dans le discours. Ils sont associés au cluster C4b-11c, typique des textes narratifs, notamment des biographies, où le référent central – généralement un humain individuel – est connu du lecteur et reste accessible tout au long du récit.

Les patrons P6 et P7, composés de SN sans déterminant, seuls ou alternés avec des SN définis, sont particulièrement caractéristiques des textes juridiques, où la structure syntaxique des CR est souvent figée et fortement contrainte. Ils correspondent aux clusters C2a-11c et C2b-11c et sont utilisés pour désigner des référents abstraits, fréquemment en position de complément du nom. Ces patrons identifieraient des chaînes d'arrière-plan, qui ne contribuent pas ou peu à la création de liens de cohésion.

Enfin, nous pouvons observer la présence de patrons constitués uniquement d'expressions de haute accessibilité, tels les pronoms et les possessifs (P10, P12, P20). Ces patrons indiquent des chaînes homogènes et propres aux textes narratifs. Nous pouvons également supposer qu'il s'agit de référents centraux, qui restent actifs. Pour vérifier cette hypothèse, il serait intéressant de croiser les patrons avec les caractéristiques globales des CR pour analyser la longueur, la distance entre les maillons et la couverture de la CR.

Pour conclure, les patrons les plus homogènes, tels que P1, P6 ou P10, apparaissent dans des contextes discursifs bien définis – respectivement les textes non narratifs avec référents abstraits, les textes juridiques, et les textes narratifs centrés sur des personnages humains. À l'inverse, les patrons plus mixtes (P2, P4, P5) traduisent des dynamiques d'activation plus complexes, dans lesquelles l'accessibilité du référent varie

au fil du texte. Leur distribution montre que les genres textuels, le type de texte et la nature du référent ont un impact sur la composition des chaînes de référence.

L'analyse des patrons permet ainsi de poser les premiers jalons d'une typologie des chaînes référentielles, en identifiant des configurations récurrentes selon différents facteurs de variation. Toutefois, pour élaborer une typologie plus fine et opérationnelle, il serait nécessaire de croiser ces patrons avec les caractéristiques globales des CR, telles que leur longueur, leur couverture ou encore la distance entre les maillons. Ces dimensions quantitatives sont en effet, des indices précieux de la saillance ou du rôle discursif du référent. Elles pourraient notamment permettre de distinguer les chaînes principales et globales, structurant l'ensemble du texte, des chaînes d'arrière-plan et locales, plus limitées en portée.

Conclusions et perspectives

Dans cette thèse, nous avons proposé une étude des chaînes de référence pour mieux en comprendre le fonctionnement textuel, à partir de l'analyse de deux corpus diversifiés : AnnoDis et Democrat. En nous positionnant dans une perspective fonctionnelle, nous nous sommes intéressée notamment aux stratégies d'introduction et de maintien des référents. L'objectif principal de notre étude était de rendre compte des régularités et des différences qui caractérisent les CR pour dégager des patrons, c'est-à-dire des types de composition similaires (Schnedecker, 2021).

Notre point de départ repose sur plusieurs constats. Tout d'abord, nous avons relevé l'absence d'études systématiques et à large échelle sur les CR en français écrit. Pendant longtemps, cette situation a été liée au manque de corpus de grande taille et annoté de manière exhaustive en chaînes de référence pour le français écrit. La mise à disposition du corpus Democrat (Landragin, Poibeau, & Victorri, 2016) a marqué une avancée importante, ouvrant la voie à de nouvelles investigations sur les CR en français écrit, dans une perspective à la fois descriptive et computationnelle.

Nous avons également souligné que, face à la complexité du phénomène et à la diversité des aspects à prendre en considération, les travaux existants se concentrent généralement sur un nombre restreint de chaînes ou d'aspects précis. Ces travaux ont permis d'identifier des critères de description pertinents, en montrant que la structure des CR varie selon plusieurs facteurs : le canal, la langue, la période, le type et le genre textuels, la nature du référent. Cependant, rares sont les études qui articulent ces facteurs à partir d'un corpus suffisamment large et varié. Dans ce travail, nous nous sommes focalisée sur trois facteurs : le type, le genre textuel et la nature du référent. Pour pouvoir analyser la nature du référent, un enrichissement des données a été nécessaire : nous avons proposé une annotation semi-automatique pour AnnoDis et une annotation manuelle pour Democrat (chapitre 6).

Un autre constat central concerne la difficulté à combiner plusieurs corpus annotés en raison de l'hétérogénéité des modèles d'annotation adoptés. En effet, les annotations dépendent du cadre théorique mobilisé et des objectifs de chaque étude. De plus, les corpus ne recouvrent pas toujours les mêmes phénomènes ou les mêmes types de référents. Face à ce constat, nous avons fait le choix d'exploiter deux corpus annotés, AnnoDis et Democrat, malgré leur hétérogénéité. Plutôt que de percevoir cette diversité comme un obstacle, nous avons cherché à en faire un levier pour mieux appréhender les différentes facettes des CR. Cela nous a conduit à interroger l'influence des modèles d'annotation sur la description des chaînes. Nous avons montré que le

modèle d'annotation peut être considéré comme un facteur de variation à part entière ([chapitre 3](#)). Nous avons notamment étudié l'impact du modèle d'annotation sur les caractéristiques globales des CR dans les deux ressources ([chapitre 4](#)). Contrairement à nos attentes initiales, les résultats montrent que le modèle d'annotation n'a pas d'effet significatif sur la longueur des chaînes. En revanche, un effet plus marqué est observé sur la distance entre les maillons. Il serait nécessaire d'élargir l'analyse en considérant davantage de critères de description des CR, mais ces premiers résultats montrent qu'il est essentiel, tant pour la constitution que pour la comparaison de ressources, d'anticiper et de documenter ces effets. Plus largement, il convient de garder à l'esprit que les phénomènes linguistiques observés dans les corpus annotés sont en partie tributaires du modèle d'annotation adopté, ce qui invite à une interprétation nuancée des résultats.

Enfin, concernant la description des chaînes de référence, nous avons constaté que les études existantes adoptent principalement une approche globale, décrivant les CR à travers des caractéristiques telles que leur longueur ou la proportion des différentes catégories grammaticales, en considérant les CR comme des ensembles d'expressions équivalentes. Ces caractéristiques permettent d'avoir une vision générale de la dimension et de la composition des CR. Bien que certains critères, tels la distance, permettent de rendre compte des effets de linéarité, nous en avons souligné les limites dès lors qu'il s'agit de décrire les stratégies d'introduction et de maintien du référent. Nous avons ainsi proposé une approche complémentaire, fondée sur la linéarité des chaînes et l'enchaînement des maillons.

Dans cette perspective, nous avons proposé une méthode outillée qui permet une description systématique et à grande échelle des enchaînements des maillons. Grâce à cette description, il a été possible de rendre compte des régularités et des différences qui caractérisent les CR et d'en dégager des patrons. L'un des principaux apports de ce travail est donc méthodologique. Afin de mettre en œuvre l'approche linéaire, nous avons proposé une méthode outillée fondée sur l'analyse des séquences, généralement utilisée en Sciences sociales ([chapitre 5](#)). Cette méthode permet de représenter chaque chaîne comme une séquence d'états (chaque état correspondant à un maillon caractérisé selon un critère donné, ici sa nature grammaticale). Cette formalisation favorise la comparaison des chaînes entre elles, leur regroupement automatique selon leur composition en appliquant des méthodes de clustering, et l'identification des patrons récurrents.

L'analyse de séquences proposée, combinée avec le clustering, est une analyse exploratoire qui permet de surmonter les limites de l'approche globale et de mettre en évidence comment les maillons s'enchaînent. Cela permet de rendre compte des différentes stratégies mises en œuvre pour l'introduction et le maintien des référents, dépassant ainsi la simple description de la catégorie la plus fréquente en position de premier maillon. Ces stratégies ne sont pas visibles avec une approche purement globale. En formalisant chaque chaîne comme une séquence d'états, nous rendons possible une analyse à large échelle, fiable et répliquable, ouvrant la possibilité de l'appliquer à d'autres données, quel que soit le modèle d'annotation, le type de textes ou la langue des textes.

Un autre avantage de l'analyse des séquences est qu'elle est modulaire : on peut

affiner les états ou changer le critère de description ou intégrer des variables pour tester l'influence des facteurs de variation.

L'analyse des séquences appliquée aux CR dans les ressources AnnoDis et DEMRC a permis de dégager des régularités et identifier plusieurs types de composition des CR, que nous avons ensuite confrontés aux facteurs de variation : le type textuel, le genre discursif et la nature du référent. Rappelons que l'analyse a été effectuée en prenant en considération les sept premiers maillons des CR. Ainsi, la longueur des séquences analysées varie entre 3 (2 pour AnnoDis) et 7. Les stratégies que nous avons mis en évidence, peuvent, dans ce sens, être considérées comme des stratégies propres aux premiers maillons de la CR. Néanmoins, nous avons montré que la plupart des CR présentes dans les corpus ne dépassent pas le 7 maillons et les CR plus longues sont très rares, ce qui signifie que même en considérant uniquement les 7 premiers maillons, les stratégies observées devraient ne pas être biaisées.

Nous avons formalisé les stratégies pour l'introduction et le maintien des référents par des patrons en les mettant en correspondance avec des types potentiels de CR. Les résultats ainsi obtenus constituent une bonne base de départ pour proposer à terme une typologie des CR. Nous proposons de distinguer quatre profils de CR :

- A ;
- B ;
- C ;
- D.

A : Les chaînes homogènes de faible accessibilité correspondent principalement à trois patrons :

- $SN_{def} > SN_{def}\{1:6\}$: identifié dans les deux ressources et très fréquent.
- $Np > Np\{1:6\}$: identifié dans les deux ressources mais peu fréquent dans AnnoDis.
- $SN_{\emptyset} > SN_{\emptyset}\{2:6\}$: identifié dans DEMRC.

L'introduction et le maintien par des SN_{def} ont déjà été observés dans d'autres travaux, notamment ceux de [Longo et Todirascu \(2014\)](#), sur des textes non narratifs. Nos observations vont dans le même sens. Ce même type de composition a également été observé à l'oral, par [Antoine et al. \(2016\)](#). L'enchaînement des Np a été mis en avant dans les travaux de [Schneidecker \(1997\)](#), pour les CR à référents humains. Dans nos résultats, nous avons pu observer que ce type de composition est également présent lorsqu'il s'agit de toponymes, que ce soit comme lieux ou comme référents humains collectifs. Cela permet de mettre en évidence que ce type de composition à lui seul ne permet pas de distinguer avec certitude les CR à référent humain des autres. Par ailleurs, nous avons pu mettre en évidence un type de CR qui n'est pas souvent analysé, illustré par le troisième patron. Nous avons montré que ce type de CR semble être typique des textes juridiques et qu'il semble correspondre à des CR d'arrière-plan, ne disposant pas de pouvoir structurant. En effet, nous avons vu que les maillons correspondent souvent à des compléments du nom, fonction syntaxique qui indique une faible saillance. Une analyse croisée avec la fonction syntaxique des maillons confirmerait cette hypothèse. Nous avons également vu que le patron comportant uniquement les SN sans déterminant est davantage mobilisé pour référer à des entités abstraites.

B : Les chaînes homogènes de haute accessibilité semblent concerner un référent central et qui reste accessible. Ce type de composition est formalisé principalement par trois patrons, identifiés dans DEMRC :

- $\text{Pro} > \text{Pro}\{2 : 6\} \text{SN}_{\text{def}} > \text{SN}_{\text{def}}\{1 : 6\}$;
- $\text{Poss} > \text{Poss}\{2 : 6\}$;
- $\text{Poss} (> \text{Poss}\{1 : 5\}) > \text{Pro}\{1 : 6\} (> \text{Poss}\{1 : 4\})$;

L'introduction du référent par des expressions comme les pronoms et les possessifs peut surprendre. Dans le [chapitre 8](#), nous avons expliqué que la présence de pronoms en première position dépend du modèle d'annotation Democrat. Nous avons montré que plusieurs cas correspondent à des pronoms pluriels, indiquant des cas de fusion. Cela indique que ces CR seraient liées à d'autres, à savoir celles des référents pris individuellement, en dehors du groupe. Concernant la présence des possessifs en premier maillon, nous avons montré que la grande majorité des cas correspondent à des SN_{poss} . L'introduction d'un référent par le biais de possessifs indique que le référent est dépendant d'un autre référent : nous pouvons donc supposer que les CR qui présentent cette composition sont des CR secondaires. Par ailleurs, nous avons vu que ce type de composition est préféré pour des référents concrets, notamment des humains individuels. Il est également utilisé de préférence dans des textes narratifs, notamment dans les romans, les nouvelles et les biographies, ce qui est cohérent avec le fait d'instancier des référents humains (cf. [8.4](#)).

C : Les chaînes homogènes, de faible et de haute accessibilité semblent concerner un référent central qui reste accessible. Ce type de composition est formalisé principalement par les patrons suivants :

- $\text{SN}_{\{\text{def}\}} (> \text{SN}_{\text{def}}\{1 : 5\}) > \text{Poss}\{1 : 5\} | \text{Pro}\{1 : 6\}$: identifié aussi bien dans AnnoDis que dans DEMRC ;
- $\text{SN}_{\text{ind}} (> \text{SN}_{\text{ind}}\{1 : 1\}) > \text{Pro} | \text{Poss}\{1 : 6\} (> \text{Poss}\{1 : 3\} | \emptyset\{1 : 2\})$: identifié dans DEMRC ;
- $\text{SN}_{\text{def}} > \text{Pro}\{1 : 4\} > \text{Poss}\{1 : 3\} > \text{Pro}\{1 : 2\}$: identifié dans DEMRC ;

Nous pouvons associer ce type de composition au patron proposé par [Givón \(1983\)](#) pour le topique continu. Nous avons vu, que ce patron est globalement utilisé pour référer à des humains, aussi bien dans AnnoDis que dans DEMRC, sans qu'une vraie préférence se dégage. Les trois patrons sont difficilement utilisés pour référer à des entités abstraites. Par ailleurs, il ne semble pas y avoir une préférence entre les textes narratifs et non narratifs.

D : Les chaînes plus hétérogènes sont formalisées par les patrons suivants :

- $\text{SN}_{\text{def}} (> \text{SN}_{\text{def}}\{1 : 3\}) > \text{Poss} | \text{Pro} | \emptyset\{1 : 4\} > \text{SN}_{\text{def}}\{1 : 4\} (> \text{Pro}\{1 : 4\} | \text{Poss}\{1 : 2\})$: identifié dans les deux ressources ;
- $\text{SN}_{\text{ind}} (> \text{SN}_{\text{ind}}\{1 : 1\}) > \text{Pro}\{1 : 5\} > \text{SN}_{\text{def}}\{1 : 4\} | \text{SN}_{\text{dem}}\{1 : 2\} (> \text{Pro}\{1 : 4\})$;
- $\text{SN}_{\text{ind}} (> \text{SN}_{\text{ind}}\{1 : 1\}) > \text{SN}_{\text{def}}\{1 : 6\} (> \text{Pro}\{1 : 3\})$;

Nous pouvons associer ce type de composition au patron proposé par [Givón \(1983\)](#) pour le topique discontinu. L'alternance entre formes de faible et de haute accessibilité pourrait, en effet, signaler que le référent a besoin d'être réinstancié et réactivé.

Une dernière contribution concerne l'enrichissement des données et leur mise à disposition. Afin de prendre en considération tous les facteurs de variation, garantir la comparabilité des résultats AnnoDis et Democrat, un travail d'harmonisation et d'enrichissement des données a été nécessaire (notamment par l'ajout d'informations sur la nature des référents). Nous avons mis à disposition la ressource DEMRC dans le format d'analyse exploité dans la thèse, avec les annotations harmonisées, actuellement disponible sur [Ortolang](#). Nous envisageons de mettre à disposition une version enrichie de la ressource avec les annotations sur la nature du référent. Par ailleurs, nous avons proposé une plateforme d'exploration des CR où il est possible de visualiser toutes les chaînes de la ressource DEMRC et d'effectuer des recherches selon plusieurs critères (disponible [ici](#)).

Ce travail comporte évidemment des limites. Nous pouvons notamment mettre en évidence le nombre restreint d'occurrences que les patrons permettent de couvrir en l'état actuel et le grand nombre de patrons proposés, notamment pour le corpus DEMRC. Nous avons pu mettre en évidence que certains patrons présentent des motifs qui se répètent, comme la transition $SN_{\text{def}} > \text{Pro}$. Une perspective intéressante pour pouvoir repérer ces motifs dans les chaînes et les exploiter pour mieux généraliser les patrons serait d'exploiter l'analyse des sous-séquences à l'aide de TraMineR. Cette analyse repose sur un algorithme pour l'extraction des motifs récurrents ([Masseglia et al., 2004](#)), permettant ainsi de repérer les sous-séquences qui composent les CR. Cela permettrait de proposer de meilleures généralisations des patrons et couvrir davantage de CR. Afin de généraliser ultérieurement les patrons, nous avons commencé à réfléchir à la possibilité de formaliser ces patrons sous forme d'expressions régulières. Cela permettrait, par exemple, de généraliser certaines régularités, comme les motifs qui se répètent, favorisant des regroupements ultérieurs, et surtout une meilleure exploitation des patrons.

Perspectives

Si ce travail a permis d'apporter une contribution à la description des CR, il met également en évidence l'ampleur des dimensions qu'il reste à explorer pour en saisir toute la complexité. Les perspectives que nous mentionnons ici visent à prolonger, approfondir ou réinvestir les axes explorés dans le cadre de la thèse, en tenant compte des limites identifiées et des questionnements restés ouverts.

Dans le prolongement immédiat de ce travail, il serait pertinent de croiser la description des classes obtenues et des patrons proposés avec les caractéristiques globales des CR, en considérant plusieurs critères à la fois. Cette analyse permettrait ainsi de compléter les observations fournies et de proposer une typologie des CR qui repose à la fois sur la dimension globale et la composition linéaire de ces structures. Par ailleurs, il serait intéressant de compléter la description de la dimension des CR en tenant compte de leur dimension séquentielle. Pour ce faire, une piste qui nous semble pertinente est d'intégrer des mesures telles que la complexité, la turbulence, l'entropie longitudinale et le nombre de transitions ([Gabadinho, Ritschard, & Studer, 2010](#)).

D'un point de vue méthodologique, l'approche linéaire adoptée pourrait être enrichie en intégrant d'autres dimensions de description. En effet, la méthode proposée

permet de considérer un seul critère à la fois. Nous nous sommes focalisée sur la nature des maillons. Nous avons également effectué l'analyse en considérant la fonction syntaxique et les types de reprise, mais les résultats obtenus ne permettent pas en l'état d'identifier de manière claire des types de composition. Or, comme nous l'avons vu, plusieurs aspects entrent en jeu simultanément, ayant un impact sur les caractéristiques des CR. Ainsi, une analyse multi-critère permettrait de mieux saisir les régularités complexes qui structurent les chaînes. En ce sens, l'exploitation de techniques d'analyse des séquences multidimensionnelle ([Ritschard, Liao, & Struffolino, 2023](#)) constitue une piste prometteuse, notamment pour considérer plusieurs dimensions.

Il serait également intéressant de prendre davantage en compte la structure textuelle. Nous avons évoqué la possibilité d'analyser les maillons situés en début de paragraphe ou dans des positions clés du discours (cadres, transitions), afin d'identifier des patrons caractéristiques de l'ouverture ou de la fermeture des chaînes. Il serait également pertinent de baliser les titres et les discours directs afin d'analyser plus finement certaines stratégies d'introduction et de maintien du référent. Cette perspective permettrait d'interroger le lien entre organisation textuelle et CR, notamment dans des genres où la structuration discursive est marquée. L'analyse de la cohabitation des chaînes dans les textes, ainsi que leur articulation à des relations de discours (par exemple, dans le cadre de la RST ([Mann & Thompson, 1987](#) ; [Taboada & Mann, 2006](#))), constitue également une piste pour mieux comprendre les liens entre les CR et la structuration des textes. Considérer les modalités de cohabitation permettrait également de mieux interpréter les patrons et identifier quand l'alternance entre expressions de faible et de haute accessibilité dépend de facteurs internes, telle la présence de référents en compétition et quand elle dépend de facteurs externes, comme le type ou le genre textuel. L'analyse des séquences multidimensionnelle pourrait, par exemple, permettre de traiter simultanément la nature des maillons et la présence d'autres référents. S'intéresser à la cohabitation pourrait également faciliter la description des stratégies mises en place lors d'un changement de référent, ce que notre méthode actuelle ne permet pas de faire.

Un autre prolongement naturel serait de mettre à l'épreuve la typologie élaborée ici sur d'autres corpus, notamment le corpus RésolCo ([Garcia-Debanc et al., 2017](#)), pour analyser les stratégies de continuité référentielle. RésolCo étant un corpus de textes d'apprenants, la description des stratégies mises en place pourrait avoir un potentiel de valorisation dans le domaine didactique. En tant qu'outil d'analyse, elle pourrait être mobilisée pour l'enseignement de la rédaction ou de la révision de textes où la maîtrise de la continuité référentielle constitue un enjeu central.

Il serait également intéressant d'appliquer l'analyse des séquences au corpus ANCOR ([Muzerelle et al., 2013](#)), ce qui offrirait l'occasion de comparer les stratégies de construction de la continuité référentielle entre l'écrit et l'oral. Cela permettrait également d'explorer des phénomènes peu accessibles dans les corpus écrits, comme la co-construction de la référence en interaction. Une telle comparaison entre modalités ouvrirait la voie à une réflexion sur la spécificité des patrons en fonction des contraintes propres à chaque mode de production. Il serait également pertinent d'exploiter des corpus des langues différentes (i.e. Scolinter ([Ponton et al., 2021](#)), CorefUD ([Nedoluzhko et al., 2022](#))) pour pouvoir mener une analyse contrastive des patrons et en étudier les spécificités.

Du point de vue applicatif, les résultats de cette thèse ouvrent des perspectives dans le domaine du traitement automatique des langues (TAL). Les patrons identifiés pourraient être mobilisés pour évaluer les performances des systèmes de résolution. En particulier, ils pourraient alimenter une typologie des erreurs de ces systèmes, en fonction de la complexité ou de la structure des chaînes. Il serait également intéressant d'évaluer les performances des systèmes en fonction des types de chaînes et d'examiner dans quelle mesure les modèles d'annotation influencent ces performances. Cela pourrait contribuer à une réflexion plus large sur les biais d'apprentissage induits par les corpus annotés.

Dans le prolongement de ce travail, bien d'autres perspectives restent à explorer pour appréhender toute la complexité des chaînes de référence. Nous espérons néanmoins avoir apporté notre pierre à l'édifice en soulignant l'importance d'envisager les chaînes de référence comme des structures discursives organisées par la linéarité du discours, et non plus seulement comme des ensembles d'expressions équivalentes.

Bibliographie

- Aassve, A., Billari, F. C., & Piccarreta, R. (2007, octobre). Strings of Adulthood : A Sequence Analysis of Young British Women's Work-Family Trajectories. *European Journal of Population / Revue européenne de Démographie*, 23(3), 369–388. doi: 10.1007/s10680-007-9134-6
- Abbott, A. (1983, octobre). Sequences of Social Events : Concepts and Methods for the Analysis of Order in Social Processes. *Historical Methods : A Journal of Quantitative and Interdisciplinary History*, 16(4), 129–147. doi: 10.1080/01615440.1983.10594107
- Abbott, A., & Barman, E. (1997, août). 2. Sequence Comparison Via alignment and Gibbs Sampling : A Formal Analysis of the Emergence of the Modern Sociological Article. *Sociological Methodology*, 27(1), 47–87. doi: 10.1111/1467-9531.271019
- Abbott, A., & Tsay, A. (2000, août). Sequence Analysis and Optimal Matching Methods in Sociology : Review and Prospect. *Sociological Methods & Research*, 29(1), 3–33. doi: 10.1177/0049124100029001001
- Achim, A. M., Fossard, M., Couture, S., & Achim, A. (2015, juin). Adjustment of speaker's referential expressions to an addressee's likely knowledge and link with theory of mind abilities. *Frontiers in Psychology*, 6. doi: 10.3389/fpsyg.2015.00823
- Ackerman, M., & Ben-David, S. (2016). A Characterization of Linkage-Based Hierarchical Clustering. , *Journal of Machine Learning Research*(17), 1–17.
- Aisenbrey, S., & Fasang, A. E. (2010, février). New Life for Old Ideas : The "Second Wave" of Sequence Analysis Bringing the "Course" Back Into the Life Course. *Sociological Methods & Research*, 38(3), 420–462. doi: 10.1177/0049124109357532
- Aktaş, B., & Stede, M. (2022, décembre). Anaphoric Distance in Oral and Written Language : Experimental Evidence. *Discours. Revue de linguistique, psycholinguistique et informatique. A journal of linguistics, psycholinguistics and computational linguistics*(31). doi: 10.4000/discours.12383
- Amoia, M., Kunz, K., & Lapshinova-Koltunski, E. (2012). Coreference in Spoken vs. Written Texts : A Corpus-based Analysis. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)* (pp. 158–164). Istanbul, Turkey : European Language Resources Association (ELRA).
- Angur Mahmud Jarman. (2020). *Hierarchical Cluster Analysis : Comparison of Single linkage, Complete linkage, Average linkage and Centroid Linkage Method*. Unpublished. doi: 10.13140/RG.2.2.11388.90240

- Anikina, T., Binder, A., Harbecke, D., Varanasi, S., Hennig, L., Ostermann, S., ... van Genabith, J. (2025, janvier). *Reverse Probing : Evaluating Knowledge Transfer via Finetuned Task Embeddings for Coreference Resolution* (N° arXiv :2501.19316). arXiv. doi: 10.48550/arXiv.2501.19316
- Anikina, T., Oguz, C., Skachkova, N., Tao, S., Upadhyaya, S., & Kruijff-Korabayova, I. (2021, novembre). Anaphora Resolution in Dialogue : Description of the DFKI-TalkingRobots System for the CODI-CRAC 2021 Shared-Task. In S. Khosla, R. Manuvinakurike, V. Ng, M. Poesio, M. Strube, & C. Rosé (Eds.), *Proceedings of the CODI-CRAC 2021 Shared Task on Anaphora, Bridging, and Discourse Deixis in Dialogue* (pp. 32–42). Punta Cana, Dominican Republic : Association for Computational Linguistics. doi: 10.18653/v1/2021.codi-sharedtask.3
- Antoine, J.-Y., Lefeuvre, A., & Schang, E. (2016). Codage en chaîne ou en première mention de la coréférence : approcher la structure des chaînes de référence par comparaison des deux annotations. *SHS Web of Conferences*, 27, 02001. doi: 10.1051/shsconf/20162702001
- Ariel, M. (1990). *Accessing noun-phrase antecedents*. London : Routledge.
- Ariel, M. (2001). Accessibility theory : An overview. In T. Sanders, J. Schilperoord, & W. Spooren (Eds.), *Human Cognitive Processing* (Vol. 8, pp. 29–87). Amsterdam : John Benjamins Publishing Company. doi: 10.1075/hcp.8.04ari
- Batool, F., & Hennig, C. (2021, juin). Clustering with the Average Silhouette Width. *Computational Statistics & Data Analysis*, 158, 107190. doi: 10.1016/j.csda.2021.107190
- Baumer, E. (2012, avril). Etude contrastive français / anglais des anaphores lexicales dans la presse et dans la fiction littéraire. *Corela*(HS-11). doi: 10.4000/corela.2233
- Baumer, E. (2019). *Reference Chains across Languages : A Cross-Linguistic study*. Beijing, China.
- Baumer, E., Dias, D., Gardelle, L., & Prak-Derrington, E. (2021). Quelles régularités pour les chaînes de référence dans le genre « nouvelles d’actualité » ? Exploration contrastive allemand / anglais / français :. *Travaux de linguistique*, n° 82(1), 17–43. doi: 10.3917/tl.082.0017
- Baumer, E., Dias, D., & Schnedecker, C. (2023). Approche contrastive exploratoire « corpus based » des séquences de clôture romanesque : les chaînes de référence comme marqueurs linguistiques au service de l’analyse littéraire. *Travaux de linguistique*, 86(1), 85–113. doi: 10.3917/tl.086.0085
- Bevacqua, L., Loáiciga, S., Rohde, H., & Hardmeier, C. (2021, décembre). Event and Entity Coreference Across Five Languages : Effects of Context and Referring Expression. *Dialogue & Discourse*, 12(2), 192–226. doi: 10.5210/dad.2021.207
- Biber, D., Connor, U., & Upton, T. A. (2007). *Discourse on the move : Using corpus analysis to describe discourse structure* (N° 28). Amsterdam : J. Benjamins.
- Blanchard, P. (2019). Sequence Analysis. In *Encyclopedia of Research Methods* (P.A. Atkinson, R.A. Williams and A. Cernat éd.). London : Sage.
- Blanchard, P., Bühlmann, F., & Gauthier, J.-A. (Eds.). (2014). *Advances in Sequence Analysis : Theory, Method, Applications* (Vol. 2). Cham : Springer International Publishing. doi: 10.1007/978-3-319-04969-4

- Bornstein, A., Cattani, A., & Dagan, I. (2020, octobre). *CoRefi : A Crowd Sourcing Suite for Coreference Annotation* (N° arXiv :2010.02588). arXiv. doi: 10.48550/arXiv.2010.02588
- Boudreau, S., & Kittredge, K. (2005). Résolution d'anaphores et identification des chaînes de coréférence : une approche « minimaliste ».
- Brzinsky-Fay, C. (2007, avril). Lost in Transition ? Labour Market Entry Sequences of School Leavers in Europe. *European Sociological Review*, 23(4), 409–422. doi: 10.1093/esr/jcm011
- Brzinsky-Fay, C., Kohler, U., & Luniak, M. (Eds.). (2006). Sequence analysis with Stata. *Stata Journal*. doi: 10.22004/ag.econ.118169
- Capin, D., Glikman, J., Schnedecker, C., & Todira, A. (2021, décembre). Le rôle des chaînes de référence dans la structuration textuelle : étude diachronique de l'ancien français au français moderne :. *Langages*, N° 224(4), 87–107. doi: 10.3917/lang.224.0087
- Chai, H., & Strube, M. (2022). Incorporating Centering Theory into Neural Coreference Resolution. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies* (pp. 2996–3002). Seattle, United States : Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.218
- Chamberlain, J., Poesio, M., & Kruschwitz, U. (2008, janvier). Phrase detectives : A web-based collaborative annotation game.
- Charolles, M. (1988). Les plans d'organisation textuelle : périodes, chaînes, portées et séquences. *Pratiques*, 57(1), 3–13. doi: 10.3406/prati.1988.1468
- Charolles, M. (2002). *La référence et les expressions référentielles en français*. Gap Paris : Ophrys.
- Chastain, C. (1975). Reference and Context. In *Language Mind and Knowledge* (K. Gunderson éd., pp. 194–269). Minneapolis : University of Minnesota Press.
- Cohen, J. (1960, avril). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1), 37–46. doi: 10.1177/001316446002000104
- Colléter, M., Fabre, C., Ho-Dac, L. M., Pery-Woodley, M.-P., Rebeyrolle, J., & Tanguy, L. (2012). *La ressource ANNODIS multi-échelle : Guide d'annotation et "bonus"* (Rapport technique N° 20).
- Corblin, F. (1987). *Indéfini, défini et démonstratif*. Genève : Droz.
- Corblin, F. (1995). *Les formes de reprise dans le discours. Anaphores et chaînes de référence*. Rennes : Presses universitaires de Rennes.
- Cornish, F. (1998). Les “chaînes topicales” : leur rôle dans la gestion et la structuration du discours. , *Cahiers de Grammaire* 23, 19–40.
- Cornish, F. (2021, juin). L'anaphore revisitée : Fonctionnements discursifs et interactionnels :. *Langue française*, N° 210(2), 9–20. doi: 10.3917/lf.210.0009
- Davis, F., & van Schijndel, M. (2020). Discourse structure interacts with reference but not syntax in neural language models. In *Proceedings of the 24th Conference on Computational Natural Language Learning* (pp. 396–407). Online : Association for Computational Linguistics. doi: 10.18653/v1/2020.conll-1.32
- Delaborde, M. (2020). *Analyse en corpus de chaînes de coréférence : la coréférence non-strictes à l'épreuve de la linguistique outillée* (Thèse de doctorat non

- publiée). Université de la Sorbonne nouvelle - Paris III.
- De Marneffe, M.-C., Recasens, M., & Potts, C. (2015, mars). Modeling the Lifespan of Discourse Entities with Application to Coreference Resolution. *Journal of Artificial Intelligence Research*, 52, 445–475. doi: 10.1613/jair.4565
- Doddington, G., Mitchell, A., Przybocki, M., Ramshaw, L., Strassel, S., & Weischedel, R. (2004, mai). The Automatic Content Extraction (ACE) Program – Tasks, Data, and Evaluation. In M. T. Lino, M. F. Xavier, F. Ferreira, R. Costa, & R. Silva (Eds.), *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*. Lisbon, Portugal : European Language Resources Association (ELRA).
- Elzinga, C. (2007). Sequence analysis : Metric representations of categorical time series.
- Elzinga, C. (2010, avril). Complexity of Categorical Time Series. *Sociological Methods & Research*, 38, 655–656. doi: 10.1177/0049124109357535
- Elzinga, C. H. (2003, août). Sequence Similarity : A Nonaligning Technique. *Sociological Methods & Research*, 32(1), 3–29. doi: 10.1177/0049124103253373
- Ezugwu, A. E., Ikotun, A. M., Oyelade, O. O., Abualigah, L., Agushaka, J. O., Eke, C. I., & Akinyelu, A. A. (2022, avril). A comprehensive survey of clustering algorithms : State-of-the-art machine learning applications, taxonomy, challenges, and future research prospects. *Engineering Applications of Artificial Intelligence*, 110, 104743. doi: 10.1016/j.engappai.2022.104743
- Federzoni, S. (2019). *Évaluation et exploitation de la Ressource ANNODIS pour la détection des chaînes de référence* (Mémoire de Master 2). Université de Toulouse - Jean Jaurès, Toulouse.
- Federzoni, S., Ho-Dac, L. M., & Fabre, C. (2023, février). A linear approach of chain composition. In *Reference : From conventions to pragmatics* (Vol. 228, pp. 107–126). Amsterdam : John Benjamins Publishing Company. doi: 10.1075/slcs.228
- Federzoni, S., Ho-Dac, L.-M., & Rebeyrolle, J. (2020). Les chaînes topicales dans la ressource ANNODIS. *SHS Web of Conferences*, 78. doi: 10.1051/shsconf/20207811005
- Fossard, M., Achim, A. M., Rousier-Vercruyssen, L., Gonzalez, S., Bureau, A., & Champagne-Lavau, M. (2018, février). Referential Choices in a Collaborative Storytelling Task : Discourse Stages and Referential Complexity Matter. *Frontiers in Psychology*, 9, 176. doi: 10.3389/fpsyg.2018.00176
- Gabadinho, A., Ritschard, G., Müller, N. S., & Studer, M. (2011). Analyzing and Visualizing State Sequences in R with **TraMineR**. *Journal of Statistical Software*, 40(4). doi: 10.18637/jss.v040.i04
- Gabadinho, A., Ritschard, G., & Studer, M. (2010). Indice de complexité pour le tri et la comparaison de séquences catégorielles.
- Gabadinho, A., Ritschard, G., Studer, M., & Muller, N. S. (2010). Mining sequence data in R with the TraMineR package : A user's guide.
- Garcia-Debanc, C., Ho-Dac, L.-M., Bras, M., & Rebeyrolle, J. (2017, janvier). Vers l'annotation discursive de textes d'élèves. *Corpus*(16). doi: 10.4000/corpus.2783

- Gardelle, L., & Vincent-Durroux, L. (2022, janvier). Anaphore et pronoms en anglais : Convergences, différences et complémentarité de quelques approches linguistiques. *Corela*(HS-35). doi: 10.4000/corela.14439
- Gardent, C., & Manuélian, H. (2005). Création d'un corpus annoté pour le traitement des descriptions définies. *Revue TAL : traitement automatique des langues*, 46(1), 115.
- Ghaddar, A., & Langlais, P. (2016, mai). WikiCoref : An English Coreference-annotated Corpus of Wikipedia Articles. In N. Calzolari et al. (Eds.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)* (pp. 136–142). Portorož, Slovenia : European Language Resources Association (ELRA).
- Givón, T. (Ed.). (1983). *Topic continuity in discourse : A quantitative cross-language study* (N° 3). Amsterdam : Benjamins.
- Glikman, J., Guillot-Barbance, C., & Obry, V. (2014, septembre). Les chaînes de référence dans un corpus de textes narratifs médiévaux : Traits généraux et facteurs de variation :. *Langages*, N° 195(3), 43–60. doi: 10.3917/lang.195.0043
- Grishman, R., & Sundheim, B. (1996). Message Understanding Conference- 6 : A Brief History.
- Grobol, L. (2020). *Coreference resolution for spoken French* (Thèse de doctorat non publiée).
- Grosz, B. J., Joshi, A. K., & Weinstein, S. (1995). Centering : A Framework for Modeling the Local Coherence of Discourse. *Computational Linguistics*, 21(2), 203–225.
- Guillot-Barbance, C., & Quignard, M. (2019, décembre). Chaînes de référence et structure textuelle dans les Essais sur la peinture de Diderot. *Discours*(25). doi: 10.4000/discours.10421
- Gundel, E. J., & Abbott, B. (2019). *The Oxford Handbook of Reference*.
- Gundel, J. K., Hegarty, M., & Borthen, K. (2003). Cognitive Status, Information Structure, and Pronominal Reference to Clausally Introduced Entities.
- Guo, C. (2021). Chaînes de référence dans les textes encyclopédiques : étude contrastive en français et en chinois :. *Travaux de linguistique*, n° 82(1), 129–152. doi: 10.3917/tl.082.0129
- Halliday, M. A. (1985). *An introduction to Functional Grammar*. London : Edward Arnold.
- Halliday, M. A., & Hasan, R. (1976). *Cohesion in English* (N° 9). London New York : Routledge, Taylor & Francis Group.
- Hamming, R. W. (1950). Error-Detecting and Error-Correcting Codes. , *Bell System Technical Journal*(26), 147–160.
- Heiden, S., Magué, J.-P., & Pincemin, B. (2010). TXM : Une plateforme logicielle open-source pour la textométrie - conception et développement. In *10th International Conference on the Statistical Analysis of Textual Data* (pp. 1021–1032). Rome, Italie.
- Held, W., Iter, D., & Jurafsky, D. (2021). Focus on what matters : Applying Discourse Coherence Theory to Cross Document Coreference. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (pp. 1406–1417). doi: 10.18653/v1/2021.emnlp-main.106

- Hirschman, L., & Chinchor, N. (1998). Appendix F : MUC-7 Coreference Task Definition (version 3.0). In *Seventh Message Understanding Conference (MUC-7) : Proceedings of a Conference Held in Fairfax, Virginia, April 29 - May 1, 1998*.
- Ho-Dac, L.-M. (2007). *La position initiale dans l'organisation du discours :
une exploration en corpus* (Thèse de doctorat non publiée). Université Toulouse le Mirail - Toulouse II.
- Ho-Dac, L.-M., & Péry-Woodley, M.-P. (2014). Annotation des structures discursives : L'expérience ANNODIS. *SHS Web of Conferences*, 8, 2647–2661. doi: 10.1051/shsconf/20140801286
- Ho-Dac, L.-M., Péry-Woodley, M.-P., & Tanguy, L. (2010, juillet). Anatomie des structures énumératives. In P. Langlais & M. Gagnon (Eds.), *Actes de la 17e conférence sur le Traitement Automatique des Langues Naturelles. Articles longs* (pp. 151–160). Montréal, Canada : ATALA.
- Hou, J., & Landragin, F. (2020). Conceptions lexicale et cognitive de la notion d'antécédent : une étude contrastive de l'anaphore pronominale en français et en chinois. *SHS Web of Conferences*, 78, 12011. doi: 10.1051/shsconf/20207812011
- Hovy, E., Marcus, M., Palmer, M., Ramshaw, L., & Weischedel, R. (2006, juin). OntoNotes : The 90% Solution. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume : Short Papers* (pp. 57–60). New York City, USA : Association for Computational Linguistics.
- Kaufman, L., & Rousseeuw, P. J. (2009). *Finding Groups in Data : An Introduction to Cluster Analysis*. John Wiley & Sons.
- Kibrik, A. A. (2011). *Reference in discourse*. Oxford New York : Oxford university press.
- Kleiber, G. (1994). *Anaphores et pronoms*. Louvain-la-Neuve : Duculot.
- Kleiber, G., & Riegel, M. (2003). Les pronoms évoluent-ils avec les référents ? In *La syntaxe raisonnée* (pp. 117–149). De Boeck Supérieur. doi: 10.3917/dbu.berre.2003.01.0117
- Kunz, K. (2010). *Variation in English and German nominal coreference : A study of political essays* (N° 21). Frankfurt am Main Berlin Bern Wien : Lang.
- Kunz, K., & Lapshinova-Koltunski, E. (2015, mars). Cross-linguistic analysis of discourse variation across registers. *Nordic Journal of English Studies*, 14(1), 258–288. doi: 10.35360/njes.347
- Kunz, K., Lapshinova-Koltunski, E., & Martínez, J. M. (2016). Beyond Identity Coreference : Contrasting Indicators of Textual Coherence in English and German. In *Proceedings of the Workshop on Coreference Resolution Beyond OntoNotes (CORBON 2016)* (pp. 23–31). Ann Arbor, Michigan : Association for Computational Linguistics. doi: 10.18653/v1/W16-0704
- Kunz, K., & Steiner, E. (2012, décembre). Towards a comparison of cohesive reference in English and German. *Linguistics and the Human Sciences*, 6(1-3), 219–251. doi: 10.1558/lhs.v6i1-3.219
- Landragin, F. (2011, novembre). Une procédure d'analyse et d'annotation des chaînes de coréférence dans des textes écrits. *Corpus*(10), 61–80. doi: 10.4000/corpus.2010
- Landragin, F. (2012). La saillance : questions méthodologiques autour d'une notion multifactorielle. *Faits de Langues*, 39(1), 15–31. doi: 10.1163/19589514-039

- 01-9000000003
- Landragin, F. (2014). Anaphores et coréférences : analyse assistée par ordinateur. Peter Lang.
- Landragin, F. (2016). Description, modélisation et détection automatique des chaînes de référence (DEMOCRAT). , 11–15.
- Landragin, F. (2017, septembre). Analyse, visualisation et identification automatique des chaînes de coréférences : Des questions interdépendantes ? : *Langue française*, N° 195(3), 17–34. doi: 10.3917/lf.195.0017
- Landragin, F. (2021a). Méthodologie pour la préparation d’une campagne d’annotation manuelle d’expressions référentielles. In *Des corpus numériques à l’analyse linguistique en langues de spécialité* (pp. 37–60). doi: 10.4000/books.ugaeditions.24215
- Landragin, F. (Ed.). (2021b). *Un corpus annoté en chaînes de référence et son exploitation : le projet DEMOCRAT* (N° 224 (décembre 2021)). Malakoff : Armand Colin.
- Landragin, F., Glikman, J., Schnedecker, C., & Todirascu, A. (2024, janvier). Chaînes de référence dans le corpus Democrat : une analyse en diachronie longue. *Corpus*(25). doi: 10.4000/corpus.8581
- Landragin, F., Poibeau, T., & Victorri, B. (2016, juillet). Analyse des références et des transitions référentielles : l’apport de la linguistique outillée. In B. Combettes, L. Sarda, & D. Vigier (Eds.), *Connexion et indexation : Ces liens qui tissent le texte*. Lyon : ENS Éditions. doi: 10.4000/books.enseditions.6880
- Landragin, F., & Schnedecker, C. (2014). *Les chaînes de référence* (N° 195). Paris : Larousse.
- Lapshinova-Koltunski, E., & Kunz, K. (2020). Exploring Coreference Features in Heterogeneous Data. In *Proceedings of the First Workshop on Computational Approaches to Discourse* (pp. 53–64). Online : Association for Computational Linguistics. doi: 10.18653/v1/2020.codi-1.6
- LDC. (2008). *ACE (Automatic Content Extraction) English Annotation Guidelines for Entities* (Version 6.6 2008.06.13).
- Lecolle, M. (2006). Polyvalence des toponymes et interprétation en contexte. *Pratiques*, 129(1), 107–122. doi: 10.3406/prati.2006.2100
- Lecolle, M. (2016, octobre). Noms collectifs humains : nomination et prédication. *Argumentation et analyse du discours*(17). doi: 10.4000/aad.2208
- Lecolle, M. (2019). *Les noms collectifs humains en français : Enjeux sémantiques, lexicaux et discursifs*. Limoges : Lambert-Lucas.
- Lesnard, L. (2010, février). Setting Cost in Optimal Matching to Uncover Contemporary Socio-Temporal Patterns. *Sociological Methods & Research*, 38(3), 389–419. doi: 10.1177/0049124110362526
- Lesnard, L., & de Saint Pol, T. (2006, avril). Introduction aux méthodes d’appariement optimal (Optimal Matching Analysis). *Bulletin de méthodologie sociologique. Bulletin of sociological methodology*(90), 5–25.
- Levenshtein, V. (1966). Binary Codes Capable of Correcting Deletion, Insertions, and Reversals. , 10(8), 707–710.
- Li, S., Kobayashi, H., & Ng, V. (2021, novembre). The CODI-CRAC 2021 Shared Task on Anaphora, Bridging, and Discourse Deixis Resolution in Dialogue :

- A Cross-Team Analysis. In S. Khosla, R. Manuvinakurike, V. Ng, M. Poesio, M. Strube, & C. Rosé (Eds.), *Proceedings of the CODI-CRAC 2021 Shared Task on Anaphora, Bridging, and Discourse Deixis in Dialogue* (pp. 71–95). Punta Cana, Dominican Republic : Association for Computational Linguistics. doi: 10.18653/v1/2021.codi-sharedtask.8
- Liao, T. F., Bolano, D., Brzinsky-Fay, C., Cornwell, B., Fasang, A. E., Helske, S., ... Studer, M. (2022). Sequence analysis : Its past, present, and future. *Social Science Research*, 107, 102772. doi: 10.1016/j.ssresearch.2022.102772
- Lloyd, S. (1982, mars). Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28(2), 129–137. doi: 10.1109/TIT.1982.1056489
- Longo, L. (2013). *Vers des moteurs de recherche « intelligents » : Un outil de détection automatique de thèmes. Méthode basée sur l'identification automatique des chaînes de référence* (Thèse de doctorat non publiée). Université de Strasbourg.
- Longo, L., & Todirascu, A. (2010, juin). Genre-based Reference Chains Identification for French. *Investigationes Linguisticae*, 21, 57. doi: 10.14746/il.2010.21.4
- Longo, L., & Todirascu, A. (2014, septembre). Vers une typologie des chaînes de référence dans des textes administratifs et juridiques :. *Langages*, N° 195(3), 79–98. doi: 10.3917/lang.195.0079
- Lundquist, L. (2005). Noms, verbes et anaphores (in) fidèles. Pourquoi les Danois sont plus fidèles que les Français. *Langue française*, 145(1), 73–91. doi: 10.3917/lf.145.0073
- Macqueen, J. (1967). Some methods for classification and analysis of multivariate observations. In L. M. Le Cam & J. Neyman (Eds.), *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability* (Vol. 1, pp. 281–297). California : University of California Press.
- Mann, W. C., & Thompson, S. A. (1987). Rhetorical Structure Theory : Description and Construction of Text Structures. In G. Kempen (Ed.), *Natural Language Generation : New Results in Artificial Intelligence, Psychology and Linguistics* (pp. 85–95). Dordrecht : Springer Netherlands. doi: 10.1007/978-94-009-3645-4_7
- Manuélian, H. (2003, juin). Une analyse des emplois du démonstratif en corpus. In B. Daille & E. Morin (Eds.), *Actes de la 10ème conférence sur le Traitement Automatique des Langues Naturelles. Articles longs* (pp. 205–214). Batz-sur-Mer, France : ATALA.
- Manuélian, H., & Schnedecker, C. (2021, décembre). Introduction des référents et premier maillon de suites coréférentielles : La réalité des textes :. *Langages*, N° 224(4), 67–86. doi: 10.3917/lang.224.0067
- Manuélian, H., & Schnedecker, C. (2022, mai). Identifier les « singletons » dans des corpus français annotés en coréférence : Peut-on prévoir l'absence de reprise coréférentielle ? *Discours*(29). doi: 10.4000/discours.11729
- Masseglia, F., Teisseire, M., & Poncelet, P. (2004, août). Extraction de motifs séquentiels. Problèmes et méthodes. *Ingénierie des systèmes d'information*, 9(3-4), 183–210. doi: 10.3166/isi.9.3-4.183-210
- Matthias, S., & Gilbert, R. (2014). A comparative review of sequence dissimilarity measures. doi: 10.12682/LIVES.2296-1658.2014.33

- Mélanie-Becquet, F., & Landragin, F. (2014, septembre). Linguistique outillée pour l'étude des chaînes de référence : questions méthodologiques et solutions techniques. *Langages*, 195(3), 117–137. doi: 10.3917/lang.195.0117
- Mitkov, R. (2002). *Anaphora resolution*. London New York : Routledge Taylor & Francis Group.
- Mitkov, R. (2005). Anaphora Resolution. In *The Oxford handbook of computational linguistics* (Reprinted éd., pp. 266–283). Oxford : Oxford Univ. Press.
- Mitkov, R., Evans, R., Orasan, C., Barbu, C., Jones, L., & Sotirova, V. (2000). Co-reference and anaphora : Developing annotating tools, annotated resources and annotation strategies. In *Discourse Anaphora and Anaphora Resolution Colloquium (DAARC2000)*. Lancaster, Royaume-Uni. .
- Müller, C., & Strube, M. (2006). Multi-level annotation of linguistic data with MMAX2.
- Muzerelle, J., Lefevre, A., Antoine, J.-Y., Schang, E., Maurel, D., Villaneau, J., & Eshkol, I. (2013, juin). ANCOR, the first large French speaking corpus of conversational speech annotated in coreference to be freely available (ANCOR, premier corpus de français parlé d'envergure annoté en coréférence et distribué librement) [in French]. In *Proceedings of TALN 2013 (Volume 2 : Short Papers)* (pp. 555–563). Les Sables d'Olonne, France : ATALA.
- Nedoluzhko, A., Novák, M., Popel, M., Žabokrtský, Z., Zeldes, A., & Zeman, D. (2022). CorefUD 1.0 : Coreference Meets Universal Dependencies.
- Nedoluzhko, A., Novák, M., Popel, M., Žabokrtský, Z., & Zeman, D. (2021). CorefUD 0.1 – a pilot experiment on harmonizing coreference datasets for 11 languages.
- Ng, V. (2007). Shallow Semantics for Coreference Resolution. In *Twentieth International Joint Conference on Artificial Intelligence (IJCAI)* (pp. 1689–1694). Hyderabad, India.
- Ng, V., & Cardie, C. (2002, juillet). Improving machine learning approaches to coreference resolution. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics* (pp. 104–111). USA : Association for Computational Linguistics. doi: 10.3115/1073083.1073102
- Oberlé, B. (2016). *Étude des chaînes de référence dans les articles de recherche de format IMRaD : Problèmes d'annotation, analyse quantitative et qualitative* (Mémoire de Master 2 Sciences Du Langage,). Université de Strasbourg.
- Oberlé, B., Schnedecker, C., Baumer, E., Capin, D., Glikman, J., Guo, C., ... Tushkova, J. (2018). Les chaînes de référence dans les textes encyclopédiques du 12e au 21e siècle : étude longitudinale. *Travaux de linguistique*, 77(2), 67–141. doi: 10.3917/tl.077.0067
- Obry, V., Glikman, J., Guillot-Barbance, C., & Pincemin, B. (2017). Les chaînes de référence dans les récits brefs en français : étude diachronique (XIIIe-XVIe s.). *Langue française*, 195(3), 91–110. doi: 10.3917/lf.195.0091
- Perret, M. (2000). Quelques remarques sur l'anaphore nominale aux XIVe et XVe siècles. *L'information grammaticale*, 87(1), 17–23. doi: 10.3406/igram.2000.2740
- Péry-Woodley, M.-P. (2000). Cadrer ou centrer son discours ? Introduteurs de cadres et centrage. , 59–78.

- Péry-Woodley, M.-P., Afantenos, S., Ho-Dac, L.-M., & Asher, N. (2011). La ressource ANNODIS, un corpus enrichi d'annotations discursives. , 52(3), 71–101.
- Pincemin, B., & Malrieu, D. (2014). Caractérisation quantitative de textes. Application à l'oral représenté, en diachronie. In *Driss ABLALI, Sémir BADIR, Dominique DUCARD (dir) Documents, textes, œuvres -Perspectives sémiotiques*, 43–56.
- Poesio, M. (2000). *GNOME Annotation Scheme Manual* (Version 4).
- Poesio, M., Bruneseaux, F., & Romary, L. (1999). The MATE meta-scheme for coreference in dialogues in multiple languages. In *Towards Standards and Tools for Discourse Tagging*.
- Poesio, M., Pradhan, S., Recasens, M., Rodriguez, K., & Versley, Y. (2016). Annotated Corpora and Annotation Tools. In *Anaphora Resolution* (pp. 98–140). Berlin, Heidelberg : Springer Berlin Heidelberg. doi: 10.1007/978-3-662-47909-4
- Ponton, C., Rafaela, G.-C., Lilia, T., Elisa, F., Brissaud, C., & Wolfarth, C. (2021). Scolinter : un corpus trilingue. L'exemple de la segmentation en mots. *Langue française*, 211(3), 37. doi: 10.3917/lf.211.0037
- Potau, M. R. (2010). *Coreference : Theory, Annotation, Resolution and Evaluation* (Thèse de doctorat non publiée). Universitat de Barcelona.
- Pradhan, S., Moschitti, A., Xue, N., Uryupina, O., & Zhang, Y. (2012, juillet). CoNLL-2012 Shared Task : Modeling Multilingual Unrestricted Coreference in OntoNotes. In *Proceedings of the Joint Conference on EMNLP and CoNLL : Shared Task* (pp. 1–40). Jeju Island, Korea, : Association for Computational Linguistics.
- Qi, P., Zhang, Y., Zhang, Y., Bolton, J., & Manning, C. D. (2020, avril). *Stanza : A Python Natural Language Processing Toolkit for Many Human Languages* (N° arXiv :2003.07082). arXiv. doi: 10.48550/arXiv.2003.07082
- Quignard, M., Le Mené, M., & Landragin, F. (2021, décembre). Élaboration du corpus DEMOCRAT : Procédures d'annotation et d'évaluation :. *Langages*, N° 224(4), 25–46. doi: 10.3917/lang.224.0025
- Recasens, M., & Hovy, E. (2010, juillet). Coreference Resolution across Corpora : Languages, Coding Schemes, and Preprocessing Information. In J. Hajič, S. Carberry, S. Clark, & J. Nivre (Eds.), *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics* (pp. 1423–1432). Uppsala, Sweden : Association for Computational Linguistics.
- Recasens, M., Hovy, E., & Martí, M. A. (2010, mai). A Typology of Near-Identity Relations for Coreference (NIDENT). In N. Calzolari et al. (Eds.), *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*. Valletta, Malta : European Language Resources Association (ELRA).
- Recasens, M., Màrquez, L., Sapena, E., Martí, M. A., Taulé, M., Hoste, V., ... Versley, Y. (2010, juillet). SemEval-2010 Task 1 : Coreference Resolution in Multiple Languages. In K. Erk & C. Strapparava (Eds.), *Proceedings of the 5th International Workshop on Semantic Evaluation* (pp. 1–8). Uppsala, Sweden : Association for Computational Linguistics.
- Recasens, M., & Martí, M. A. (2010, décembre). AnCora-CO : Coreferentially annotated corpora for Spanish and Catalan. *Language Resources and Evaluation*, 44(4), 315–345. doi: 10.1007/s10579-009-9108-x
- Reynolds, A. P., Richards, G., & Rayward-Smith, V. J. (2004). The Application of

- K-Medoids and PAM to the Clustering of Rules. In Z. R. Yang, H. Yin, & R. M. Everson (Eds.), *Intelligent Data Engineering and Automated Learning – IDEAL 2004* (pp. 173–178). Berlin, Heidelberg : Springer. doi: 10.1007/978-3-540-28651-6_25
- Ritschard, G., Liao, T. F., & Struffolino, E. (2023, août). Strategies for Multidomain Sequence Analysis in Social Research. *Sociological Methodology*, 53(2), 288–322. doi: 10.1177/00811750231163833
- Robette, N. (2011). *Explorer et décrire les parcours de vie : les typologies de trajectoires*. CEPED.
- Robette, N. (2018, avril). *L'analyse de séquences : une introduction avec le logiciel R et le package TraMineR* [Billet]. doi: 10.58079/t4av
- Rousier-Vercruyssen, L., & Landragin, F. (2019, décembre). Interdistance et instabilité au sein des chaînes de référence : Indices textuels ? *Discours*(25). doi: 10.4000/discours.10522
- Rousseeuw, P. J. (1987, novembre). Silhouettes : A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20, 53–65. doi: 10.1016/0377-0427(87)90125-7
- Salles, M. (2015a, septembre). Anaphore possessive et anaphore associative : le cas des noms collectifs. *Discours*(16). doi: 10.4000/discours.8981
- Salles, M. (2015b). Chaînes de référence : la deuxième mention. L'exemple des entités inanimées dans les narrations littéraires. *Travaux de linguistique*, 71(2), 111–133. doi: 10.3917/tl.071.0111
- Schmid, H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceedings of International Conference on New Methods in Language Processing*. Manchester, UK.
- Schnedecker, C. (1997). *Nom propre et chaînes de référence* (Klincksieck, Ed.). Librairie KLINCKSIECK.
- Schnedecker, C. (2005). Les chaînes de référence dans les portraits journalistiques : éléments de description. *Travaux de linguistique*, 51(2), 85–133. doi: 10.3917/tl.051.0085
- Schnedecker, C. (2014, septembre). Chaînes de référence et variations selon le genre : *Langages*, N° 195(3), 23–42. doi: 10.3917/lang.195.0023
- Schnedecker, C. (2017, septembre). Les chaînes de référence : Une configuration d'indices pour distinguer et identifier les genres textuels : *Langue française*, N° 195(3), 53–72. doi: 10.3917/lf.195.0053
- Schnedecker, C. (2019, juin). De l'intérêt de la notion de chaîne de référence par rapport à celles d'anaphore et de coréférence. *Cahiers de praxématique*(72). doi: 10.4000/praxématique.5339
- Schnedecker, C. (2021). *Les chaînes de référence en français*. Paris : Éditions Ophrys.
- Schnedecker, C., Glikman, J., & Landragin, F. (2017, septembre). Les chaînes de référence : annotation, application et questions théoriques : *Langue française*, N° 195(3), 5–16. doi: 10.3917/lf.195.0005
- Schnedecker, C., & Landragin, F. (2014, septembre). Les chaînes de référence : Présentation : *Langages*, N° 195(3), 3–22. doi: 10.3917/lang.195.0003
- Schnedecker, C., & Longo, L. (2012). Impact des genres sur la composition des

- chaînes de référence : le cas des faits divers. *SHS Web of Conferences*, 1, 1957–1972. doi: 10.1051/shsconf/20120100061
- Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, 27(3), 379–423. doi: 10.1002/j.1538-7305.1948.tb01338.x
- Soon, W. M., Ng, H. T., & Lim, D. C. Y. (2001, décembre). A Machine Learning Approach to Coreference Resolution of Noun Phrases. *Computational Linguistics*, 27(4), 521–544. doi: 10.1162/089120101753342653
- Strube, M., & Hahn, U. (1996). Functional centering. In *Proceedings of the 34th annual meeting on Association for Computational Linguistics* - (pp. 270–277). Santa Cruz, California : Association for Computational Linguistics. doi: 10.3115/981863.981899
- Studer, M. (2012a). *Étude des inégalités de genre en début de carrière académique à l'aide de méthodes innovatrices d'analyse de données séquentielles. Chapitre : Le manuel de la librairie WeightedCluster : Un guide pratique pour la création de typologies de trajectoires en sciences sociales avec R.* (Thèse SES 777, Faculté des sciences économiques et sociales, Université de Genève). doi: 10.13097/ARCHIVE-OUVERTE/UNIGE:22054
- Studer, M. (2012b). Le manuel de la librairie WeightedCluster. Un guide pratique pour la création de typologies de trajectoires en sciences sociales avec R.
- Studer, M., & Ritschard, G. (2016, février). What Matters in Differences Between Life Trajectories : A Comparative Review of Sequence Dissimilarity Measures. *Journal of the Royal Statistical Society Series A : Statistics in Society*, 179(2), 481–511. doi: 10.1111/rssa.12125
- Sukthanker, R., Poria, S., Cambria, E., & Thirunavukarasu, R. (2020, juillet). Anaphora and coreference resolution : A review. *Information Fusion*, 59, 139–162. doi: 10.1016/j.inffus.2020.01.010
- Taboada, M., & Mann, W. C. (2006, juin). Rhetorical Structure Theory : Looking back and moving ahead. *Discourse Studies*, 8(3), 423–459. doi: 10.1177/1461445606061881
- Todirascu, A., François, T., Bernhard, D., Gala, N., Ligozat, A.-L., & Khobzi, R. (2017, octobre). Chaînes de référence et lisibilité des textes : le projet ALLuSIF. *Langue française*, 195(3), 35–52. doi: 10.3917/lf.195.0035
- Toshniwal, S., Xia, P., Wiseman, S., Livescu, K., & Gimpel, K. (2021, septembre). *On Generalization in Coreference Resolution* (N° arXiv :2109.09667). arXiv.
- Tutin, A., Trouilleux, F., Clouzot, C., Gaussier, É., Zaenen, A., Rayot, S., & Antoniadis, G. (2000). Annotating a large corpus with anaphoric links. In *Proceedings of the Third International Conference on Discourse Anaphora and Anaphor Resolution (DAARC2000)* (p. 2). United Kingdom.
- Urieli, A. (2013). *Robust French syntax analysis : Reconciling statistical methods and linguistic knowledge in the Talismane toolkit* (Thèse de doctorat non publiée). Université Toulouse 2 Le Mirail.
- Uryupina, O., Artstein, R., Bristot, A., Cavicchio, F., Delogu, F., Rodriguez, K. J., & Poesio, M. (2020, janvier). Annotating a broad range of anaphoric phenomena, in a variety of genres : The ARRAU Corpus. *Natural Language Engineering*, 26(1), 95–128. doi: 10.1017/S1351324919000056
- Uryupina, O., Kabadjov, M., & Poesio, M. (2016). Detecting Non-reference and

- Non-anaphoricity. In *Anaphora Resolution* (pp. 369–392). Berlin, Heidelberg : Springer Berlin Heidelberg.
- van Deemter, K., & Kibble, R. (2000). On Coreferring : Coreference in MUC and Related Annotation Schemes. *Computational Linguistics*, 26(4), 629–637.
- Vogels, J., Krahmer, E., & Maes, A. (2013, avril). When a Stone Tries to Climb up a Slope : The Interplay between Lexical and Perceptual Animacy in Referential Choices. *Frontiers in Psychology*, 4. doi: 10.3389/fpsyg.2013.00154
- Walker, J., Walker, M., Joshi, A., Walker, M., & Prince, E. (1998). *Centering Theory in Discourse*. Clarendon Press.
- Ward, J. H. (1963, mars). Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association*, 58(301), 236–244. doi: 10.1080/01621459.1963.10500845
- Widlöcher, A., & Mathet, Y. (2009). La plate-forme Glozz : environnement d’annotation et d’exploration de corpus. In *Actes de la 16ème conférence sur le Traitement Automatique des Langues Naturelles. Articles courts* (pp. 207–216). Senlis, France : ATALA.
- Wilkens, R., Oberle, B., Landragin, F., & Todirascu, A. (2020, mai). French Coreference for Spoken and Written Language. In *Proceedings of the Twelfth Language Resources and Evaluation Conference* (pp. 80–89). Marseille, France : European Language Resources Association.
- Wolna, A., Durlik, J., & Wodniecka, Z. (2022, janvier). Pronominal anaphora resolution in Polish : Investigating online sentence interpretation using eye-tracking. *PLOS ONE*, 17(1), e0262459. doi: 10.1371/journal.pone.0262459
- Wolter, L. (2010). Teaching & Learning Guide for : Demonstratives in Philosophy and Linguistics. *Philosophy Compass*, 5(1), 108–111. doi: 10.1111/j.1747-9991.2009.00267.x
- Yu, J., Khosla, S., Manuvinakurike, R., Levin, L., Ng, V., Poesio, M., ... Rosé, C. (2022, octobre). The CODI-CRAC 2022 Shared Task on Anaphora, Bridging, and Discourse Deixis in Dialogue. In J. Yu et al. (Eds.), *Proceedings of the CODI-CRAC 2022 Shared Task on Anaphora, Bridging, and Discourse Deixis in Dialogue* (pp. 1–14). Gyeongju, Republic of Korea : Association for Computational Linguistics.
- Zhu, Y., Pradhan, S., & Zeldes, A. (2021, juin). *OntoGUM : Evaluating Contextualized SOTA Coreference Resolution on 12 More Genres* (N° arXiv :2106.00933). arXiv.

Annexe A

Balises `<xml>` utilisées pour signaler la structure textuelle dans chaque fichier du corpus DEMOCRAT

id	Siècle	<s>	<p>	<div>	<pb>	<lb>	<head>	<body>
articleswiki	21e	✓	✗	✗	✗	✓	✗	✗
conventionthontropical	21e	✓	✗	✗	✗	✓	✗	✗
aden	20e	✓	✗	✗	✗	✓	✗	✗
codeprocedurepenale	20e	✓	✗	✗	✗	✓	✗	✗
conventioninstitutiv	20e	✓	✗	✗	✗	✓	✗	✗
conventionmilieumarin	20e	✓	✗	✗	✗	✓	✗	✗
convtransportaerien	20e	✓	✗	✗	✗	✓	✗	✗
delavilleaumoulin	20e	✓	✗	✗	✗	✓	✗	✗
diablaucorps	20e	✓	✗	✗	✗	✓	✗	✗
doucelumiere	20e	✓	✗	✗	✗	✓	✗	✗
elisabethseton	20e	✓	✗	✗	✗	✓	✗	✗
estrepublikain	20e	✓	✗	✗	✗	✓	✗	✗
estrepublikain-2	20e	✓	✗	✗	✗	✓	✗	✗
jeanchristophe-1	20e	✓	✗	✗	✗	✓	✗	✗
jeanchristophe-2	20e	✓	✗	✗	✗	✓	✗	✗
nemoville	20e	✓	✗	✗	✗	✓	✗	✗
rosaliedeconstant	20e	✓	✗	✗	✗	✓	✗	✗
bouvardetpecuchet	19e	✓	✗	✗	✗	✓	✗	✗
capitainefracasse	19e	✓	✗	✗	✗	✓	✗	✗
chateaubriandgenie	19e	✗	✓	✓	✓	✓	✓	✓
codecivilsdesfrancais-1	19e	✓	✗	✗	✗	✓	✗	✗
codecivilsdesfrancais-2	19e	✓	✗	✗	✗	✓	✗	✗
madamedehautefort	19e	✓	✗	✗	✗	✓	✗	✗
mademoisellefifi-1	19e	✓	✗	✗	✗	✓	✗	✗
mademoisellefifi-2	19e	✓	✗	✗	✗	✓	✗	✗
mademoisellefifi-3	19e	✓	✗	✗	✗	✓	✗	✗
morteamoureuse	19e	✓	✗	✗	✗	✓	✗	✗
pauline	19e	✓	✗	✗	✗	✓	✗	✗
sarrasine	19e	✓	✗	✗	✗	✓	✗	✗
ventredeparis	19e	✓	✗	✗	✗	✓	✗	✗
diderotessais	18e	✗	✓	✓	✓	✓	✓	✓
fables	18e	✓	✗	✗	✗	✓	✗	✗
montesquieuulois	18e	✗	✓	✓	✓	✓	✓	✓
pauletvirginie	18e	✓	✗	✗	✗	✓	✗	✗
ramsaycyrus	18e	✗	✓	✓	✓	✓	✓	✓
voltaireessai	18e	✗	✓	✓	✓	✓	✓	✓
bossuetdiscours	17e	✗	✓	✓	✓	✓	✓	✓
coeffeteauhistoire	17e	✗	✓	✓	✓	✓	✓	✓
descartesdiscours	17e	✗	✓	✓	✓	✓	✓	✓
lafayettecleves	17e	✗	✓	✓	✓	✓	✓	✓
perraultcontes-1	17e	✗	✓	✓	✓	✓	✓	✓
perraultcontes-2	17e	✗	✓	✓	✓	✓	✓	✓
serresagriculture	17e	✗	✓	✓	✓	✓	✓	✓
desperriersrecre	16e	✗	✓	✓	✓	✓	✓	✓
dubellaydeffense	16e	✗	✓	✓	✓	✓	✓	✓
lerybresil	16e	✗	✓	✓	✓	✓	✓	✓
rabelaispantagruel-v2	16e	✗	✓	✓	✓	✓	✓	✓
servitudevolontaire	16e	✓	✗	✗	✗	✓	✗	✗
commynes	15e	✓	✓	✓	✓	✓	✓	✓
jehandeparis	15e	✓	✓	✓	✓	✓	✓	✓
livredestroisvertus	15e	✗	✓	✓	✓	✓	✓	✓
daudin	14e	✗	✓	✓	✓	✓	✓	✓
moree	14e	✓	✓	✓	✓	✓	✗	✓
regcrim1	14e	✗	✓	✓	✓	✓	✓	✓
chartes-hain13	13e	✗	✓	✓	✓	✓	✓	✓
sbath1	13e	✗	✓	✓	✓	✓	✓	✓
eneas	12e	✓	✓	✓	✓	✓	✗	✓
roland	environ 11e	✓	✗	✓	✓	✓	✗	✓
Total		37✓-21✗	25✓- 33✗	26✓-32✗	26✓-32✗	58✓-0✗	23✓- 35✗	26✓-32✗

TABLEAU A.1 – Balises xml utilisées pour signaler la structure textuelle dans chaque fichier du corpus DEMOCRAT

Annexe B

Exemples complets des chaînes de référence

B.1 diableaucorps_RC_19–21_s–CHAINE–5

[diableaucorps_RC_19-21_s-CHAINE-5] - id : s-CHAINE-5, nb_maillons : 895 mentions, from : w_dd_10, to : w_dd_14592

Raymond Radiguet Le Diable au corps Grasset, 1923 **Je** [u-MENTION-5] vais encourir bien des reproches..

Mais qu'y puis **-je** [u-MENTION-8] ??

Est -ce **ma** [u-MENTION-10] faute si **j'** [u-MENTION-11] eus douze ans quelques mois avant la déclaration de la guerre??

Sans doute, les troubles qui me vinrent de cette période extraordinaire furent d'une sorte qu'on n'éprouve jamais à cet âge ; mais comme il n'existe rien d'assez fort pour nous vieillir malgré les apparences, c'est en enfant que **je** [u-MENTION-26] devais me conduire dans une aventure où déjà un homme eût éprouvé de l'embarras..

Je [u-MENTION-31] ne suis pas le seul..

Et **mes** [u-MENTION-33] camarades garderont de cette époque un souvenir qui n'est pas celui de leurs aînés..

Que ceux déjà qui **m'** [u-MENTION-42] en veulent se représentent ce que fut la guerre pour tant de très jeunes garçons : quatre ans de grandes vacances. . [phrase] [phrase]

Mes [u-MENTION-53] parents condamnaient plutôt la camaraderie mixte.. [phrase]

Je [u-MENTION-61] n'ai jamais été un rêveur..

Ce qui semble rêve aux autres, plus crédules, **me** [u-MENTION-66] paraissait à **moi** [u-MENTION-67] aussi réel que le fromage au chat, malgré la cloche de verre.. [phrase] [phrase]

Jusqu'à douze ans, **je** [u-MENTION-83] ne me vois aucune amourette, sauf pour une petite fille, nommée Carmen, à qui **je** [u-MENTION-87] fis tenir, par un gamin plus jeune que **moi** [u-MENTION-89], une lettre dans laquelle **je** [u-MENTION-92] lui exprimais mon amour..

Je [u-MENTION-95] m'autorisai de cet amour pour solliciter un rendez -vous..

Ma [u-MENTION-99] lettre lui avait été remise le matin avant qu'elle se rendît en classe..

J' [u-MENTION-105] avais distingué la seule fillette qui **me** [u-MENTION-108] ressemblât, parce qu'elle était propre, et allait à l'école accompagnée d'une petite, comme **moi** [u-MENTION-113] de **mon** [u-MENTION-115] petit frère..

Afin que ces deux témoins se tussent, **j'** [u-MENTION-117] imaginai de les marier, en quelque sorte..

À **ma** [u-MENTION-120] lettre, **j'** [u-MENTION-121] en joignis donc une de la part de **mon** [u-MENTION-125] frère, qui ne savait pas écrire, pour Mlle Fauvette..

J' [u-MENTION-128] expliquai à **mon** [u-MENTION-130] frère **mon** [u-MENTION-132] entremise, et notre chance de tomber juste sur deux sœurs de nos âges et douées de noms de baptêmes aussi exceptionnels..

J' [u-MENTION-140] eus la tristesse de voir que **je** [u-MENTION-142] ne m'étais pas mépris sur le bon genre de Carmen, lorsque, après avoir déjeuné avec **mes** [u-MENTION-147] parents qui **me** [u-MENTION-149] gâtaient et ne **me** [u-MENTION-150] grondaient jamais, **je** [u-MENTION-152] rentraï en classe. .

À peine **mes** [u-MENTION-155] camarades à leurs pupitres — **moi** [u-MENTION-158] en haut de la classe, accroupi pour prendre dans un placard, en **ma** [u-MENTION-162] qualité de premier, les volumes de la lecture à haute voix —, le directeur entra.. [phrase] [phrase]

Mes [u-MENTION-171] jambes fléchirent, les volumes tombèrent, et **je** [u-MENTION-173] les ramassai, tandis que le directeur s'entretenait avec le maître..

Déjà, les élèves des premiers bancs se tournaient vers **moi** [u-MENTION-178], écarlate, au fond de la classe, car ils entendaient chuchoter **mon** [u-MENTION-182] nom..

Enfin, le directeur **m'** [u-MENTION-184] appela, et pour **me** [u-MENTION-185] punir finement, tout en n'éveillant, croyait -il, aucune mauvaise idée chez les élèves, **me** [u-MENTION-189] félicita d'avoir écrit une lettre de douze lignes sans aucune faute..

Il **me** [u-MENTION-195] demanda si **je** [u-MENTION-196] l'avais bien écrite seul, puis il **me** [u-MENTION-199] pria de le suivre dans son bureau.. [phrase]

Il **me** [u-MENTION-206] morigéna dans la cour, sous l'averse..

Ce qui troubla fort **mes** [u-MENTION-211] notions de morale, fut qu'il considérait comme aussi grave d'avoir compromis la jeune fille (dont les parents lui avaient communiqué **ma** [u-MENTION-217] déclaration), que d'avoir dérobé une feuille de papier à lettres..

Il **me** [u-MENTION-220] menaça d'envoyer cette feuille chez **moi** [u-MENTION-222] ..

Je [u-MENTION-223] le suppliai de n'en rien faire..

Il céda, mais **me** [u-MENTION-227] dit qu'il conservait la lettre, et qu'à la première récidive il ne pourrait plus cacher **ma** [u-MENTION-234] mauvaise conduite. Ce mélange d'effronterie et de timidité déroutait les miens et les trompait, comme, à l'école, ma facilité, véritable paresse, **me** [u-MENTION-241] faisait prendre pour un bon élève. .

Je [u-MENTION-243] rentraï en classe..

Le professeur, ironique, **m'** [u-MENTION-246] appela Don Juan..

J' [u-MENTION-248] en fus extrêmement flatté, surtout de ce qu'il **me** [u-MENTION-251] citât le nom d'une œuvre que **je** [u-MENTION-255] connaissais et que ne connaissaient pas **mes** [u-MENTION-258] camarades..

Son « Bonjour, Don Juan » et **mon** [u-MENTION-264] sourire entendu transformèrent la classe à **mon** [u-MENTION-267] égard..

Peut-être avait -elle déjà su que **j'** [u-MENTION-269] avais chargé un enfant des petites classes de porter une lettre à une « fille », comme disent les écoliers dans leur dur langage..

Cet enfant s'appelait Messenger ; **je** [u-MENTION-279] ne l'avais pas élu d'après son nom, mais, quand même, ce nom **m'** [u-MENTION-284]

avait inspiré confiance. .

À une heure, **j'** [u-MENTION-285] avais supplié le directeur de ne rien dire à **mon** [u-MENTION-288] père ; à quatre, **je** [u-MENTION-289] brûlais de lui raconter tout..

Rien ne **m'** [u-MENTION-291] y obligeait..

Je [u-MENTION-293] mettrais cet aveu sur le compte de la franchise..

Sachant que mon père ne se fâcherait pas, **j'** [u-MENTION-297] étais, somme toute, ravi qu'il connût **ma** [u-MENTION-300] prouesse. .

J' [u-MENTION-301] avouai donc, ajoutant avec orgueil que le directeur **m'** [u-MENTION-303] avait promis une discrétion absolue (comme à une grande personne)..

Mon père voulait savoir si **je** [u-MENTION-307] n'avais pas forgé de toutes pièces ce roman d'amour.. [phrase] [phrase] [phrase]

dit alors le directeur surpris et très ennuyé ; **il** [u-MENTION-318] vous a raconté cela??

Il [u-MENTION-320] m'avait supplié de me taire, disant que vous **le** [u-MENTION-323] tueriez. .

Ce mensonge du directeur l'excusait ; il contribua encore à **mon** [u-MENTION-329] ivresse d'homme..

J' [u-MENTION-330] y gagnai séance tenante l'estime de **mes** [u-MENTION-335] camarades et des clignements d'yeux du maître.. [phrase]

Le malheureux ignorait ce que **je** [u-MENTION-343] savais déjà : **mon** [u-MENTION-345] père, choqué par sa conduite, avait décidé de **me** [u-MENTION-348] laisser finir **mon** [u-MENTION-350] année scolaire, et de **me** [u-MENTION-351] reprendre.. [phrase]

Ma [u-MENTION-357] mère ne voulant pas que cela influât sur **mes** [u-MENTION-360] prix, **mes** [u-MENTION-362] couronnes, se réservait de dire la chose, après la distribution..

Ce jour venu, grâce à une injustice du directeur qui craignait confusément les suites de son mensonge, **seul de la classe** [u-MENTION-373] , **je** [u-MENTION-375] reçus la couronne d'or que méritait aussi le prix d'excellence.. [phrase] [phrase]

Ma [u-MENTION-394] mère **me** [u-MENTION-395] jugeait trop jeune pour aller à Henri Iv.. [phrase]

Je [u-MENTION-401] restai deux ans à la maison et **travaillai** [u-MENTION-404] seul. .

Je [u-MENTION-405] me promettais des joies sans bornes, car, réussissant à faire en quatre heures le travail que ne fournissaient pas en deux jours mes anciens condisciples, **j'** [u-MENTION-412] étais libre plus de la moitié du jour..

Je [u-MENTION-415] me promenais seul au bord de la Marne qui était tellement notre rivière que **mes** [u-MENTION-421] sœurs disaient, en parlant de la Seine, « une Marne »..

J' [u-MENTION-424] allais même dans le bateau de **mon** [u-MENTION-427] père, malgré sa défense ; mais **je** [u-MENTION-430] ne ramais pas, et sans m'avouer que **ma** [u-MENTION-432] peur n'était pas celle de lui désobéir, mais la peur tout court..

Je [u-MENTION-436] lisais, couché dans ce bateau.. [phrase] [phrase]

Aussi, bien plus tard, à l'âge où l'adolescent méprise les livres de la Bibliothèque rose, **je** [u-MENTION-449] pris goût à leur charme enfantin, alors qu'à cette époque **je** [u-MENTION-453] ne les aurais voulu lire pour rien au monde. .

Le désavantage de ces récréations alternant avec le travail était de transformer pour **moi** [u-MENTION-458] toute l'année en fausses vacances..

Ainsi, **mon** [u-MENTION-463] travail de chaque jour était -il peu de chose, mais, comme, travaillant moins de temps que les autres, **je** [u-MENTION-467] travaillais en plus pendant leurs vacances, ce peu de chose était le bouchon de liège qu'un chat garde toute sa vie au bout de la queue, alors qu'il préférerait sans doute un mois de casserole. .

Les vraies vacances approchaient, et **je** [u-MENTION-480] m'en occupais fort peu puisque c'était pour **moi** [u-MENTION-482] le même régime.. [phrase] [phrase] [phrase] [phrase] [phrase] [phrase] [phrase] [phrase] [phrase] [phrase] [phrase]

Tout cet ensemble **me** [u-MENTION-527] laisse un souvenir de feu d'artifice.. [phrase] [phrase] [phrase]

Mes [u-MENTION-539] frères et **mes** [u-MENTION-541] sœurs commençaient d'en vouloir à la guerre, ils la trouvaient longue.. [phrase] [phrase] [phrase]

Au lieu de quitter la table où les grandes personnes s'attardent, ils y restent pour entendre **mon** [u-MENTION-561] père parler de départ.. [phrase] [phrase]

Mes [u-MENTION-566] frères plaisaient **ma** [u-MENTION-568] petite sœur.. [phrase]

» **Ma** [u-MENTION-577] sœur sanglote.. [phrase] [phrase] [phrase] [phrase]

Tandis que chacun s'étonne, **je** [u-MENTION-586] découvre enfin les mobiles de ce patriotisme : un voyage à bicyclette!! [phrase] [phrase] [phrase] [phrase] [phrase] [phrase] [phrase] [phrase]

Voici comment : Nous nous moquions, **mes** [u-MENTION-623] frères et **moi** [u-MENTION-624] , d'un de nos voisins, homme grotesque, nain à barbiche blanche et à capuchon, conseiller municipal, nommé Maréchaud.. [phrase] [phrase] [phrase] [phrase] [phrase] [phrase] En revenant de l'école, et en y allant, **mes** [u-MENTION-653] frères tiraient sa sonnette, avec d'autant plus d'audace que le chien, qui pouvait avoir **mon** [u-MENTION-659] âge, n'était pas à craindre. .

La veille du 14 juillet 1914, en allant à la rencontre de **mes** [u-MENTION-662] frères, quelle ne fut pas ma surprise de voir un attroupement devant la grille des Maréchaud.. [phrase] [phrase] [phrase] [phrase] [phrase]

J' [u-MENTION-695] eusse voulu pouvoir rester là toujours, mais notre bonne, envoyée par **ma** [u-MENTION-701] mère, vint nous rappeler au travail..

Sans cela, **je** [u-MENTION-704] serais privé de fête..

Je [u-MENTION-706] partis la mort dans l'âme, et priant Dieu que la bonne fût encore sur le toit, lorsque **j'** [u-MENTION-709] irais chercher **mon** [u-MENTION-711] père à la gare. . [phrase]

Ils [u-MENTION-720] ne lui accordaient qu'une minute distraite. . [phrase] [phrase] [phrase] [phrase] [phrase] [phrase] [phrase] [phrase] [phrase] [phrase] [phrase] [phrase] [phrase] [phrase] [phrase] [phrase]

La folle disait des choses que **j'** [u-MENTION-861] ai oubliées, avec cette profonde mélancolie résignée que donne aux voix la certitude qu'on a raison, que tout le monde se trompe.. [phrase] [phrase] [phrase] [phrase] [phrase] [phrase] [phrase] [phrase] [phrase]
Je [u-MENTION-917] pensai à quelque fille, capitaine corsaire, restant seule sur son bateau qui sombre. . [phrase]
J' [u-MENTION-924] avais voulu rester avec **mon** [u-MENTION-926] père, tandis que **ma** [u-MENTION-928] mère, pour assouvir ce besoin de mal au cœur qu'ont les enfants, conduisait les siens au manège en montagnes russes..
Certes, **j'** [u-MENTION-934] éprouvais cet étrange besoin plus vivement que **mes** [u-MENTION-938] frères..
J' [u-MENTION-939] aimais que **mon** [u-MENTION-941] cœur batte plus vite et irrégulièrement..
Ce spectacle, d'une poésie profonde, **me** [u-MENTION-944] satisfaisait davantage..
« Comme **tu** [u-MENTION-945] es pâle », avait dit **ma** [u-MENTION-947] mère..
Je [u-MENTION-948] trouvais le prétexte des feux de Bengale..
Ils **me** [u-MENTION-952] donnaient, dis **-je** [u-MENTION-953], une couleur verte. .
— Je crains tout de même que cela **l'** [u-MENTION-957] impressionne trop, dit -elle à **mon** [u-MENTION-960] père. . [phrase]
Il [u-MENTION-963] peut regarder n'importe quoi, sauf un lapin qu'on écorche. .
Mon [u-MENTION-969] père disait cela pour que **je** [u-MENTION-971] restasse..
Mais il savait que ce spectacle **me** [u-MENTION-974] bouleversait..
Je [u-MENTION-975] sentais qu'il le bouleversait aussi..
Je [u-MENTION-978] lui demandai de **me** [u-MENTION-980] prendre sur ses épaules pour mieux voir..
En réalité, **j'** [u-MENTION-983] allais m'évanouir, **mes** [u-MENTION-985] jambes ne **me** [u-MENTION-986] portaient plus. . [phrase] [phrase]
[phrase] [phrase] [phrase]
Jusqu'ici **j'** [u-MENTION-1018] avais essayé de supporter tout, bien que **mes** [u-MENTION-1021] oreilles tintassent et que le cœur **me** [u-MENTION-1023] manquât..
Mais quand **j'** [u-MENTION-1024] entendis des gens crier : « Elle vit encore », **je** [u-MENTION-1027] tombai, sans connaissance, des épaules de mon père. .
Revenu à **moi** [u-MENTION-1031], il **m'** [u-MENTION-1033] entraîna au bord de la Marne.. [phrase]
Au retour, **je** [u-MENTION-1039] crus voir derrière la grille une silhouette blanche, le fantôme de la bonne!! [phrase]
Si **j'** [u-MENTION-1063] insiste sur un tel épisode, c'est qu'il fait comprendre mieux que tout autre l'étrange période de la guerre, et combien, plus que le pittoresque, **me** [u-MENTION-1070] frappait la poésie des choses. . [phrase] [phrase] [phrase]
Tandis que **ma** [u-MENTION-1085] tante parlait d'une amie, enfuie dès les premiers jours, après avoir enterré dans son jardin des pendules, des boîtes de sardines, **je** [u-MENTION-1095] demandai à **mon** [u-MENTION-1097] père le moyen d'emporter nos vieux livres ; c'est ce qu'il **me** [u-MENTION-1102] coûtait le plus de perdre. . [phrase]
Mes [u-MENTION-1110] sœurs, maintenant, allaient à J [phrase] [phrase] [phrase] [phrase]
Je [u-MENTION-1124] devais entrer au lycée Henri IV ; mais **mon** [u-MENTION-1127] père préféra **me** [u-MENTION-1128] garder encore un an à la campagne..
Ma [u-MENTION-1132] seule distraction de ce morne hiver fut de courir chez notre marchande de journaux, pour être sûr d'avoir un exemplaire du Mot, journal qui **me** [u-MENTION-1141] plaisait et paraissait le samedi..
Ce jour -là, **je** [u-MENTION-1144] n'étais jamais levé tard. .
Mais le printemps arriva, qu'égayèrent **mes** [u-MENTION-1149] premières incartades..
Sous prétexte de quêtes, ce printemps, plusieurs fois, **je** [u-MENTION-1152] me promenai, endimanché, une jeune personne à **ma** [u-MENTION-1155] droite..
Je [u-MENTION-1156] tenais le tronc ; elle, la corbeille d'insignes..
Dès la seconde quête, des confrères **m'** [u-MENTION-1163] apprirent à profiter de ces journées libres où l'on **me** [u-MENTION-1167] jetait dans les bras d'une petite fille.. [phrase]
Pour la première fois, **j'** [u-MENTION-1182] eus un ami..
J' [u-MENTION-1184] aimais à quêter avec sa sœur..
Pour la première fois, **je** [u-MENTION-1188] m'entendais avec un garçon aussi précoce que **moi** [u-MENTION-1190], admirant même sa beauté, son effronterie.. [phrase] [phrase] [phrase] [phrase]
René allait au lycée Henri IV, et **je** [u-MENTION-1210] serais dans sa classe, en troisième..
Il ne devait pas apprendre le grec ; il **me** [u-MENTION-1216] fit cet extrême sacrifice de convaincre ses parents de le lui laisser apprendre.. [phrase] [phrase] [phrase]
Ils y virent l'effet de **ma** [u-MENTION-1244] bonne influence, et, s'ils supportaient ses autres camarades, **j'** [u-MENTION-1248] étais, du moins, le seul ami qu'ils approuvassent. .
Pour la première fois, nul jour des vacances de cette année ne **me** [u-MENTION-1256] fut pesant..
Je [u-MENTION-1257] connus donc que personne n'échappe à son âge, et que **mon** [u-MENTION-1262] dangereux mépris s'était fondu comme glace dès que quelqu'un avait bien voulu prendre garde à **moi** [u-MENTION-1265], de la façon qui **me** [u-MENTION-1268] convenait.. [phrase]
Le jour de la rentrée des classes, René **me** [u-MENTION-1280] fut un guide précieux. .
Avec lui tout **me** [u-MENTION-1284] devenait plaisir, et **moi** [u-MENTION-1285] **qui** [u-MENTION-1286], seul, ne pouvais avancer d'un pas, **j'** [u-MENTION-1288] aimais faire à pied, deux fois par jour, le trajet qui sépare Henri IV de la gare de la Bastille, où nous prenions notre

train. .

Trois ans passèrent ainsi, sans autre amitié et sans autre espoir que les polissonneries du jeudi – avec les petites filles que les parents de **mon** [u-MENTION-1309] ami nous fournissaient innocemment, invitant ensemble à goûter les amis de leur fils et les amies de leur fille –, menues faveurs que nous dérobiaient, et qu'elles nous dérobaient, sous prétexte de jeux à gages. .

La belle saison venue, **mon** [u-MENTION-1328] père aimait à nous emmener, **mes** [u-MENTION-1332] frères et **moi** [u-MENTION-1333] , dans de longues promenades..

Un de nos buts favoris était Ormesson, et de suivre le Morbras, rivière large d'un mètre, traversant des prairies où poussent des fleurs qu'on ne rencontre nulle part ailleurs, et dont **j'** [u-MENTION-1350] ai oublié le nom.. [phrase] [phrase] [phrase] [phrase]

Mon [u-MENTION-1377] père **me** [u-MENTION-1378] dit que nous retrouverions à La Varenne des gens agréables, les Grangier..

Je [u-MENTION-1383] les connaissais pour avoir vu le nom de leur fille, Marthe, dans le catalogue d'une exposition de peinture..

Un jour, **j'** [u-MENTION-1392] avais entendu **mes** [u-MENTION-1394] parents parler de la visite d'un M. Grangier.. [phrase] [phrase]

Son père aurait voulu lui faire une surprise : que ses aquarelles figurassent dans une exposition de charité dont **ma** [u-MENTION-1413] mère était présidente.. [phrase] [phrase] [phrase]

Mais Mme Grangier paraissait l'aînée de son mari ; son inélégance, sa taille courte, firent qu'elle **me** [u-MENTION-1440] déplut au premier coup d'œil. .

Au cours de cette promenade, **je** [u-MENTION-1442] devais remarquer qu'elle fronçait souvent les sourcils, ce qui couvrait son front de rides auxquelles il fallait une minute pour disparaître..

Afin qu'elle eût tous les motifs de **me** [u-MENTION-1453] déplaire, sans que **je** [u-MENTION-1454] me reprochasse d'être injuste, **je** [u-MENTION-1455] souhaitais qu'elle employât des façons de parler assez communes..

Sur ce point, elle **me** [u-MENTION-1460] déçut. . [phrase] [phrase]

Je [u-MENTION-1469] tremblais à la perspective d'une promenade sans autre compagnie que celle de ses parents.. [phrase] [phrase] [phrase] [phrase]

Cette imprudente **me** [u-MENTION-1493] charma. . [phrase] [phrase] [phrase]

Sur la route, Marthe et **moi** [u-MENTION-1519] marchions en tête..

Mon [u-MENTION-1521] père marchait derrière, entre les Grangier. .

Mes [u-MENTION-1525] frères, eux, bâillaient avec ce nouveau petit camarade chétif, à qui l'on défendait de courir. .

Comme **je** [u-MENTION-1530] complimentais Marthe sur ses aquarelles, elle **me** [u-MENTION-1535] répondit modestement que c'étaient des études.. [phrase]

Elle **me** [u-MENTION-1542] montrerait mieux, des fleurs « stylisées »..

Je [u-MENTION-1544] jugeai bon, pour la première fois, de ne pas lui dire que **je** [u-MENTION-1547] trouvais ces sortes de fleurs ridicules.

Sous son chapeau, elle ne pouvait bien **me** [u-MENTION-1552] voir..

Moi [u-MENTION-1553] . **Je** [u-MENTION-1554] l'observais. .

— Vous ressemblez peu à madame votre mère, lui dis **-je** [u-MENTION-1560] .. [phrase]

— On me le dit quelquefois ; mais, quand **vous** [u-MENTION-1566] viendrez à la maison, je **vous** [u-MENTION-1569] montrerai des photographies de maman lorsqu'elle était jeune, je lui ressemble beaucoup. .

Je [u-MENTION-1575] fus attristé de cette réponse, et **je** [u-MENTION-1577] priai Dieu de ne point voir Marthe quand elle aurait l'âge de sa mère. .

Voulant dissiper le malaise de cette réponse pénible, et ne comprenant pas que, pénible, elle ne pouvait l'être que pour **moi** [u-MENTION-1587] , puisque heureusement Marthe ne voyait point sa mère avec **mes** [u-MENTION-1593] yeux, **je** [u-MENTION-1594] lui dis :

— Vous avez tort de vous coiffer de la sorte, les cheveux lisses vous iraient mieux. .

Je [u-MENTION-1600] restai terrifié, n'ayant jamais dit pareille chose à une femme..

Je [u-MENTION-1604] pensais à la façon dont **j'** [u-MENTION-1607] étais coiffé, **moi** [u-MENTION-1608] . .

— **Vous** [u-MENTION-1609] pourrez le demander à maman (comme si elle avait besoin de se justifier!! [phrase] [phrase]

« Quelle fille était -ce donc, pensais **-je** [u-MENTION-1624] , pour admettre qu'un gamin la querelle à propos de ses mèches??

» **J'** [u-MENTION-1629] essayais de deviner ses goûts en littérature ; **je** [u-MENTION-1633] fus heureux qu'elle connût Baudelaire et Verlaine, charmé de la façon dont elle aimait Baudelaire, qui n'était pourtant pas la mienne..

J' [u-MENTION-1644] y discernais une révolte.. [phrase] [phrase] [phrase] [phrase]

Désagréablement surpris d'apprendre qu'elle était fiancée, **je** [u-MENTION-1672] me réjouis de savoir qu'elle désobéissait à un soldat assez nigaud pour craindre Baudelaire..

Je [u-MENTION-1677] fus heureux de sentir qu'il devait souvent choquer Marthe..

Après la première surprise désagréable, **je** [u-MENTION-1682] me félicitai de son étroitesse, d'autant mieux que **j'** [u-MENTION-1685] eusse craint, s'il avait lui aussi goûté Les Fleurs du mal, que leur futur appartement ressemblât à celui de La Mort des amants..

Je [u-MENTION-1693] me demandai ensuite ce que cela pouvait bien **me** [u-MENTION-1695] faire. . [phrase]

Moi [u-MENTION-1700] **qui** [u-MENTION-1701] n'y allais jamais, **je** [u-MENTION-1704] lui proposai de l'y conduire, ajoutant que **j'** [u-MENTION-1707] y travaillais souvent..

Mais, craignant ensuite que **mon** [u-MENTION-1710] mensonge fût découvert, **je** [u-MENTION-1711] la priai de n'en point parler à **mon** [u-MENTION-1715] père..

Il ignorait, dis **-je** [u-MENTION-1717] , que **je** [u-MENTION-1718] manquais des cours de gymnastique pour me rendre à la Grande-

Chaumière..

Car **je** [u-MENTION-1721] ne voulais pas qu'elle pût se figurer que **je** [u-MENTION-1723] cachais l'académie à **mes** [u-MENTION-1726] parents, parce qu'ils **me** [u-MENTION-1728] défendaient de voir des femmes nues..

J' [u-MENTION-1730] étais heureux qu'il se fit un secret entre nous, et **moi** [u-MENTION-1733] , timide, me sentais déjà tyrannique avec elle. .

J' [u-MENTION-1735] étais fier aussi d'être préféré à la campagne, car nous n'avions pas encore fait allusion au décor de notre promenade.. [phrase] [phrase] [phrase]

Dans **ma** [u-MENTION-1763] candeur, **je** [u-MENTION-1764] regrettais d'avoir été si loin, et d'avoir tellement précipité les choses..

« Après une conversation moins sentimentale, plus naturelle, pensai -**je** [u-MENTION-1769] , **je** [u-MENTION-1770] pourrais éblouir Marthe, et m'attirer la bienveillance de ses parents, en racontant le passé de ce village..

» **Je** [u-MENTION-1777] m'en abstins..

Je [u-MENTION-1779] croyais avoir des raisons profondes, et **pensais** [u-MENTION-1781] qu'après tout ce qui s'était passé, une conversation tellement en dehors de nos inquiétudes communes ne pourrait que rompre le charme..

Je [u-MENTION-1788] croyais qu'il s'était passé des choses graves..

C'était d'ailleurs vrai, simplement, **je** [u-MENTION-1790] le sus dans la suite, parce que Marthe avait faussé notre conversation dans le même sens que **moi** [u-MENTION-1797] ..

Mais **moi** [u-MENTION-1798] **qui** [u-MENTION-1799] ne pouvais m'en rendre compte, **je** [u-MENTION-1801] me figurais lui avoir adressé des paroles significatives..

Je [u-MENTION-1804] croyais avoir déclaré **mon** [u-MENTION-1806] amour à une personne insensible..

J' [u-MENTION-1808] oubliais que M. et Mme Grangier eussent pu entendre sans le moindre inconvénient tout ce que **j'** [u-MENTION-1812] avais dit à leur fille ; mais, **moi** [u-MENTION-1815] , aurais -**je** [u-MENTION-1816] pu le lui dire en leur présence? ?

— Marthe ne **m'** [u-MENTION-1822] intimide pas, me répétais -**je** [u-MENTION-1823] ..

Donc, seuls, ses parents et **mon** [u-MENTION-1828] père **m'** [u-MENTION-1829] empêchent de me pencher sur son cou et de l'embrasser. . Profondément en **moi** [u-MENTION-1833] , un autre garçon se félicitait de ces trouble-fête..

Celui -ci pensait : — Quelle chance que **je** [u-MENTION-1839] ne me trouve pas seul avec elle!!

Car **je** [u-MENTION-1841] n'oserais pas davantage l'embrasser, et n'**aurais** [u-MENTION-1843] aucune excuse. . [phrase] [phrase] [phrase]

Je [u-MENTION-1856] dus subir les compliments de Mme Grangier..

Ils **m'** [u-MENTION-1860] humiliaient..

Ils rappelaient à sa fille que **je** [u-MENTION-1864] n'étais encore qu'un lycéen, qui passerait son baccalauréat dans un an..

Marthe voulut boire de la grenadine ; **j'** [u-MENTION-1872] en commandai aussi..

Le matin encore, **je** [u-MENTION-1875] me serais cru déshonoré en buvant de la grenadine..

Mon [u-MENTION-1878] père n'y comprenait rien..

Il **me** [u-MENTION-1881] laissait toujours servir des apéritifs..

Je [u-MENTION-1884] tremblai qu'il **me** [u-MENTION-1886] plaisantât sur **ma** [u-MENTION-1888] sagesse..

Il le fit, mais à mots couverts, de façon que Marthe ne devinât pas que **je** [u-MENTION-1892] buvais de la grenadine pour faire comme elle. . [phrase] [phrase]

Je [u-MENTION-1898] promis à Marthe de lui porter, le jeudi suivant, la collection du journal Le Mot et Une saison en enfer. . [phrase] [phrase] [phrase]

Mon [u-MENTION-1920] père et **mes** [u-MENTION-1922] frères s'étaient ennuyés, qu'importe!! [phrase]

Le lendemain, au lycée, **je** [u-MENTION-1926] n'éprouvai pas le besoin de raconter à René, à qui **je** [u-MENTION-1930] disais tout, **ma** [u-MENTION-1932] journée du dimanche..

Mais **je** [u-MENTION-1934] n'étais pas d'humeur à supporter qu'il **me** [u-MENTION-1936] raillât de n'avoir pas embrassé Marthe en cachette..

Autre chose **m'** [u-MENTION-1939] étonnait ; c'est qu'aujourd'hui **je** [u-MENTION-1941] trouvai René moins différent de **mes** [u-MENTION-1944] camarades. .

Ressentant de l'amour pour Marthe, **j'** [u-MENTION-1948] en ôtais à René, à **mes** [u-MENTION-1952] parents, à **mes** [u-MENTION-1954] sœurs. .

Je [u-MENTION-1955] me promettais bien cet effort de volonté de ne pas venir la voir avant le jour de notre rendez -vous..

Pourtant, le mardi soir, ne pouvant attendre, **je** [u-MENTION-1963] sus trouver à **ma** [u-MENTION-1965] faiblesse de bonnes excuses qui **me** [u-MENTION-1968] permissent de porter après le dîner le livre et les journaux..

Dans cette impatience, Marthe verrait la preuve de **mon** [u-MENTION-1978] amour, disais -**je** [u-MENTION-1979] , et si elle refuse de la voir, **je** [u-MENTION-1982] saurais bien l'y contraindre. .

Pendant un quart d'heure, **je** [u-MENTION-1986] courus comme un fou jusqu'à sa maison..

Alors, craignant de la déranger pendant son repas, **j'** [u-MENTION-1994] attendis, en nage, dix minutes, devant la grille..

Je [u-MENTION-1998] pensais que pendant ce temps **mes** [u-MENTION-2001] palpitations de cœur s'arrêteraient.. [phrase]

Je [u-MENTION-2003] manquai tourner bride, mais depuis quelques minutes, d'une fenêtre voisine, une femme **me** [u-MENTION-2007] regardait curieusement, voulant savoir ce que **je** [u-MENTION-2009] faisais, réfugié contre cette porte..

Elle **me** [u-MENTION-2012] décida..

Je [u-MENTION-2013] sonnai..

J' [u-MENTION-2014] entrai dans la maison..

Je [u-MENTION-2016] demandai à la domestique si Madame était chez elle..

Presque aussitôt, Mme Grangier parut dans la petite pièce où l'on **m'** [u-MENTION-2024] avait introduit. .

Je [u-MENTION-2025] sursautai, comme si la domestique eût dû comprendre que **j'** [u-MENTION-2027] avais demandé « Madame » par convenance et que **je** [u-MENTION-2029] voulais voir « Mademoiselle »..

Rougissant, **je** [u-MENTION-2031] priai Mme Grangier de **m'** [u-MENTION-2033] excuser de la déranger à pareille heure, comme s'il eût été une heure du matin : ne pouvant venir jeudi, **j'** [u-MENTION-2038] apportais le livre et les journaux à sa fille. .

— Cela tombe à merveille, **me** [u-MENTION-2045] dit Mme Grangier, car Marthe n'aurait pu **vous** [u-MENTION-2048] recevoir.. [phrase]
[phrase]

Je [u-MENTION-2060] m'en allai donc, et puisque **je** [u-MENTION-2061] n'avais plus de chance de la revoir jamais, croyais **-je** [u-MENTION-2064] , m'efforçais de ne plus penser à Marthe, et, par cela même, ne pensant qu'à elle. .

Pourtant, un mois après, un matin, sautant de **mon** [u-MENTION-2071] wagon à la gare de la Bastille, **je** [u-MENTION-2074] la vis qui descendait d'un autre.. [phrase]

Je [u-MENTION-2083] lui demandai de **m'** [u-MENTION-2085] accompagner jusqu'à Henri Iv. . [phrase]

Vexé qu'elle **me** [u-MENTION-2096] parlât études, comme si aucune autre conversation n'eût été de **mon** [u-MENTION-2100] âge, **je** [u-MENTION-2101] lui répondis aigrement que ce serait assez drôle. . [phrase]

Je [u-MENTION-2106] pensai à sa mère. .

Nous arrivions à Henri Iv, et, ne voulant pas la quitter sur ces paroles que **je** [u-MENTION-2114] croyais blessantes, **je** [u-MENTION-2115] décidai d'entrer en classe une heure plus tard, après le cours de dessin..

Je [u-MENTION-2120] fus heureux qu'en cette circonstance Marthe ne montrât pas de sagesse, ne **me** [u-MENTION-2124] fit aucun reproche, et, plutôt, semblât **me** [u-MENTION-2128] remercier d'un tel sacrifice, en réalité nul..

Je [u-MENTION-2130] lui fus reconnaissant qu'en échange elle ne **me** [u-MENTION-2133] proposât point de l'accompagner dans ses courses, mais qu'elle **me** [u-MENTION-2138] donnât son temps comme **je** [u-MENTION-2141] lui donnais le mien. . [phrase]

Je [u-MENTION-2150] renonçai au lycée..

J' [u-MENTION-2152] avais dans **ma** [u-MENTION-2154] poche, par miracle, plus d'argent que n'en a d'habitude un collégien en deux ans, ayant la veille vendu **mes** [u-MENTION-2161] timbres-poste les plus rares à la Bourse aux timbres, qui se tient derrière le Guignol des Champs-Élysées. .

Au cours de la conversation, Marthe **m'** [u-MENTION-2169] ayant appris qu'elle déjeunait chez ses beaux-parents, **je** [u-MENTION-2172] décidai de la résoudre à rester avec **moi** [u-MENTION-2174] .. [phrase] [phrase]

Mais, voyant que **je** [u-MENTION-2184] restais sur **ma** [u-MENTION-2186] chaise de fer, elle n'eut pas le courage de **me** [u-MENTION-2189] rappeler que **j'** [u-MENTION-2190] aurais dû être assis sur les bancs de Henri Iv. . [phrase] [phrase] [phrase] [phrase] [phrase]

J' [u-MENTION-2213] en aurai mal pour notre entente. .

— Ces chaises sont trop dures, **me** [u-MENTION-2218] dit -elle, comme pour s'excuser d'être debout. . [phrase]

Je [u-MENTION-2223] ne pus m'empêcher d'imaginer les dessins que le cannage imprime sur la peau. .

— Allons, accompagnez -moi dans les magasins, puisque **vous** [u-MENTION-2229] êtes décidé à ne pas aller en classe, dit Marthe, faisant pour la première fois allusion à ce que **je** [u-MENTION-2234] négligeais pour elle. .

Je [u-MENTION-2236] l'accompagnai dans plusieurs maisons de lingerie, l'empêchant de commander ce qui lui plaisait et ne **me** [u-MENTION-2242] plaisait pas ; par exemple, évitant le rose, qui **m'** [u-MENTION-2245] importune, et qui était sa couleur favorite. . [phrase]

Ne pensant pas qu'elle pouvait leur mentir pour le simple plaisir de rester en **ma** [u-MENTION-2259] compagnie, **je** [u-MENTION-2260] cherchai ce qui la déterminerait à **me** [u-MENTION-2263] suivre dans l'école buissonnière.. [phrase] [phrase] [phrase]

Je [u-MENTION-2274] tenais **mon** [u-MENTION-2276] prétexte..

À son refus, empreint d'une véritable déception, **je** [u-MENTION-2280] pensai qu'elle viendrait..

Au bout d'une demi-heure, ayant usé de tout pour la convaincre, et n'insistant même plus, **je** [u-MENTION-2284] l'accompagnai chez ses beaux-parents, dans l'état d'esprit d'un condamné à mort espérant jusqu'au dernier moment qu'un coup de main se fera sur la route du supplice..

Je [u-MENTION-2294] voyais s'approcher la rue, sans que rien ne se produisît.. [phrase]

Elle **me** [u-MENTION-2303] dit : — Attendez -moi une seconde.. [phrase]

Au bout de quelques minutes, n'en pouvant plus d'impatience, **j'** [u-MENTION-2313] avisai une marchande de fleurs et **je** [u-MENTION-2316] choisis une à une des roses rouges, dont **je** [u-MENTION-2319] fis faire une botte..

Je [u-MENTION-2321] ne pensais pas tant au plaisir de Marthe qu'à la nécessité pour elle de mentir encore ce soir pour expliquer à ses parents d'où venaient les roses..

Notre projet, lors de la première rencontre, d'aller à une académie de dessin ; le mensonge du téléphone qu'elle répéterait, ce soir, à ses parents, mensonge auquel s'ajouterait celui des roses, **m'** [u-MENTION-2343] étaient des faveurs plus douces qu'un baiser..

Car, ayant souvent embrassé, sans grand plaisir, des lèvres de petites filles, et oubliant que c'était parce que **je** [u-MENTION-2349] ne les aimais pas, **je** [u-MENTION-2351] désirais peu les lèvres de Marthe..

Tandis qu'une telle complicité **m'** [u-MENTION-2355] était restée, jusqu'à ce jour, inconnue. . [phrase]

Je [u-MENTION-2361] donnai au chauffeur l'adresse d'un bar de la rue Daunou. . [phrase]

Elle respirait de temps en temps les roses rouges dont elle se promettait de faire une aquarelle, qu'elle **me** [u-MENTION-2382] donnerait en souvenir de cette journée..

Je [u-MENTION-2385] lui demandai de **me** [u-MENTION-2387] montrer une photographie de son fiancé..

Je [u-MENTION-2391] le trouvai beau..

Sentant déjà quelle importance elle attachait à **mes** [u-MENTION-2396] opinions, **je** [u-MENTION-2397] poussai l'hypocrisie jusqu'à lui dire qu'il était très beau, mais d'un air peu convaincu, pour lui donner à penser que **je** [u-MENTION-2403] le lui disais par politesse..

Ce qui, selon **moi** [u-MENTION-2407], devait jeter le trouble dans l'âme de Marthe, et, de plus, **m'** [u-MENTION-2411] attirer sa reconnaissance. . [phrase] [phrase] [phrase] [phrase] [phrase]

Bien que **je** [u-MENTION-2438] me fusse promis de ne montrer d'extrême plaisir ou déplaisir à aucune des paroles de Marthe, il **me** [u-MENTION-2443] fallut faire un effort pour continuer de marcher sur le boulevard d'un pas tranquille qui maintenant ne s'accordait plus avec le rythme de **mon** [u-MENTION-2451] cœur. .

Cette obligation d'accompagner Marthe **m'** [u-MENTION-2454] apparut comme une malchance.. [phrase]

Puis, **j'** [u-MENTION-2461] entrevis le moyen de choisir une chambre pour Marthe et pour **moi** [u-MENTION-2465] . .

J' [u-MENTION-2466] oubliais si vite son fiancé, qu'au bout d'un quart d'heure de marche, on **m'** [u-MENTION-2471] aurait surpris en **me** [u-MENTION-2472] rappelant que, dans cette chambre, un autre dormirait auprès d'elle. . [phrase] [phrase]

Il **me** [u-MENTION-2483] fallut donc les combattre tous deux.. [phrase]

Au moindre mot de Marthe, devinant ce qui la tentait, il **me** [u-MENTION-2490] fallait lui désigner le contraire, qui ne **me** [u-MENTION-2494] plaisait pas toujours, afin de **me** [u-MENTION-2496] donner l'apparence de céder à ses caprices, quand **j'** [u-MENTION-2500] abandonnerais un meuble pour un autre, qui dérangeait moins son œil. . [phrase]

» N'osant même plus **m'** [u-MENTION-2510] avouer ses propres goûts, elle les attribuait à son fiancé..

Je [u-MENTION-2517] devinai que dans quelques jours nous les raillerions ensemble. .

Pourtant **je** [u-MENTION-2522] ne comprenais pas bien cette faiblesse..

« Si elle ne **m'** [u-MENTION-2525] aime pas, pensai -**je** [u-MENTION-2526], quelle raison a-t-elle de **me** [u-MENTION-2530] céder, de sacrifier ses préférences, et celles de ce jeune homme, aux miennes??

» **Je** [u-MENTION-2536] n'en trouvai aucune..

La plus modeste eût été encore de **me** [u-MENTION-2538] dire que Marthe **m'** [u-MENTION-2540] aimait..

Pourtant **j'** [u-MENTION-2541] étais sûr du contraire. .

Marthe **m'** [u-MENTION-2544] avait dit : « Au moins laissons -lui l'étoffe rose.. [phrase]

» Rien que pour ce mot, **je** [u-MENTION-2549] me sentais près de lâcher prise.. [phrase]

Je [u-MENTION-2553] représentai à Marthe combien ces murs roses gâcheraient les meubles simples que « nous avions choisis », et, reculant encore devant le scandale, lui conseillai de faire peindre les murs de sa chambre à la chaux! ! [phrase] [phrase]

Elle se contenta de **me** [u-MENTION-2574] dire : « En effet, **vous** [u-MENTION-2575] avez raison..

» À la fin de cette journée éreintante, **je** [u-MENTION-2577] me félicitai du pas que **j'** [u-MENTION-2580] avais fait..

J' [u-MENTION-2581] étais parvenu à transformer, meuble à meuble, ce mariage d'amour, ou plutôt d'amourette, en un mariage de raison, et le quel!! [phrase]

En **me** [u-MENTION-2596] quittant, ce soir -là, au lieu d'éviter désormais **mes** [u-MENTION-2599] conseils, elle **m'** [u-MENTION-2601] avait prié de l'aider les jours suivants dans le choix de ses autres meubles..

Je [u-MENTION-2607] le lui promis, mais à condition qu'elle **me** [u-MENTION-2611] jurât de ne jamais le dire à son fiancé, puisque la seule raison qui pût à la longue lui faire admettre ces meubles, s'il avait de l'amour pour Marthe, c'était de penser que tout sortait d'elle, de son bon plaisir, qui deviendrait le leur. .

Quand **je** [u-MENTION-2628] rentrai à la maison, **je** [u-MENTION-2630] crus lire dans le regard de **mon** [u-MENTION-2633] père qu'il avait déjà appris **mon** [u-MENTION-2636] escapade.. [phrase] [phrase] [phrase]

En **me** [u-MENTION-2644] couchant, **je** [u-MENTION-2645] me répétais que, si elle songeait à son mariage avant de dormir, elle devait, ce soir, l'envisager de tout autre sorte qu'elle ne l'avait fait les jours précédents..

Pour **moi** [u-MENTION-2656], quelle que fût l'issue de cette idylle, **j'** [u-MENTION-2660] étais, d'avance, bien vengé de son Jacques : **je** [u-MENTION-2663] pensais à la nuit de noces dans cette chambre austère, dans « **ma** [u-MENTION-2667] » chambre! !

Le lendemain matin, **je** [u-MENTION-2669] guettai dans la rue le facteur qui devait apporter une lettre d'absence..

Il **me** [u-MENTION-2675] la remit, **je** [u-MENTION-2677] l'empochai, jetant les autres dans la boîte de notre grille.. [phrase]

Manquer la classe voulait dire, selon **moi** [u-MENTION-2687], que **j'** [u-MENTION-2688] étais amoureux de Marthe..

Je [u-MENTION-2690] me trompais..

Marthe ne **m'** [u-MENTION-2692] était que le prétexte de cette école buissonnière..

Et la preuve, c'est qu'après avoir goûté en compagnie de Marthe aux charmes de la liberté, **je** [u-MENTION-2700] voulus y goûter seul, puis faire des adeptes..

La liberté **me** [u-MENTION-2704] devint vite une drogue. .

L'année scolaire touchait à sa fin, et **je** [u-MENTION-2709] voyais avec terreur que **ma** [u-MENTION-2711] paresse allait rester impunie, alors que **je** [u-MENTION-2712] souhaitais le renvoi du collège, un drame, enfin, qui clôturât cette période. . [phrase]

Certes, **je** [u-MENTION-2726] ne cherchais pas à faire de la peine à **mon** [u-MENTION-2729] père ; pourtant, **je** [u-MENTION-2730] souhaitais

la chose qui pourrait lui en faire le plus..

Les classes **m'** [u-MENTION-2736] avaient toujours été un supplice ; Marthe et la liberté avaient achevé de **me** [u-MENTION-2741] les rendre intolérables..

Je [u-MENTION-2743] me rendais bien compte que, si **j'** [u-MENTION-2744] aimais moins René, c'était simplement parce qu'il **me** [u-MENTION-2748] rappelait quelque chose du collège..

Je [u-MENTION-2751] souffrais, et cette crainte **me** [u-MENTION-2753] rendait même physiquement malade, à l'idée de me retrouver, l'année suivante, dans la niaiserie de **mes** [u-MENTION-2758] condisciples. .

Pour le malheur de René, **je** [u-MENTION-2761] lui avais trop bien fait partager **mon** [u-MENTION-2764] vice..

Aussi, lorsque, moins habile que **moi** [u-MENTION-2765] , il **m'** [u-MENTION-2767] annonça qu'il était renvoyé de Henri Iv, **je** [u-MENTION-2770] crus l'être **moi-même** [u-MENTION-2772] ..

Il fallait l'apprendre à **mon** [u-MENTION-2775] père, car il **me** [u-MENTION-2777] saurait gré de le lui dire **moi-même** [u-MENTION-2780] , avant la lettre du censeur, lettre trop grave à subtiliser. . [phrase]

Le lendemain, jour de congé, **j'** [u-MENTION-2789] attendis que **mon** [u-MENTION-2791] père fût à Paris pour prévenir **ma** [u-MENTION-2794] mère.. [phrase]

Puis, **je** [u-MENTION-2802] partis au bord de la Marne, où Marthe **m'** [u-MENTION-2807] avait dit qu'elle **me** [u-MENTION-2809] rejoindrait peut-être.. [phrase] [phrase]

Mon, amour puisant dans cette rencontre une mauvaise énergie, **j'** [u-MENTION-2817] aurais pu, ensuite, lutter contre **mon** [u-MENTION-2819] père ; tandis que l'orage éclatant après une journée de vide, de tristesse, **je** [u-MENTION-2824] rentrai le front bas, comme il convenait..

Je [u-MENTION-2826] revins chez nous un peu après l'heure où **je** [u-MENTION-2832] savais que **mon** [u-MENTION-2834] père avait coutume d'y être.. [phrase]

Je [u-MENTION-2837] me promenai dans le jardin, attendant que **mon** [u-MENTION-2840] père **me** [u-MENTION-2841] fît venir..

Mes [u-MENTION-2843] sœurs jouaient en silence.. [phrase]

Un de **mes** [u-MENTION-2848] frères, assez excité par l'orage, **me** [u-MENTION-2850] dit de me rendre dans la chambre où **mon** [u-MENTION-2854] père s'était étendu. .

Des éclats de voix, des menaces, **m'** [u-MENTION-2858] eussent permis la révolte.. [phrase]

Mon [u-MENTION-2862] père se taisait ; ensuite, sans aucune colère, avec une voix même plus douce que de coutume, il **me** [u-MENTION-2866] dit : — Eh bien que comptes **-tu** [u-MENTION-2868] faire maintenant? ?

Les larmes qui ne pouvaient s'enfuir par **mes** [u-MENTION-2873] yeux, comme un essaim d'abeilles, bourdonnaient dans **ma** [u-MENTION-2876] tête..

À une volonté, **j'** [u-MENTION-2878] eusse pu opposer la mienne, même impuissante..

Mais devant une telle douceur, **je** [u-MENTION-2882] ne pensais qu'à me soumettre. .

— Ce que tu **m'** [u-MENTION-2885] ordonneras de faire. . [phrase]

Je t' [u-MENTION-2887] ai toujours laissé agir comme **tu** [u-MENTION-2889] voulais ; continue..

Sans doute auras **-tu** [u-MENTION-2890] à cœur de m'en faire repentir. . [phrase]

Mon [u-MENTION-2897] père ne **me** [u-MENTION-2898] demandait même pas de larmes..

Devant sa générosité, **j'** [u-MENTION-2903] avais honte du présent et de l'avenir..

Car **je** [u-MENTION-2907] sentais que quoi que **je** [u-MENTION-2910] lui dise, **je** [u-MENTION-2912] mentirais..

« Au moins que ce mensonge le reconforte, pensai **-je** [u-MENTION-2915] , en attendant de lui être une source de nouvelles peines..

» Ou plutôt non, **je** [u-MENTION-2919] cherche encore à me mentir à **moi-même** [u-MENTION-2920] ..

Ce que **je** [u-MENTION-2922] voulais, c'était faire un travail, guère plus fatigant qu'une promenade, et qui laissât comme elle, à mon esprit, la liberté de ne pas se détacher de Marthe une minute..

Je [u-MENTION-2931] feignis de vouloir peindre et de n'avoir jamais osé le dire..

Encore une fois, **mon** [u-MENTION-2935] père ne dit pas non, à condition que **je** [u-MENTION-2936] continuasse d'apprendre chez nous ce que **j'** [u-MENTION-2939] aurais dû apprendre au collège, mais avec la liberté de peindre. . [phrase]

À force de penser à Marthe, **j'** [u-MENTION-2946] y pensai de moins en moins.. [phrase] [phrase] [phrase]

J' [u-MENTION-2956] avais même pris goût au travail..

Je [u-MENTION-2958] n'avais pas menti comme **je** [u-MENTION-2959] le craignais. .

Lorsque quelque chose, venu de l'extérieur, m'obligeait à penser moins paresseusement à Marthe, **j'** [u-MENTION-2966] y pensais sans amour, avec la mélancolie que l'on éprouve pour ce qui aurait pu être.. [phrase]

me disais **-je** [u-MENTION-2973] , c'eût été trop beau.. [phrase]

» Une chose étonnait **mon** [u-MENTION-2979] père.. [phrase]

Il **me** [u-MENTION-2982] fît à ce sujet sa première scène, croyant que **j'** [u-MENTION-2986] avais soustrait la lettre, que **j'** [u-MENTION-2988] avais feint ensuite de lui annoncer gratuitement la nouvelle, que **j'** [u-MENTION-2992] avais ainsi obtenu son indulgence.. [phrase]

Je [u-MENTION-2997] me croyais renvoyé du collège, mais **je** [u-MENTION-2999] me trompais..

Aussi, **mon** [u-MENTION-3001] père ne comprit -il rien lorsque, au début des vacances, nous reçûmes une lettre du proviseur. .

Il demandait si **j'** [u-MENTION-3009] étais malade et s'il fallait **m'** [u-MENTION-3010] inscrire pour l'année suivante. .

La joie de donner enfin satisfaction à **mon** [u-MENTION-3014] père comblait un peu le vide sentimental dans lequel je me trouvais car, si **je** [u-MENTION-3017] croyais ne plus aimer Marthe, **je** [u-MENTION-3019] la considérais du moins comme le seul amour qui eût été digne de **moi** [u-MENTION-3023] ..

C'est dire que **je** [u-MENTION-3024] l'aimais encore. .

J' [u-MENTION-3026] étais dans ces dispositions de cœur quand, à la fin de novembre, un mois après avoir reçu une lettre de faire-part de son mariage, **je** [u-MENTION-3036] trouvais, en rentrant chez nous, une invitation de Marthe qui commençait par ces lignes : « Je ne comprends rien à **votre** [u-MENTION-3044] silence..

Pourquoi ne venez **-vous** [u-MENTION-3045] pas me voir??

Sans doute avez **-vous** [u-MENTION-3047] oublié que **vous** [u-MENTION-3048] avez choisi mes meubles?? [phrase] [phrase] [phrase]

Je [u-MENTION-3059] m'étonnai que la sienne fût si grande.. [phrase]

Quand **j'** [u-MENTION-3067] arrivais pour goûter, il faisait déjà nuit.. [phrase]

À voir cette fenêtre illuminée par des flammes inégales, comme des vagues, **je** [u-MENTION-3075] crus à un commencement d'incendie.. [phrase]

Je [u-MENTION-3080] m'étonnai d'une semblable négligence..

Je [u-MENTION-3082] cherchais la sonnette : **je** [u-MENTION-3084] ne la trouvais point..

Enfin, gravissant les trois marches du perron, **je** [u-MENTION-3088] me décidai à frapper contre les vitres du rez-de-chaussée de droite, derrière lesquelles **j'** [u-MENTION-3092] entendais des voix..

Une vieille femme ouvrit la porte : **je** [u-MENTION-3096] lui demandai où demeurait Mme Lacombe (tel était le nouveau nom de Marthe) : « C'est au-dessus..

» **Je** [u-MENTION-3104] montai l'escalier dans le noir, trébuchant, me cognant, et mourant de crainte qu'il fût arrivé quelque malheur..

Je [u-MENTION-3108] frappai..

C'est Marthe qui vint **m'** [u-MENTION-3110] ouvrir..

Je [u-MENTION-3111] faillis lui sauter au cou, comme les gens qui se connaissent à peine, après avoir échappé au naufrage.. [phrase]

Sans doute **me** [u-MENTION-3120] trouva-t-elle l'air égaré, car, avant toute chose, **je** [u-MENTION-3122] lui demandai pourquoi « il y avait le feu ».. .

— C'est qu'en **vous** [u-MENTION-3125] attendant, j'avais fait dans la cheminée du salon un feu de bois d'olivier, à la lueur duquel je lisais. .

En entrant dans la petite chambre qui lui servait de salon, peu encombrée de meubles, et que les tentures, les gros tapis doux comme un poil de bête, rétrécissaient jusqu'à lui donner l'aspect d'une boîte, **je** [u-MENTION-3149] fus à la fois heureux et malheureux comme un dramaturge qui, voyant sa pièce, y découvre trop tard des fautes. . [phrase]

— **Vous** [u-MENTION-3162] n'aimez peut-être pas l'odeur de l'olivier?? [phrase]

Marthe semblait s'excuser d'un détail de son cru, dans cette chambre qui était **mon** [u-MENTION-3181] œuvre.. [phrase] [phrase]

Ce feu **me** [u-MENTION-3188] ravit, et aussi de voir qu'elle attendait comme **moi** [u-MENTION-3190] de se sentir brûlante d'un côté, pour se retourner de l'autre..

Son visage calme et sérieux ne **m'** [u-MENTION-3195] avait jamais paru plus beau que dans cette lumière sauvage.. [phrase] [phrase] [phrase] [phrase]

Mon [u-MENTION-3210] esprit s'engourdissait peu à peu auprès d'elle, **je** [u-MENTION-3212] la trouvais différente..

C'est que, maintenant que **j'** [u-MENTION-3215] étais sûr de ne plus l'aimer, **je** [u-MENTION-3217] commençais à l'aimer..

Je [u-MENTION-3219] me sentais incapable de calculs, de machinations, de tout ce dont, jusqu'alors, et encore à ce moment -là, **je** [u-MENTION-3225] croyais que l'amour ne peut se passer..

Tout à coup, **je** [u-MENTION-3228] me sentais meilleur..

Ce brusque changement aurait ouvert les yeux de tout autre : **je** [u-MENTION-3232] ne vis pas que **j'** [u-MENTION-3233] étais amoureux de Marthe..

Au contraire, **j'** [u-MENTION-3235] y vis la preuve que **mon** [u-MENTION-3239] amour était mort, et qu'une belle amitié le remplacerait..

Cette longue perspective d'amitié **me** [u-MENTION-3244] fit admettre soudain combien un autre sentiment eût été criminel, lésant un homme qui l'aimait, à qui elle devait appartenir, et qui ne pouvait la voir. .

Pourtant, autre chose **m'** [u-MENTION-3254] aurait dû renseigner sur **mes** [u-MENTION-3256] véritables sentiments..

Il y a quelques mois, quand **je** [u-MENTION-3257] rencontrais Marthe, **mon** [u-MENTION-3260] prétendu amour ne **m'** [u-MENTION-3261] empêchait pas de la juger, de trouver laides la plupart des choses qu'elle trouvait belles, la plupart des choses qu'elle disait, enfantines.. Aujourd'hui, si **je** [u-MENTION-3270] ne pensais pas comme elle, **je** [u-MENTION-3272] me donnais tort..

Après la grossièreté de **mes** [u-MENTION-3276] premiers désirs, c'était la douceur d'un sentiment plus profond qui **me** [u-MENTION-3279] trompait..

Je [u-MENTION-3280] ne me sentais plus capable de rien entreprendre de ce que **je** [u-MENTION-3282] m'étais promis..

Je [u-MENTION-3283] commençais à respecter Marthe, parce que **je** [u-MENTION-3285] commençais à l'aimer. .

Je [u-MENTION-3287] revins tous les soirs ; **je** [u-MENTION-3289] ne pensai même pas à la prier de **me** [u-MENTION-3291] montrer sa chambre, encore moins à lui demander comment Jacques trouvait nos meubles..

Je [u-MENTION-3298] ne souhaitais rien d'autre que ces fiançailles éternelles, nos corps étendus près de la cheminée, se touchant l'un l'autre, et **moi** [u-MENTION-3306] , n'osant bouger, de peur qu'un seul de **mes** [u-MENTION-3309] gestes suffît à chasser le bonheur. .

[phrase]

Dans **ma** [u-MENTION-3316] paresse heureuse, elle lut de l'indifférence..

Pensant que **je** [u-MENTION-3319] ne l'aimais pas, elle s'imagina que **je** [u-MENTION-3322] me laisserais vite de ce salon silencieux, si elle ne faisait rien pour **m'** [u-MENTION-3325] attacher à elle. . [phrase]

J' [u-MENTION-3328] y voyais une preuve du bonheur. .

Je [u-MENTION-3332] me sentais tellement près de **Marthe** [u-MENTION-3334] , avec la certitude que nous pensions en même temps aux mêmes choses, que lui parler **m'** [u-MENTION-3339] eût semblé absurde, comme de parler haut quand on est seul.. [phrase]

La sagesse eût été de **me** [u-MENTION-3344] servir de moyens de correspondre aussi grossiers que la parole ou le geste, tout en déplorant qu'il n'en existât point de plus subtils. .

À **me** [u-MENTION-3350] voir tous les jours m'enfoncer de plus en plus dans ce mutisme délicieux, Marthe se figura que **je** [u-MENTION-3354] m'ennuyais de plus en plus..

Elle se sentait prête à tout pour **me** [u-MENTION-3358] distraire. . [phrase]

Ou plutôt **je** [u-MENTION-3364] croyais qu'elle dormait..

Son sommeil lui était prétexte, pour mettre ses bras autour de **mon** [u-MENTION-3373] cou, et une fois réveillée, les yeux humides, **me** [u-MENTION-3376] dire qu'elle venait d'avoir un rêve triste..

Elle ne voulait jamais **me** [u-MENTION-3381] le raconter..

Je [u-MENTION-3383] profitais de son faux sommeil pour respirer ses cheveux, son cou, ses joues brûlantes, et en les effleurant à peine pour qu'elle ne se réveillât point ; toutes caresses qui ne sont pas, comme on croit, la menue monnaie de l'amour, mais, au contraire, la plus rare, et auxquelles seule la passion puisse recourir..

Moi [u-MENTION-3402] . **Je** [u-MENTION-3403] les croyais permises à **mon** [u-MENTION-3406] amitié..

Pourtant, **je** [u-MENTION-3407] commençai à me désespérer sérieusement de ce que seul l'amour nous donnât des droits sur une femme..

Je [u-MENTION-3414] me passerai bien de l'amour, pensai -**je** [u-MENTION-3416] , mais jamais de n'avoir aucun droit sur Marthe..

Et, pour en avoir, **j'** [u-MENTION-3421] étais même décidé à l'amour, tout en croyant le déplorer..

Je [u-MENTION-3424] désirais Marthe et ne le comprenais pas. .

Quand elle dormait ainsi, sa tête appuyée contre un de **mes** [u-MENTION-3433] bras, **je** [u-MENTION-3434] me penchais sur elle pour voir son visage entouré de flammes.. [phrase]

Un jour que **je** [u-MENTION-3441] m'approchais trop sans pourtant que **mon** [u-MENTION-3443] visage touchât le sien, **je** [u-MENTION-3445] fus comme l'aiguille qui dépasse d'un millimètre la zone interdite et appartient à l'aimant.. [phrase]

C'est ainsi que **je** [u-MENTION-3454] sentis **mes** [u-MENTION-3456] lèvres contre les siennes.. [phrase]

Je [u-MENTION-3462] l'embrassai, stupéfait de **mon** [u-MENTION-3465] audace, alors qu'en réalité c'était elle qui, lorsque **j'** [u-MENTION-3468] approchais de son visage, avait attiré **ma** [u-MENTION-3472] tête contre sa bouche..

Ses deux mains s'accrochaient à **mon** [u-MENTION-3478] cou ; elles ne se seraient pas accrochées plus furieusement dans un naufrage..

Et **je** [u-MENTION-3482] ne comprenais pas si elle voulait que **je** [u-MENTION-3484] la sauve, ou bien que **je** [u-MENTION-3486] me noie avec elle. .

Maintenant, elle s'était assise, elle tenait **ma** [u-MENTION-3492] tête sur ses genoux, caressant **mes** [u-MENTION-3496] cheveux, et **me** [u-MENTION-3497] répétant très doucement : « Il faut que **tu** [u-MENTION-3499] t'en ailles, il ne faut plus jamais revenir..

» **Je** [u-MENTION-3501] n'osais pas la tutoyer ; lorsque **je** [u-MENTION-3503] ne pouvais plus me taire, **je** [u-MENTION-3504] cherchais longuement **mes** [u-MENTION-3507] mots, construisant **mes** [u-MENTION-3509] phrases de façon à ne pas lui parler directement, car si **je** [u-MENTION-3511] ne pouvais pas la tutoyer, **je** [u-MENTION-3513] sentais combien il était encore plus impossible de lui dire vous..

Mes [u-MENTION-3517] larmes **me** [u-MENTION-3518] brûlaient..

S'il en tombait une sur la main de Marthe, **je** [u-MENTION-3523] m'attendais toujours à l'entendre pousser un cri..

Je [u-MENTION-3527] m'accusai d'avoir rompu le charme, me disant qu'en effet **j'** [u-MENTION-3529] avais été fou de poser **mes** [u-MENTION-3531] lèvres contre les siennes, oubliant que c'était elle qui **m'** [u-MENTION-3534] avait embrassé..

« Il faut que **tu** [u-MENTION-3535] t'en ailles, ne plus jamais revenir..

» **Mes** [u-MENTION-3538] larmes de rage se mêlaient à **mes** [u-MENTION-3541] larmes de peine.. [phrase]

Si **j'** [u-MENTION-3549] avais parlé, ç'aurait été pour injurier Marthe.. [phrase]

« Puisqu'il est trop tard, la faisais -**je** [u-MENTION-3557] penser, dans **mon** [u-MENTION-3559] injustice peut-être clairvoyante, après tout, **j'** [u-MENTION-3560] aime autant qu'il [u-MENTION-3561] souffre..

» Dans ce feu, **je** [u-MENTION-3563] grelottais, **je** [u-MENTION-3564] claquais des dents..

À **ma** [u-MENTION-3567] véritable peine qui **me** [u-MENTION-3569] sortait de l'enfance, s'ajoutaient des sentiments enfantins..

J' [u-MENTION-3572] étais le spectateur qui ne veut pas s'en aller parce que le dénouement lui déplait..

Je [u-MENTION-3577] lui dis : « **Je** [u-MENTION-3579] ne m'en irai pas..

Vous vous êtes moquée de **moi** [u-MENTION-3581] ..

Je [u-MENTION-3582] ne veux plus vous voir..

» Car si **je** [u-MENTION-3584] ne voulais pas rentrer chez **mes** [u-MENTION-3586] parents, **je** [u-MENTION-3587] ne voulais pas non plus revoir Marthe..

Je [u-MENTION-3589] l'aurais plutôt chassée de chez elle !

Mais elle sanglotait : « **Tu** [u-MENTION-3593] es un enfant..

Tu [u-MENTION-3595] ne comprends donc pas que si je **te** [u-MENTION-3597] demande de t'en aller, c'est que je **t'** [u-MENTION-3599] aime..
» Haineusement, **je** [u-MENTION-3601] lui dis que **je** [u-MENTION-3603] comprenais fort bien qu'elle avait des devoirs et que son mari était à la guerre. .
Elle secouait la tête : « Avant **toi** [u-MENTION-3612] , j'étais heureuse, je croyais aimer mon fiancé.. [phrase]
C'est **toi** [u-MENTION-3620] qui m'as montré que je ne t'aimais pas..
Mon devoir n'est pas celui que **tu** [u-MENTION-3628] penses..
Ce n'est pas de ne pas mentir à mon mari, mais de ne pas **te** [u-MENTION-3632] mentir..
Va-t'en et ne me crois pas méchante ; bientôt **tu** [u-MENTION-3634] m'auras oubliée..
Mais je ne veux pas causer le malheur de **ta** [u-MENTION-3639] vie..
Je pleure, parce que je suis trop vieille pour **toi** [u-MENTION-3642] !! [phrase]
Et, quelles que soient les passions que **j'** [u-MENTION-3650] éprouve dans la suite, jamais ne sera plus possible l'émotion adorable de voir une fille de dix-neuf ans pleurer parce qu'elle se trouve trop vieille. .
La saveur du premier baiser **m'** [u-MENTION-3659] avait déçu comme un fruit que l'on goûte pour la première fois.. [phrase]
Quelques minutes après, non seulement **j'** [u-MENTION-3670] étais habitué à la bouche de Marthe, mais encore **je** [u-MENTION-3673] ne pouvais plus m'en passer..
Et c'est alors qu'elle parlait de **m'** [u-MENTION-3676] en priver à tout jamais. .
Ce soir -là, Marthe **me** [u-MENTION-3681] reconduisit jusqu'à la maison..
Pour **me** [u-MENTION-3683] sentir plus près d'elle, **je** [u-MENTION-3686] me blottissais sous cape, et **je** [u-MENTION-3688] la tenais par la taille.. [phrase]
Elle **me** [u-MENTION-3697] faisait lui jurer mille folies. .
Devant la maison de **mes** [u-MENTION-3703] parents, **je** [u-MENTION-3704] ne voulus pas laisser Marthe repartir seule, et l'accompagnai jusque chez elle..
Sans doute ces enfantillages n'eussent -ils jamais pris fin, car elle voulait **m'** [u-MENTION-3713] accompagner encore..
J' [u-MENTION-3714] acceptai, à condition qu'elle **me** [u-MENTION-3716] laisserait à moitié route. .
J' [u-MENTION-3718] arrivai une demi-heure en retard pour le dîner.. [phrase]
Je [u-MENTION-3724] mis ce retard sur le compte du train..
Mon [u-MENTION-3729] père fit semblant de le croire. .
Plus rien ne **me** [u-MENTION-3732] pesait..
Dans la rue, **je** [u-MENTION-3734] marchais aussi légèrement que dans **mes** [u-MENTION-3736] rêves. .
Jusqu'ici tout ce que **j'** [u-MENTION-3739] avais convoité, enfant, il en avait fallu faire **mon** [u-MENTION-3743] deuil..
D'autre part, la reconnaissance **me** [u-MENTION-3745] gâtait les jouets offerts.. [phrase]
J' [u-MENTION-3752] étais ivre de passion..
Marthe était à moi ; ce n'est pas **moi** [u-MENTION-3757] **qui** [u-MENTION-3758] l'avais dit, c'était elle..
Je [u-MENTION-3762] pouvais toucher sa figure, embrasser ses yeux, ses bras, l'habiller, l'abîmer, à **ma** [u-MENTION-3773] guise..
Dans **mon** [u-MENTION-3775] délire, **je** [u-MENTION-3776] la mordais aux endroits où sa peau était nue, pour que sa mère la soupçonnât d'avoir un amant..
J' [u-MENTION-3786] aurais voulu pouvoir y marquer **mes** [u-MENTION-3789] initiales..
Ma [u-MENTION-3791] sauvagerie d'enfant retrouvait le vieux sens des tatouages.. [phrase]
» **J'** [u-MENTION-3800] aurais voulu pouvoir embrasser ses seins..
Je [u-MENTION-3803] n'osais pas le lui demander, pensant qu'elle saurait les offrir elle -même, comme ses lèvres..
Au bout de quelques jours, l'habitude d'avoir ses lèvres étant venue, **je** [u-MENTION-3814] n'envisageai pas d'autre délice. . [phrase]
[phrase] [phrase]
Je [u-MENTION-3833] ne voyais pas flamber ces lettres sans malaise..
Elles grandissaient une seconde le feu et, somme toute, **j'** [u-MENTION-3839] avais peur de voir plus clair. .
Marthe, qui souvent maintenant **me** [u-MENTION-3843] demandait s'il était vrai que **je** [u-MENTION-3844] l'avais aimée dès notre première rencontre, **me** [u-MENTION-3847] reprochait de ne le lui avoir pas dit avant son mariage.. [phrase] [phrase] [phrase] [phrase] [phrase]
[phrase] [phrase] [phrase] [phrase] [phrase] [phrase]
Les jours où elle ne **m'** [u-MENTION-3921] accompagnait pas à Paris, Marthe, nue sous un peignoir, attendait que **je** [u-MENTION-3925] revinsse de **mes** [u-MENTION-3927] cours de dessin, étendue devant la cheminée où brûlait toujours l'olivier de ses beaux-parents..
[phrase]
Je [u-MENTION-3940] ne sais quelle timidité, si ce n'est celle que l'on éprouve en face de ce qu'on n'a jamais fait, **me** [u-MENTION-3949] retenait..
Je [u-MENTION-3950] pensais à Daphnis.. [phrase]
Au fait, ne considérais -**je** [u-MENTION-3959] pas Marthe plutôt comme une vierge, livrée, la première quinzaine de ses noces, à un inconnu et plusieurs fois prise par lui de force. .
Le soir, seul dans **mon** [u-MENTION-3971] lit, **j'** [u-MENTION-3972] appelais Marthe, m'en voulant, **moi** [u-MENTION-3974] **qui** [u-MENTION-3975] me croyais un homme, de ne l'être pas assez pour finir d'en faire **ma** [u-MENTION-3980] maîtresse..
Chaque jour, allant chez elle, **je** [u-MENTION-3983] me promettais de ne pas sortir qu'elle ne le fût. Le jour de l'anniversaire de **mes** [u-

MENTION-3989] seize ans, au mois de mars 1918, tout en **me** [u-MENTION-3992] suppliant de ne pas me fâcher, elle **me** [u-MENTION-3994] fit cadeau d'un peignoir, semblable au sien, qu'elle voulait **me** [u-MENTION-4000] voir mettre chez elle..

Dans **ma** [u-MENTION-4003] joie, **je** [u-MENTION-4004] faillis faire un calembour, **moi** [u-MENTION-4006] **qui** [u-MENTION-4007] n'en faisais jamais..

Ma [u-MENTION-4011] robe prétexte!!

Car il **me** [u-MENTION-4012] semblait que ce qui jusqu'ici avait entravé **mes** [u-MENTION-4016] désirs, c'était la peur du ridicule, de **me** [u-MENTION-4020] sentir habillé, lorsqu'elle ne l'était pas..

D'abord **je** [u-MENTION-4023] pensai à mettre cette robe le jour même..

Puis, **je** [u-MENTION-4026] rougis, comprenant ce que son cadeau contenait de reproches. .

Dès le début de notre amour, Marthe **m'** [u-MENTION-4035] avait donné une clef de son appartement, afin que **je** [u-MENTION-4039] n'eusse pas à l'attendre dans le jardin, si, par hasard, elle était en ville..

Je [u-MENTION-4044] pouvais me servir moins innocemment de cette clef.. [phrase]

Je [u-MENTION-4048] quittai Marthe en lui promettant de venir déjeuner le lendemain avec elle..

Mais **j'** [u-MENTION-4053] étais décidé à revenir le soir aussitôt que possible. .

À dîner, **j'** [u-MENTION-4055] annonçai à **mes** [u-MENTION-4057] parents que **j'** [u-MENTION-4058] entreprendrais le lendemain avec René une longue promenade dans la forêt de Sénart..

Je [u-MENTION-4064] devais pour cela partir à cinq heures du matin..

Comme toute la maison dormirait encore, personne ne pourrait deviner l'heure à laquelle **j'** [u-MENTION-4071] étais parti, et si **j'** [u-MENTION-4072] avais découché. .

À peine avais -**je** [u-MENTION-4073] fait part de ce projet à **ma** [u-MENTION-4076] mère, qu'elle voulut préparer elle -même un panier rempli de provisions, pour la route..

J' [u-MENTION-4082] étais consterné, ce panier détruisait tout le romanesque et le sublime de **mon** [u-MENTION-4087] acte..

Moi [u-MENTION-4088] **qui** [u-MENTION-4089] goûtais d'avance l'effroi de Marthe quand **j'** [u-MENTION-4092] entrerais dans sa chambre, **je** [u-MENTION-4095] pensais maintenant à ses éclats de rire en voyant paraître ce Prince Charmant, un panier de ménagère à son bras..

J' [u-MENTION-4104] eus beau dire à **ma** [u-MENTION-4106] mère que René s'était muni de tout, elle ne voulut rien entendre.. [phrase]

Tandis que ma mère emplissait le panier qui **me** [u-MENTION-4118] gâtait d'avance **ma** [u-MENTION-4120] première nuit d'amour, **je** [u-MENTION-4122] voyais les yeux pleins de convoitise de **mes** [u-MENTION-4126] frères..

Je [u-MENTION-4127] pensai bien à le leur offrir en cachette, mais une fois tout mangé, au risque de se faire fouetter, et pour le plaisir de **me** [u-MENTION-4133] perdre, ils eussent tout raconté. .

Il fallait donc **me** [u-MENTION-4135] résigner, puisque nulle cachette ne semblait assez sûre. .

Je [u-MENTION-4137] m'étais juré de ne pas partir avant minuit pour être sûr que **mes** [u-MENTION-4141] parents dormissent..

J' [u-MENTION-4142] essayai de lire..

Mais comme dix heures sonnaient à la mairie, et que **mes** [u-MENTION-4146] parents étaient couchés depuis quelque temps déjà, **je** [u-MENTION-4148] ne pus attendre..

Ils habitaient au premier étage, **moi** [u-MENTION-4151] au rez-de-chaussée..

Je [u-MENTION-4153] n'avais pas mis **mes** [u-MENTION-4155] bottines afin d'escalader le mur le plus silencieusement possible..

Les tenant d'une main, tenant de l'autre ce panier fragile à cause des bouteilles, **j'** [u-MENTION-4163] ouvris avec précaution une petite porte d'office.. [phrase] [phrase] [phrase]

Apercevant que la lumière n'était pas encore éteinte dans la chambre de **mes** [u-MENTION-4172] parents, **je** [u-MENTION-4173] fus sur le point de me recoucher..

Mais **j'** [u-MENTION-4174] étais en route..

Déjà la précaution des bottines était impossible ; à cause de la pluie **je** [u-MENTION-4178] dus les remettre..

Ensuite, il **me** [u-MENTION-4180] fallait escalader le mur pour ne point ébranler la cloche de la grille..

Je [u-MENTION-4184] m'approchai du mur, contre lequel **j'** [u-MENTION-4187] avais pris soin, après le dîner, de poser une chaise de jardin pour faciliter **mon** [u-MENTION-4193] évaison.. [phrase] [phrase]

Comme **je** [u-MENTION-4200] m'y suspendais, l'une d'elles tomba..

Mon [u-MENTION-4205] angoisse décupla le bruit de sa chute.. [phrase]

Je [u-MENTION-4211] tenais le panier avec **mes** [u-MENTION-4214] dents ; **je** [u-MENTION-4215] tombai dans une flaque..

Une longue minute, **je** [u-MENTION-4218] restai debout, les yeux levés vers la fenêtre de **mes** [u-MENTION-4222] parents, pour voir s'ils bougeaient, s'étant aperçus de quelque chose.. [phrase]

J' [u-MENTION-4226] étais sauf! !

Pour me rendre jusque chez Marthe, **je** [u-MENTION-4228] suivis la Marne..

Je [u-MENTION-4230] comptais cacher **mon** [u-MENTION-4232] panier dans un buisson et le reprendre le lendemain.. [phrase] [phrase]

J' [u-MENTION-4246] hésitai longtemps, plus pâle qu'un homme qui pose une cartouche de dynamite..

Je [u-MENTION-4252] cachai tout de même **mes** [u-MENTION-4254] victuailles. . [phrase]

Je [u-MENTION-4257] pris la clef qu'on laissait toujours dans la boîte aux lettres..

Je [u-MENTION-4263] traversai le petit jardin sur la pointe des pieds, puis montai les marches du perron..
J' [u-MENTION-4269] ôtai encore **mes** [u-MENTION-4271] bottines avant de prendre l'escalier. . [phrase]
Peut-être s'évanouirait -elle en **me** [u-MENTION-4276] voyant dans sa chambre..
Je [u-MENTION-4279] tremblai ; **je** [u-MENTION-4280] ne trouvai pas le trou de la serrure..
Enfin, **je** [u-MENTION-4283] tournai la clef lentement, afin de ne réveiller personne..
Je [u-MENTION-4286] butai dans l'antichambre contre le porte-parapluies..
Je [u-MENTION-4289] craignais de prendre les sonnettes pour des commutateurs..
J' [u-MENTION-4292] allai à tâtons jusqu'à la chambre..
Je [u-MENTION-4294] m'arrêtai avec, encore, l'envie de fuir..
Peut-être Marthe ne **me** [u-MENTION-4297] pardonnerait jamais..
Ou bien si **j'** [u-MENTION-4299] allais tout à coup apprendre qu'elle **me** [u-MENTION-4302] trompe, et la trouver avec un homme! !
J' [u-MENTION-4305] ouvris..
Je [u-MENTION-4306] murmurai : — Marthe? ? [phrase] [phrase]
Elle **me** [u-MENTION-4318] prenait pour Jacques! !
Or, si **je** [u-MENTION-4320] voyais de quelle façon elle l'eût accueilli, **j'** [u-MENTION-4324] apprenais du même coup qu'elle **me** [u-MENTION-4327] cachait déjà quelque chose.. [phrase]
J' [u-MENTION-4331] allumai.. [phrase]
Il était simple de dire : « **C'** [u-MENTION-4334] est **moi** [u-MENTION-4335] », et pourtant, **je** [u-MENTION-4336] ne le disais pas..
Je [u-MENTION-4338] l'embrassai dans le cou. . [phrase] [phrase] [phrase]
D'une seconde à l'autre, elle changea d'attitude et, sans prendre la peine de s'expliquer **ma** [u-MENTION-4351] présence nocturne : —
Mais **mon pauvre chéri** [u-MENTION-4352] , **tu** [u-MENTION-4354] vas prendre mal!!
Déshabille **-toi** [u-MENTION-4356] vite. . [phrase]
À son retour dans la chambre, comme **je** [u-MENTION-4363] ne bougeais pas, elle dit : — Veux **-tu** [u-MENTION-4365] que je t'aide? ?
Moi [u-MENTION-4367] **qui** [u-MENTION-4368] redoutais par-dessus tout le moment où **je** [u-MENTION-4371] devrais me déshabiller et **qui** [u-MENTION-4372] en envisageais le ridicule, **je** [u-MENTION-4375] bénissais la pluie grâce à quoi ce déshabillage prenait un sens maternel..
Mais Marthe repartait, revenait, repartait dans la cuisine, pour voir si l'eau de **mon** [u-MENTION-4384] grog était chaude..
Enfin, elle **me** [u-MENTION-4386] trouva nu sur le lit, me cachant à moitié sous l'édredon..
Elle **me** [u-MENTION-4390] gronda : c'était fou de rester nu ; il fallait me frictionner à l'eau de Cologne. .
Puis, Marthe ouvrit une armoire et **me** [u-MENTION-4395] jeta un costume de nuit..
« Il devait être de **ma** [u-MENTION-4400] taille.. [phrase]
Et **je** [u-MENTION-4403] pensais à l'arrivée, fort possible, de ce soldat, puisque Marthe y avait cru. .
J' [u-MENTION-4408] étais dans le lit..
Marthe **m'** [u-MENTION-4411] y rejoignit..
Je [u-MENTION-4413] lui demandai d'éteindre..
Car, même en ses bras, **je** [u-MENTION-4417] me méfiais de **ma** [u-MENTION-4419] timidité..
Les ténèbres **me** [u-MENTION-4421] donneraient du courage..
Marthe **me** [u-MENTION-4424] répondit doucement : — Non..
Je veux **te** [u-MENTION-4427] voir t'endormir.. <html>

B.2 articleswiki_RC_19–21_s–CHAINED–1353

[articleswiki_RC_19-21_s-CHAINE-1353] - id : s-CHAINE-1353, nb_maillons : 391 mentions, from : w_export_6682, to : w_export_10621

Ils se trouvent principalement en **3 Afrique centrale et australe** [u-MENTION-1981] .. [phrase] Bien que **4 la phylogénie des équidés** [u-MENTION-1984] soit peu connue, ce groupe est manifestement paraphylétique, c'est-à-dire que si tous ces animaux descendent bien d'**5 une espèce commune** [u-MENTION-1988] , toutes les espèces descendantes de celle -ci ne sont pas que des zèbres.. Il y a aussi des chevaux et des ânes qui sont plus ou moins proches de **2 chacune de ces espèces** [u-MENTION-1996] . . **6 Description** [u-MENTION-1998] Mammifères terrestres herbivores, **7 les caractéristiques générales des zèbres** [u-MENTION-1999] sont **8 celles des équidés du genre Equus** [u-MENTION-2001] , avec **9 des nuances comportementales et physiologiques** [u-MENTION-2004] pour **10 chaque espèce** [u-MENTION-2005] . . Les zèbres communs mesurent de **11 1, 10 mètre** [u-MENTION-2007] à **12 1, 40 mètre** [u-MENTION-2008] (**13 1, 25 mètre** [u-MENTION-2009] en moyenne) au garrot contre **15 1, 30 mètre** [u-MENTION-2011] à **16 1, 60 mètre** [u-MENTION-2012] (**17 1, 45 mètre** [u-MENTION-2013] en moyenne) au garrot pour les zèbres de Grévy, et vivent en moyenne **14 25** [u-MENTION-2017] à **18 30 ans** [u-MENTION-2018] dans la nature et jusqu'à **19 40 ans** [u-MENTION-2020] dans un zoo.. **20 La longueur du corps** [u-MENTION-2022] va de **21 2, 20 mètres** [u-MENTION-2024] à **22 2, 70 mètres** [u-MENTION-2025] pour les zèbres communs et de **23 2, 50 mètres** [u-MENTION-2027] à **24 3, 00 mètres** [u-MENTION-2028] pour les zèbres de Grévy et **25 la longueur de la queue** [u-MENTION-2030] de **26 40** [u-MENTION-2032] à **27 75 cm** [u-MENTION-2033] .. **28 La masse** [u-MENTION-2034] varie de **29 175** [u-MENTION-2035] à **30 300 kg** [u-MENTION-2036] pour les zèbres communs contre **31 300** [u-MENTION-2038] à **32 400 kg** [u-MENTION-2039] pour les zèbres de Grévy.. [phrase] On compte en **33 2016** [u-MENTION-2045] **34 moins de 800 000 zèbres** [u-MENTION-2046] au total. . [phrase] Grâce à **35 des méthodes de 36 reconnaissance de 37 formes** [u-MENTION-2052] , les scientifiques peuvent désormais lire les rayures caractéristiques des zèbres comme **39 des codes-barres** [u-MENTION-2058] pour recenser **40 une population** [u-MENTION-2059] à partir de **41 photographies** [u-MENTION-2060] .. Après avoir pris la photo d'un individu, **44 les chercheurs** [u-MENTION-2063] la transfèrent sur **42 un ordinateur équipé de 45 logiciels dédiés** [u-MENTION-2065] comme StripeSpotter, un logiciel mis au point par **52 47 l'université de 48 l'Illinois à 49 Chicago et 50 l'université de 51 Princeton** [u-MENTION-2068] .. Ils zooment ensuite sur le flanc de l'animal, où **38 chaque rayure** [u-MENTION-2078] est décomposée en **43 lignes verticales de pixels** [u-MENTION-2079] .. **53 Leurs combinaisons** [u-MENTION-2081] sont aussi uniques que **54 les empreintes digitales humaines** [u-MENTION-2083] .. Reste à faire **55 une recherche** [u-MENTION-2084] sur **56 la base de 57 données** [u-MENTION-2085] pour voir si le zèbre est **58 un nouveau venu** [u-MENTION-2088] .. StripeSpotter a été utilisé sur **60 46 des zèbres de Grévy et 59 des zèbres des plaines** [u-MENTION-2090] .. **61 D'autres travaux** [u-MENTION-2093] étendent **62 la reconnaissance** [u-MENTION-2094] à **63 d'autres animaux** [u-MENTION-2095] comme **67 64 les léopards, 65 les girafes et 66 les Pterois** [u-MENTION-2096] . . **68 Une légende africaine** [u-MENTION-2100] demande si le zèbre est blanc à rayures noires ou noir à rayures blanches.. **69 Cette question** [u-MENTION-2104] a généré **72 70 de nombreuses légendes ou 71 réponses fantaisistes** [u-MENTION-2105] , mais **73 des explications plus scientifiques** [u-MENTION-2108] existent. . **74 Formation des rayures** [u-MENTION-2109] **75 Les premières populations de zèbres** [u-MENTION-2111] étaient de **76 couleur gris-ardoise** [u-MENTION-2113] .. **77 Les spécialistes** [u-MENTION-2114] pensent généralement, en observant **78 les rayures partielles du Quagga** [u-MENTION-2115] et en tenant compte de **79 la pigmentation nécessaire aux animaux** [u-MENTION-2117] pour survivre sous **81 le soleil d'Afrique** [u-MENTION-2119] , que les zèbres étaient originellement **80 des animaux pigmentés de noir** [u-MENTION-2122] et que les raies se forment par **83 inhibition de 84 la production de 85 mélanine** [u-MENTION-2124] . . Les raies noires et blanches du zèbre sont absentes **82 au stade fœtal initial** [u-MENTION-2129] , ils sont entièrement noirs.. Les rayures finissent par apparaître par bandes d'**86 environ quatre cents micromètres** [u-MENTION-2133] (**87 vingt fois** [u-MENTION-2134] **88 une cellule** [u-MENTION-2135]).. [phrase] [phrase] Selon **89 J. B. L. Bard** [u-MENTION-2140] , **90 les espèces de zèbres** [u-MENTION-2141] différeraient selon le stade embryonnaire auquel

apparaissent les raies. . En **91 1952** [u-MENTION-2146] , Alan Turing a démontré que « même si **92 la concentration initiale en 93 morphogène** [u-MENTION-2148] est uniforme, **94 la combinaison de 95 réactions chimiques et de 96 diffusion 97 des substances à travers 98 les tissus** [u-MENTION-2150] peut faire apparaître **otif un motif** [u-MENTION-2155] » et aussi que « **otif ce motif** [u-MENTION-2156] dépend **99 du type de 100 réactions impliquées, de 101 la forme de 102 la région et 103 des concentrations initiales** [u-MENTION-2157] ». . Ainsi l'évolution vers des rayures au lieu de taches n'est pas difficile et dépendra de **104 la taille** [u-MENTION-2166] et **105 du temps de 107 gestation de 108 l'espèce** [u-MENTION-2167] . . **109 Ceci** [u-MENTION-2170] conforte la théorie du Dr Debra Kay Bennett selon laquelle les espèces de zèbres sont, **110 chacune** [u-MENTION-2176] , plus proche d'**111 une espèce de cheval** [u-MENTION-2177] que de **112 ses consœurs** [u-MENTION-2179] car « il suffit d' **113 une petite modification 114 des relations temporelles des processus** [u-MENTION-2181] qui sous-tendent **115 la formation otif du motif** [u-MENTION-2185] » pour faire apparaître des rayures au lieu de taches.. **106 Ce qui a donc pu se produire indépendamment au cours de l'évolution 116 des différentes espèces devenues des zèbres** [u-MENTION-2189] . . **117 Rôle** [u-MENTION-2193] On ignore encore exactement quelle pourrait être l'utilité des rayures : **119 de nombreuses théories** [u-MENTION-2197] sont proposées mais **120 aucune** [u-MENTION-2198] n'est validée.. **121 Les principales hypothèses sur la fonction des zébrures** [u-MENTION-2199] concerneraient **124 le camouflage, 118 l'évasion face aux prédateurs, la thermorégulation et 123 les interactions sociales** [u-MENTION-2202] ; toutefois, selon **125 une étude multifactorielle publiée en 126 2013** [u-MENTION-2208] , rien ne vient étayer de **127 manière convaincante** [u-MENTION-2210] **130 128 l'une ou 129 l'autre de 131 ces hypothèses** [u-MENTION-2211] . . **132 Au 19e siècle** [u-MENTION-2215] , **133 l'écrivain Rudyard Kipling et le naturaliste Alfred Russel Wallace** [u-MENTION-2216] ont contribué à diffuser l'hypothèse selon laquelle les rayures du zèbre lui permettaient de mieux se fondre dans la savane.. **134 Cette hypothèse longtemps considérée comme crédible dans 135 la communauté scientifique** [u-MENTION-2225] a toutefois été démentie **136 au début 137 du 21e siècle** [u-MENTION-2227] .. L'hypothèse est formellement démentie en **138 2016** [u-MENTION-2230] .. En fait, dans la savane, le zèbre est très visible, et il tendrait donc à se dresser comme **139 une exception** [u-MENTION-2234] à **140 la règle du camouflage** [u-MENTION-2235] . . Les rayures auraient aussi **141 un effet stroboscopique** [u-MENTION-2238] sur les prédateurs.. Lorsque **142 tout un troupeau** [u-MENTION-2240] s'enfuit, **143 les raies 144 des divers individus** [u-MENTION-2241] se mélangent, rendant flou **145 le contour d'146 un animal** [u-MENTION-2243] **147 aux yeux d'un lion** [u-MENTION-2245] , par exemple.. **149 Un phénomène comparable** [u-MENTION-2247] serait à **150 l'origine 151 du camouflage Dazzle** [u-MENTION-2248] . . Dans **152 les années 1970** [u-MENTION-2250] , **153 des recherches** [u-MENTION-2251] ont pointé le fait que **154 la mouche tsé-tsé** [u-MENTION-2254] , responsable de **aladieSommeil la maladie 156 du sommeil** [u-MENTION-2255] à **aladieSommeil laquelle** [u-MENTION-2257] les zèbres sont plus sensibles que **155 d'autres animaux sauvages** [u-MENTION-2259] , est attirée par **157 la vue de 158 larges zones monochromes** [u-MENTION-2260] : les rayures permettraient ainsi de se protéger du parasite.. Il est à **160 ce sujet significatif** [u-MENTION-2264] de constater que **164 161 les zones de 162 répartition des zèbres et de 163 ces glossines** [u-MENTION-2265] coïncident exactement et que les rayures s'estompent chez **165 les populations moins exposées au parasitisme par 167 les trypanosomes** [u-MENTION-2271] .. **159 Des travaux publiés en 166 2012** [u-MENTION-2274] viennent corroborer **168 cette hypothèse** [u-MENTION-2276] , en démontrant que **169 les taons** [u-MENTION-2277] sont plus attirés par **170 les monochromes** [u-MENTION-2278] , et que **171 l'effet « répulsif » des rayures** [u-MENTION-2279] est plus prononcé pour **172 des rayures semblables à celles des zèbres** [u-MENTION-2281] . . **173 Une autre hypothèse** [u-MENTION-2284] veut que les rayures contribueraient à la thermorégulation, permettant aux zèbres qui broutent pendant **122 des heures** [u-MENTION-2289] de mieux supporter **174 les fortes chaleurs de la savane africaine** [u-MENTION-2290] .. Les bandes noires et blanches, par **177 175 absorption et 176 réflexion différentielle 178 des rayons solaires** [u-MENTION-2293] , chauffent différemment, **179 ce qui provoquerait entre elles 180 un flux d'air différentiel à 181 l'origine de 182 tourbillons engendrant 183 un effet de 184 refroidissement** [u-MENTION-2297] .. **185 Ce dispositif** [u-MENTION-2304] leur permet d'avoir **186 une température corporelle**

inférieure à 187 celle d'188 herbivores de 189 taille similaire paissant dans 190 les mêmes conditions [u-MENTION-2306] (**191 29, 2 ° C** [u-MENTION-2311] contre **192 32, 5 ° C** [u-MENTION-2312]). . Les rayures favoriseraient **193 la cohésion sociale** [u-MENTION-2314] en facilitant **194 la reconnaissance** [u-MENTION-2315] et **195 l'identification de 196 chaque individu d'197 un groupe** [u-MENTION-2316] grâce **198 au dessin de 199 rayures unique** [u-MENTION-2319] .. Ainsi le zèbre de Burchell possède de **200 vingt-cinq** [u-MENTION-2322] à **201 trente raies** [u-MENTION-2323] sur **202 chacun de 203 ses flancs** [u-MENTION-2324] , le zèbre de Grévy **204 en** [u-MENTION-2328] compte environ **205 quatre-vingt** [u-MENTION-2329] et le zèbre des montagnes **206 quarante-trois** [u-MENTION-2331] . . **209 207 Comportement et 208 écologie** [u-MENTION-2332] Prédateurs **148 Les lions et les hyènes** [u-MENTION-2336] peuvent s'attaquer **210 aux adultes** [u-MENTION-2339] ; **213 211 les jeunes poulains et 212 les jeunes pré-adultes** [u-MENTION-2340] peuvent être **214 la proie 215 des lycaons, 216 guépards, 217 léopards** [u-MENTION-2343] . Les prédateurs sont opportunistes et s'attaquent **218 aux animaux vulnérables** [u-MENTION-2349] et donc peu rapides ; **219 les individus malades, âgés, blessés, isolés, jeunes ou 220 les femelles en 221 gestation** [u-MENTION-2350] , sont **222 les proies idéales** [u-MENTION-2353] . . Un zèbre en **224 bonne condition physique** [u-MENTION-2355] a **225 plusieurs moyens de défense contre 226 ses prédateurs** [u-MENTION-2356] , il possède **223 une très bonne vue diurne** [u-MENTION-2360] , **227 une ouïe excellente** [u-MENTION-2361] , **228 un bon odorat** [u-MENTION-2362] et court très vite.. Les zèbres peuvent ainsi tenir **229 une vitesse de 30 à 40 km / h** [u-MENTION-2365] sur **230 une très longue distance** [u-MENTION-2366] ou, en **231 cas de danger** [u-MENTION-2367] , galoper à **232 60 km / h** [u-MENTION-2368] en moyenne et même faire **233 des pointes** [u-MENTION-2369] à **234 80 km / h** [u-MENTION-2370] pour semer par exemple une lionne qui court presque aussi vite, mais ne tiendra pas **235 la distance** [u-MENTION-2374] .. Pour se défendre, ils peuvent aussi mordre et d'**236 un coup de 237 sabot** [u-MENTION-2376] , briser **238 la mâchoire d'239 une lionne** [u-MENTION-2378] .. Leurs ruades peuvent être mortelles, celles -ci sont encore plus puissantes que **240 celles d'un cheval** [u-MENTION-2383] .. Leurs rayures provoquent aussi **241 une sorte d'« 242 effet stroboscopique** [u-MENTION-2387] », et rendent **243 les individus** [u-MENTION-2390] difficiles à repérer lorsqu'ils courent en **244 groupe** [u-MENTION-2392] . . **245 Systématique** [u-MENTION-2393] Taxonomie **247 Les populations de zèbres** [u-MENTION-2395] varient beaucoup, et **248 les liens entre les espèces** [u-MENTION-2397] ne sont pas encore très bien compris.. [phrase] **246 Espèces habituellement citées dans 249 les classifications classiques** [u-MENTION-2401] : • Equus burchellii **250 Gray** [u-MENTION-2404] , **251 1824** [u-MENTION-2405] ou Zèbre de Burchell (zèbre des plaines, synonyme du suivant) • Equus quagga **256 Boddaert** [u-MENTION-2410] , **255 1785** [u-MENTION-2411] • Equus grevyi **254 Oustalet** [u-MENTION-2413] , **257 1882** [u-MENTION-2414] • Equus zebra **258 Linnaeus** [u-MENTION-2416] , **259 1758** [u-MENTION-2417] Les espèces sont notamment caractérisées par **260 un nombre différent de 261 paires de 262 chromosomes** [u-MENTION-2419] : **263 46** [u-MENTION-2422] pour Equus grevyi, **265 44** [u-MENTION-2424] pour **268 Equus quagga et Equus burchellii** [u-MENTION-2425] , et **269 32** [u-MENTION-2428] pour Equus zebra. . **264 Une nouvelle classification** [u-MENTION-2430] a été proposée en **266 2004** [u-MENTION-2431] par **267 les Anglais 270 C. P. Groves et 271 H. B. Bell** [u-MENTION-2432] , d'après **272 l'observation traditionnelle 273 du pelage et 274 des crânes de ces animaux** [u-MENTION-2435] : • **275 Sous-genre Dolichohippus** [u-MENTION-2439] avec **276 une seule espèce** [u-MENTION-2440] : • Equus grevyi • **277 Sous-genre Hippotigris** [u-MENTION-2442] avec **278 trois espèces** [u-MENTION-2443] : • Equus quagga, avec **279 six sous-espèces** [u-MENTION-2445] : • **280 Equus quagga quagga** [u-MENTION-2446] (éteint) • **281 Equus quagga burchellii** [u-MENTION-2447] • **282 Equus quagga boehmi** [u-MENTION-2448] , • **283 Equus quagga borensis** [u-MENTION-2449] • **284 Equus quagga chapmani** [u-MENTION-2450] • **285 Equus quagga crawshayi** [u-MENTION-2451] • Equus zebra • **287 Equus hartmannae** [u-MENTION-2453] **286 Espèces vivantes** [u-MENTION-2454] • Le zèbre de Burchell (Equus quagga, syn.. [phrase] Selon **252 les dernières recherches** [u-MENTION-2459] (**253 2004** [u-MENTION-2460]) il comporterait **288 six sous-espèces** [u-MENTION-2462] , réparties entre **290 289 l'Afrique de l'Est et du Sud** [u-MENTION-2463] .. Le « zèbre de Burchell vrai » désignerait en fait **291 une sous-espèce éteinte** [u-MENTION-2466] (Equus quagga burchelli)

ou n'ayant jamais existé ; • • Le zèbre des montagnes (*Equus zebra*), de **292 l'Afrique du Sud-Ouest** [u-MENTION-2470] , tend à avoir **293 une toison lisse** [u-MENTION-2471] , **294 un ventre blanc** [u-MENTION-2472] et **295 des rayures plus rapprochées que son cousin des plaines** [u-MENTION-2473] .. Il comprend **296 deux sous-espèces** [u-MENTION-2478] , en **297 danger d'298 extinction** [u-MENTION-2479] ; • • Le zèbre de Grévy (*Equus grevyi*) est le plus grand de tous les zèbres.. Il possède **299 300 une longue crinière hérissée ainsi qu'une tête longue et étroite** [u-MENTION-2485] qui le fait ressembler à **u-2490 une mule** [u-MENTION-2490] .. C'est un habitant **301 des savanes d'302 Éthiopie, de 303 Somalie et 304 du nord 305 du Kenya** [u-MENTION-2493] .. Il est également menacé d'**306 extinction** [u-MENTION-2499] .. • **307 Histoire évolutive** [u-MENTION-2500] Il y a **308 environ 54 millions d'années** [u-MENTION-2501] , **ammifere un petit mammifère** [u-MENTION-2502] , de **310 la taille d'311 un renard sans 312 sabot** [u-MENTION-2503] , baptisé **313 Hyracotherium** [u-MENTION-2506] par **314 les paléontologues** [u-MENTION-2507] , vivait sur **315 le continent américain** [u-MENTION-2508] .. **ammifere Il** [u-MENTION-2509] serait à **309 l'origine de tous les équidés** [u-MENTION-2510] (cheval, **318 poney** [u-MENTION-2513] , âne, zèbre).. Les zèbres sont, probablement, **316 les plus anciens représentants du genre Equus** [u-MENTION-2517] .. [phrase] Le zèbre faisait également **317 partie de la famille 320 des périssodactyles** [u-MENTION-2522] , tels que **323 321 les rhinocéros ou 322 les tapirs** [u-MENTION-2525] .. Cette famille regroupe **319 tous les animaux** [u-MENTION-2529] comportant **324 un nombre impair de 325 doigts** [u-MENTION-2530] .. **326 Au Miocène** [u-MENTION-2532] , **327 les graminées** [u-MENTION-2533] étant plus riches et plus abondantes, les équidés primitifs en profitèrent pour se multiplier et se développer : **328 leurs jambes** [u-MENTION-2535] s'allongèrent pour mieux échapper **329 aux prédateurs** [u-MENTION-2537] , et **330 leurs pieds** [u-MENTION-2538] ne comptèrent, désormais, plus qu'**331 un seul doigt** [u-MENTION-2540] , recouvert d'**332 un ongle** [u-MENTION-2541] , **333 le sabot** [u-MENTION-2542] .. [phrase] Grâce à **335 leur développement** [u-MENTION-2547] , ils gagnèrent en **336 vitesse** [u-MENTION-2550] , se déplacèrent davantage à **337 la recherche de 338 nourriture** [u-MENTION-2552] et entreprirent de **339 plus longues migrations** [u-MENTION-2555] .. C'est à ce moment -là qu'ils se répandirent en Asie, en Afrique et en Europe, passant par **334 le détroit de 340 Béring** [u-MENTION-2561] , alors recouvert de **341 glace épaisse** [u-MENTION-2563] .. De **342 nos jours** [u-MENTION-2564] , il est presque impossible de distinguer le crâne d'un zèbre de **343 celui d'un cheval** [u-MENTION-2568] , mais nous pouvons penser que les équidés qui colonisèrent les savanes tropicales devinrent des zèbres, laissant **344 les déserts arides** [u-MENTION-2575] aux ânes sauvages et **346 les zones tempérées de 347 l'hémisphère Nord** [u-MENTION-2577] aux chevaux sauvages. . Des fossiles datant du Pléistocène démontrent **345 la grande répartition de ces équidés** [u-MENTION-2582] .. Au Pléistocène, **352 348 Equus sivalensis (349 Chine) et 350 Equus sellardsi** [u-MENTION-2585] (**351 Amérique du Nord** [u-MENTION-2589]) ressemblaient au quagga, ce zèbre d'Afrique qui s'est éteint **353 au XIXe siècle** [u-MENTION-2594] .. À **354 la même époque** [u-MENTION-2595] , celui qui occupait **355 le Sud de l'Afrique** [u-MENTION-2597] est *Equus pilicatus*, ancêtre direct du zèbre de Grévy.. **356 L'évolution de 357 nombreuses espèces d'équidés** [u-MENTION-2601] est mal connue, mais on sait qu'il existait encore **360 des ânes sauvages et des zèbres** [u-MENTION-2605] en Europe à **358 la fin de 359 la dernière période glaciaire de 361 l' ère quaternaire** [u-MENTION-2609] . . Aujourd'hui, **362 les équidés sauvages** [u-MENTION-2612] sont devenus rares.. Il existe sept principales espèces équinées, dont **363 la plupart** [u-MENTION-2615] sont très proches de **364 l'extinction** [u-MENTION-2616] : les trois zèbres d'Afrique et leurs cousins, l'âne sauvage ; **365 le cheval sauvage de 366 Mongolie** [u-MENTION-2621] ; les deux ânes sauvages d'Asie, **368 le kiang** [u-MENTION-2626] et **369 l'hémione** [u-MENTION-2627] . . **371 367 Étymologie et 370 dénominations** [u-MENTION-2628] Le substantif masculin « zèbre » est **373 un emprunt au portugais zebra** [u-MENTION-2632] , substantif féminin lui -même probablement issu — comme **372 l'espagnol cebra** [u-MENTION-2635] — d'**375 un latin vulgaire * eciferus** [u-MENTION-2636] , variation vernaculaire **376 du latin equiferus** [u-MENTION-2637] (« **377 cheval sauvage** [u-MENTION-2638] »), composé de **378 equus** [u-MENTION-2639] (« **379 cheval** [u-MENTION-2640] ») et de **380 ferrus** [u-MENTION-2641] (« **381 sauvage** [u-MENTION-2642] ») . . En français, « zèbre » est attesté **382 au début 383 du XVIIe siècle** [u-MENTION-2645] :

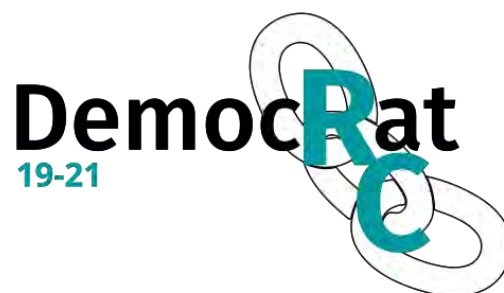
d'après **384 le Trésor de la langue française informatisé** [u-MENTION-2647] , **385 sa première occurrence** [u-MENTION-2648] (**386 graphie : « zebre** [u-MENTION-2650] ») se trouve dans **387 l' Histoire des choses plus mémorables advenues tant ez Indes orientales, que autres païs de la découverte** [u-MENTION-2651] **388 des Portugais de 389 Pierre de Jarric** [u-MENTION-2652] (en), parue à **390 Bordeaux** [u-MENTION-2654] en **391 1610** [u-MENTION-2655] . . Le portugais zebra servait initialement à désigner **374 un équidé sauvage** [u-MENTION-2657] , **392 le zebro** [u-MENTION-2658] ou **393 zebro** [u-MENTION-2659] , particulièrement abondant dans la péninsule Ibérique jusqu'**395 au XVIe siècle** [u-MENTION-2661] .. En **396 juin 2015** [u-MENTION-2662] , **397 la nature biologique précise de l'animal** [u-MENTION-2663] reste incertaine.. Quatre hypothèses ont été avancées selon lesquelles **398 le zebro ibérique** [u-MENTION-2667] serait : soit **399 l'hydrintin** [u-MENTION-2668] (**400 Equus hydruntinus** [u-MENTION-2669]), un onagre éteint ; soit **401 l'ancêtre 402 du Sorraia** [u-MENTION-2670] , une race chevaline portugaise ; soit **403 un onagre moderne** [u-MENTION-2672] , introduit dans la péninsule Ibérique par **405 les musulmans** [u-MENTION-2674] ; soit **394 un équidé domestique errant** [u-MENTION-2675] , **404 âne** [u-MENTION-2676] ou **406 cheval** [u-MENTION-2677] . . [phrase] On rencontre aussi **407 le terme zébrasse** [u-MENTION-2685] ou **408 zebrette** [u-MENTION-2686] . . [phrase] **409 Les zèbres et l'homme** [u-MENTION-2694] Tentatives de domestication S'il est possible de domestiquer **412 un zèbre pris isolément** [u-MENTION-2699] , l'espèce s'**414 y** [u-MENTION-2701] prête peu.. Vivant dans **413 un milieu comprenant de 415 grands prédateurs tels que le lion, le guépard et la hyène** [u-MENTION-2702] , le zèbre a développé **416 de puissantes techniques de 417 défense** [u-MENTION-2708] .. D'après l'universitaire Carol Hall, « le fait qu' il soit **421 un « aliment pour lion** [u-MENTION-2714] » l' a peut-être rendu moins attrayant **418 aux yeux 419 des premiers humains** [u-MENTION-2717] ».. Les tentatives de domestication s'expliquent par **410 sa résistance 411 au climat chaud** [u-MENTION-2721] , **420 aux maladies africaines** [u-MENTION-2724] et de **422 sa rapidité supérieure à 423 celle 424 du cheval rustique** [u-MENTION-2725] .. En **425 Afrique du Sud** [u-MENTION-2729] , les Boers ont essayé **426 plusieurs fois** [u-MENTION-2731] de harnacher des zèbres mais ont vu **427 la plupart de 428 leurs tentatives** [u-MENTION-2734] déjouées par **429 la nature sauvage et têtue de l' animal** [u-MENTION-2737] . . **430 Hybrides** [u-MENTION-2739] Le zébrâne est le croisement d'un zèbre et d'**434 une ânesse** [u-MENTION-2743] . . Le zébrule est le croisement d'un zèbre et d'**431 une jument** [u-MENTION-2747] . . **432 Reconstitutions** [u-MENTION-2748] Le Quagga contemporain est **433 une tentative de 435 restitution de cette sous-espèce** [u-MENTION-2750] .. Il s'agit d'**436 un élevage sélectif du Zèbre des plaines** [u-MENTION-2753] (Equus Quagga) dans **437 le but** [u-MENTION-2756] de retrouver **438 l'apparence de 439 l'animal disparu** [u-MENTION-2757] après avoir découvert **440 la grande similitude de 441 l'ADN 442 des spécimens disparus** [u-MENTION-2759] avec **443 celui des zèbres modernes** [u-MENTION-2762] . . Les zèbres dans **444 la culture** [u-MENTION-2765] **446 Vocabulaire** [u-MENTION-2766] « Zébré » est **448 un adjectif** [u-MENTION-2768] , dérivé de zèbre, qui signifie rayé, souvent de **447 rayures noires et blanches alternées** [u-MENTION-2771] mais pas obligatoirement. . Dans **450 plusieurs langues** [u-MENTION-2772] (comme **451 l'espagnol** [u-MENTION-2773] , **452 l'anglais** [u-MENTION-2774] ou **453 le néerlandais** [u-MENTION-2775]), on appelle **454 le passage piéton** [u-MENTION-2777] respectivement **455 le passage-zèbre** [u-MENTION-2778] , **456 le croisement-zèbre** [u-MENTION-2779] et **457 le sentier-zèbre** [u-MENTION-2780] . . **458 Idiotismes animaliers** [u-MENTION-2781] : **459 un drôle de zèbre** [u-MENTION-2782] est **461 un individu douteux** [u-MENTION-2784] et **462 filer comme un zèbre** [u-MENTION-2785] veut dire courir très vite. . **463 Proverbes** [u-MENTION-2787] « La pluie mouille le zèbre mais n'efface pas ses rayures » (**449 massai** [u-MENTION-2793] , **460 Kenya** [u-MENTION-2794]) • « Le noir porte **464 sa négritude** [u-MENTION-2796] comme le zèbre ses rayures » (**465 Sénégal** [u-MENTION-2801]) • « **466 Un homme sans 467 tradition** [u-MENTION-2802] est comme **468 un zèbre sans rayure** [u-MENTION-2804] » (Afrique) • « **469 L'ombre du zèbre** [u-MENTION-2807] n'a pas de rayures » (**470 René Char** [u-MENTION-2810] , poète français) • **471 Le symbolisme du zèbre** [u-MENTION-2811] Le zèbre est **472 un symbole de 473 courage** [u-MENTION-2814] car il entreprend **474 de grandes migrations annuelles** [u-MENTION-2817] pour chercher **475 des pâturages** [u-MENTION-2818] , malgré HyènesCrocodiles les lions et

les hyènes ou encore **478 les crocodiles** [u-MENTION-2822] HyenesCrocodiles qui les menacent. . Le zèbre est aussi **476 un symbole** **477 du métissage et de** **479 l'harmonie entre** **480 les races** [u-MENTION-2826] .. Ainsi, le zèbre est **481 le symbole animal du Botswana** [u-MENTION-2831] : ses rayures noires et blanches ornent **483 le drapeau** [u-MENTION-2835] depuis **484 1996** [u-MENTION-2836] et **485 deux zèbres affrontés** [u-MENTION-2837] encadrent **486 les armoiries du pays** [u-MENTION-2838] . . Le Zèbre de Grévy est aussi **482 le symbole de l'Afrique et de** **488 la faune africaine** [u-MENTION-2841] . . En **487 Angola** [u-MENTION-2844] , **489 des crinières de zèbres** [u-MENTION-2845] sont portées lors de **490 danses rituelles** [u-MENTION-2847] ayant lieu pour **491 les cérémonies d'****492 initiation** **493 des jeunes** [u-MENTION-2848] lors **494 des rites de «** **495 transformation** [u-MENTION-2851] ». . Zèbres dans **497 445 l'art et** **496 la fiction** [u-MENTION-2854] zèbre dans **498 l'art sur** **499 Commons** [u-MENTION-2858] • « **500 Les Zèbres** [u-MENTION-2860] » ou « **501 le Zèbre** [u-MENTION-2861] » est **502 un surnom** [u-MENTION-2862] ou **503 un titre souvent donné en** **504 référence à** **505 des bandes noires et blanches ou à** **506 un personnage peu fiable** [u-MENTION-2863] .. • • Drôles de zèbres est **508 un film de** **507 Guy Lux de** **509 1977** [u-MENTION-2868] où un patron d'**510 hôtel** [u-MENTION-2873] savant fou à **511 ses heures** [u-MENTION-2874] veut transformer un cheval de course en **513 crack** [u-MENTION-2877] en lui injectant notamment **512 du sang de zèbre** [u-MENTION-2879] .. • • **514 Le Zèbre** [u-MENTION-2881] est **515 un roman d'****516 Alexandre Jardin** [u-MENTION-2882] , **517 prix Femina** **518 1988** [u-MENTION-2884] .. • • Zig Zag, l'étalon zébré est **520 un film** [u-MENTION-2887] pour **521 enfants** [u-MENTION-2888] dont **519 le héros** [u-MENTION-2890] est **522 un zèbre** [u-MENTION-2891] .. • • **523 Marty le zèbre** [u-MENTION-2892] , personnage **524 des films d'animation** **525 Madagascar**, **526 Madagascar 2 et** **527 Madagascar 3** [u-MENTION-2893] .. • • **530 528 Le film et** **529 le manga Zebraman** [u-MENTION-2897] racontent **531 l'histoire d'un super-héros** [u-MENTION-2900] dont **535 532 le costume et** **533 les pouvoirs** [u-MENTION-2903] (**534 une « super-ruade** [u-MENTION-2906] ») sont inspirés du zèbre..

Annexe C

Documentation de la ressource Democrat_RC_19-21

Documentation de la ressource Democrat_RC_19–21 telle que mise à disposition sur Ortolang. Tout complément d’information est disponible à l’adresse suivante :
https://hdl.handle.net/11403/democrat_rc_19-21.



Documentation de la ressource

DEMOCRAT_RC_19-21

mai 2022

Conformément aux objectifs du projet DEMOCRAT, la ressource est diffusée gratuitement sous licence ouverte CC BY-SA. Les fichiers et la documentation de la ressource DEMOCRAT_RC_19-21 ont été déposés sur la plateforme ORTOLANG (Outils et Ressources pour un Traitement Optimisé de la LANGue), plateforme vouée à être pérenne.

Lien vers le corpus annoté DEMOCRAT : <https://hdl.handle.net/11403/democrat>
Lien vers la ressource DEMOCRAT_RC_19-21 : https://hdl.handle.net/11403/democrat_rc_19-21

DEMOCRAT_RC_19-21

1 Description

La ressource DEMOCRAT_RC_19-21¹ a été produite dans le cadre de la thèse de Silvia Federzoni au laboratoire CLLE (UMR5263, CNRS & Université de Toulouse Jean Jaurès). Cette ressource est le résultat d'une extraction automatique de toutes les chaînes de coréférence (au moins 3 maillons) telles qu'annotées dans les textes du 19^{ème}, 20^{ème} et 21^{ème} siècle constituant le corpus DEMOCRAT (Landragin, 2016 ; Lattice, LiLPa, ICAR, & IHRIM, 2019 ; Quignard, Le Mené, & Landragin, 2021)².

Les fichiers qui constituent la ressource rassemblent ainsi toutes les chaînes d'au moins 3 maillons présentes dans les textes. En s'appuyant sur l'exemple de la ressource AnnoDis (Péry-Woodley, Afantenos, Ho-Dac, & Asher, 2011), DEMOCRAT_RC_19-21 présente chaque chaîne sous forme de *segment de texte*, celui-ci correspondant uniquement à la portion de texte couverte par la chaîne en question. De ce fait, une même partie de texte peut être répétée plusieurs fois, comme le montre la figure 1, où deux chaînes couvrent à peu près la même portion de texte³. Ce format facilite l'étude de chaque chaîne indépendamment des autres, tout en laissant la possibilité de revenir aux annotations originales grâce aux identifiants des chaînes et des mentions (cf. section 3).

```
<div type="Reference_chain" xml:id="s-CHAINE-282" from="w_export_1397" id_firstMention="u-MENTION-400"
to="w_export_1462" id_lastMention="u-MENTION-416" extent="3 mentions">
  <s n="52" nbToks="25" nbWords="21" from="w_export_1397" to="w_export_1421">
    <rs type="mention" xml:id="u-MENTION-400" from="w_export_1397" to="w_export_1401">Sa langue bleue et
    préhensile</rs>
    est la plus puissante, la plus coriace et la plus longue ( 55 cm ) parmi les ongulés..
  </s>
  <s n="53" nbToks="19" nbWords="18" from="w_export_1422" to="w_export_1440">
    Elle peut
    <rs type="mention" xml:id="u-MENTION-405" from="w_export_1424" to="w_export_1424">l'</rs>
    allonger pour atteindre les pousses les plus tendres entre les barrières d'épines d'acacias..
  </s>
  <s n="54" nbToks="12" nbWords="11" from="w_export_1441" to="w_export_1452" nbMentions="0"/>
  <s n="55" nbToks="38" nbWords="35" from="w_export_1453" to="w_export_1490">
    Elle saisit donc les pousses d'acacias avec
    <rs type="mention" xml:id="u-MENTION-416" from="w_export_1461" to="w_export_1462">sa langue</rs>
    , puis les guide entre ses lèvres, referme la bouche et tire la tête en arrière pour racler les feuilles
    grâce à ses dents du bas.
    <lb n="21"/>
  </s>
</div>
<div type="Reference_chain" xml:id="s-CHAINE-285" from="w_export_1428" id_firstMention="u-MENTION-406"
to="w_export_1465" id_lastMention="u-MENTION-418" extent="3 mentions">
  <s n="53" nbToks="19" nbWords="18" from="w_export_1422" to="w_export_1440">
    Elle peut l'allonger pour atteindre
    <rs type="mention" xml:id="u-MENTION-406" from="w_export_1428" to="w_export_1432">les pousses les plus
    tendres</rs>
    entre les barrières d'épines d'acacias..
  </s>
  <s n="54" nbToks="12" nbWords="11" from="w_export_1441" to="w_export_1452" nbMentions="0"/>
  <s n="55" nbToks="38" nbWords="35" from="w_export_1453" to="w_export_1490">
    Elle saisit donc
    <rs type="mention" xml:id="u-MENTION-414" from="w_export_1456" to="w_export_1459">les pousses d'acacias</rs>
    avec sa langue, puis
    <rs type="mention" xml:id="u-MENTION-418" from="w_export_1465" to="w_export_1465">les</rs>
    guide entre ses lèvres, referme la bouche et tire la tête en arrière pour racler les feuilles grâce à ses
    dents du bas.
    <lb n="21"/>
  </s>
</div>
```

FIGURE 1 – Aperçu des chaînes de référence dans DEMOCRAT_RC_19-21

1. RC pour *reference chains* ; 19-21 pour indiquer les siècles couverts par les textes qui constituent cette ressource. Nous utiliserons cette notation aussi pour nommer les fichiers xml. On ajoute cette notation au nom du fichier source, par exemple le fichier *aden* est nommé dans la ressource *aden_RC_19-21*.

2. Site web du projet DEMOCRAT : <http://www.lattice.cnrs.fr/democrat/>

3. Si les phrases ne contiennent pas de maillons, le texte correspondant à la phrase n'est pas présent (cf. section 3).

Les fichiers qui constituent la ressource sont diffusés sous deux formats différents :

- Fichiers xml (un par fichier DEMOCRAT) : il s’agit de fichiers xml avec annotation intégrée. Ces fichiers suivent les recommandations du consortium TEI (version P5) (cf. section 3).
- Fichiers CoNLL (un par chaîne de référence) : il s’agit de fichiers au format CoNLL-U (fichiers tabulé, un token par ligne) augmentés avec des colonnes supplémentaires pour les entités nommées et les annotations des mentions. Nous avons utilisé Stanza (Qi, Zhang, Zhang, Bolton, & Manning, 2020)⁴ pour l’analyse des textes. Ces fichiers permettent de récupérer différentes informations pour caractériser les maillons et les chaînes (cf. section 4).

1.1 Composition de la ressource

Le corpus DEMOCRAT, dont la constitution est donnée dans le tableau 1, est composé de textes narratifs et non narratifs, relevant de 18 genres textuels différents.

Siècle			Nb de mots	Nb de CR	Nb d’ER
	#	%			
21^e	2	3,45	21801	711	6298
20^e	15	25,86	175434	5302	49964
19^e	13	22,41	161897	4356	45589
18^e	6	10,34	66889	2301	18956
17^e	7	12,07	86147	3167	23406
16^e	5	8,62	61216	1630	15716
15^e	3	5,17	34416	728	10723
14^e	3	5,17	34616	1221	9678
13^e	2	3,45	21729	700	7412
12^e	1	1,72	11647	124	4138
environ 11^e	1	1,72	13059	170	5043
Total	58		688851	20410	196923

TABLEAU 1 – Composition du corpus DEMOCRAT

Seuls les textes du 19^{ème}, 20^{ème} et 21^{ème} siècle ont été pris en compte dans la ressource DEMOCRAT_RC_19-21 (lignes en gras dans tableau 1). La ressource DEMOCRAT_RC_19-21 est ainsi composée de 29 fichiers qui regroupent les chaînes annotées dans les 29 textes retenus⁵ correspondants. Les 29 textes pris en compte sont de type narratif (18 textes) et non narratif (11 textes) et couvrent 7 genres textuels différents, pour un total de 347 155 mots (cf. tableau 2).

Le tableau 2 indique les 29 fichiers constituant la ressource avec leur type et genre textuel, le siècle d’appartenance, le nombre de mots du texte, le nombre de chaînes (d’au moins 3 maillons) et le nombre de maillons.

Les informations présentes dans les 5 premières colonnes du tableau sont celles indiquées dans la ressource DEMOCRAT. De ce fait, le nombre de mots indiqué ne correspond pas au nombre de mots du fichier DEMOCRAT_RC_19-21 mais au nombre de mots contenus dans le fichier DEMOCRAT. Le nombre de chaînes et de maillons a été calculé à partir de DEMOCRAT_RC_19-21 (pour les traitements réalisés voir section 2).

4. <https://stanfordnlp.github.io/stanza/index.html>.

5. Le texte *Chateaubriandgenie* a été exclu de cette sélection parce que sa structure XML-TEI est différente de celle des autres textes. En effet, c’est le seul texte où la structure textuelle (chapitres, titres etc.) est indiquée.

Fichier	Type textuel	Genre textuel	Siècle	Nb mots	Nb chaînes	Nb maillons
aden	non_narratif	pamphlet	20 ^e	13290	131	1137
articleswiki	non_narratif	articles_encyclopédiques	21 ^e	10628	119	1245
bouvardetpecuchet	narratif	roman	19 ^e	12241	103	1437
capitainefracasse	narratif	roman	19 ^e	11817	150	1235
codecivildesfrancais-1	non_narratif	texte_juridique	19 ^e	12943	218	1853
codecivildesfrancais-2	non_narratif	texte_juridique	19 ^e	14431	202	1773
codeprocedurepenale	non_narratif	texte_juridique	20 ^e	12509	330	2086
conventioninstitutiv	non_narratif	texte_juridique	20 ^e	5384	73	586
conventionmilieumarin	non_narratif	texte_juridique	20 ^e	12860	294	2053
conventionthontropical	non_narratif	texte_juridique	21 ^e	11173	200	1613
convtransportaerien	non_narratif	texte_juridique	20 ^e	9402	138	1233
delavilleaumoulin	narratif	roman	20 ^e	11945	133	2182
diablaucorps	narratif	roman	20 ^e	14596	154	2478
doucelumiere	narratif	roman	20 ^e	12268	110	1856
elisabethseton	narratif	biographie	20 ^e	12990	172	2167
estrepublikain	non_narratif	articles_de_presse	20 ^e	12254	170	888
estrepublikain-2	non_narratif	articles_de_presse	20 ^e	11605	151	811
jeanchristophe-1	narratif	roman	20 ^e	10962	154	1951
jeanchristophe-2	narratif	roman	20 ^e	10303	96	1834
madamedehautefort	narratif	biographie	19 ^e	11758	135	2083
mademoisellefifi-1	narratif	nouvelle	19 ^e	10893	79	1367
mademoisellefifi-2	narratif	nouvelle	19 ^e	9812	79	1316
mademoisellefifi-3	narratif	nouvelle	19 ^e	11624	87	1621
morteamoureuse	narratif	nouvelle	19 ^e	14270	137	2122
nemoville	narratif	roman	20 ^e	12378	130	2212
pauline	narratif	roman	19 ^e	12284	89	1620
rosaliedeconstant	narratif	biographie	20 ^e	12688	143	1944
sarrasine	narratif	nouvelle	19 ^e	15346	119	1937
ventredeparis	narratif	roman	19 ^e	12501	140	1572
Total				347 155	4236	48 212

TABLEAU 2 – Composition de la ressource DEMOCRAT_RC_19-21

2 Constitution de le ressource

La ressource DEMOCRAT_RC_19-21 est le résultat de différents traitements effectués sur les annotations d'origine. Ces traitements ont porté sur :

1. Une sélection des chaînes de référence. Nous avons fait le choix de n'intégrer dans la ressource que les chaînes constituées d'au moins 3 maillons. Les singletons (38 619 dans les 29 textes DEMOCRAT) et les chaînes constituées de deux maillons (5 651 dans les 29 textes DEMOCRAT) ne sont pas reproduits dans la ressource.
2. L'extraction et le nettoyage de la segmentation en phrases : la segmentation en phrases du corpus DEMOCRAT, effectuée automatiquement par TreeTagger, a été corrigée automatiquement dans les cas où un maillon chevauchait plusieurs phrases. Ainsi, dans l'erreur de segmentation observée dans l'exemple suivant

*<s n="244"> [...] j'ai entendu mes parents parler de la visite d'<rs type="mention"> un
M. </s> <s n="245"> Grangier</rs> [...] </s>*

les phrases ont été fusionnées pour aboutir au résultat suivant :

`<s n="244,245">[...] j'ai entendu mes parents parler de la visite d'<rs type="mention">
un M. Grangier</rs> [...] </s>.`

Par ce traitement, 196 fusions ont été réalisées dont 81,6% (160) concernent la fusion de deux phrases, 15,82% (31) la fusion de 3 phrases et 2,55% (5) la fusion de plus de 3 phrases.

3. La délimitation des maillons : dans la majorité des cas les maillons ont été indiqués tels qu'ils l'étaient dans le corpus DEMOCRAT. Dans certains cas, plusieurs maillons d'une même chaîne sont imbriqués comme dans l'exemple suivant où "du CP" est un maillon imbriqué dans le maillon "Les enfants de la maternelle et du CP", tous deux appartenant à la même chaîne : `<rs type="mention"> Les enfant de la maternelle<rs type="mention"> et du CP</rs></rs>.`

Dans des cas de ce type, le maillon imbriqué n'a pas été retenu et le maillon le plus large a été conservé (p. ex. `<rs type="mention"> Les enfants de la maternelle et du CP</rs>`). Si après ce traitement la chaîne en cours de traitement n'est plus constituée d'au moins 3 maillons, celle-ci ne figurera pas dans la ressource DEMOCRAT_RC_19-21.

Par ce traitement 262 maillons imbriqués ont été supprimés.

3 Fichiers XML

La ressource DEMOCRAT_RC_19-21 a été encodée en suivant les recommandations du consortium TEI (version P5). Certains attributs non définis par la TEI ont été ajoutés. Les principaux éléments sont énumérés ci-dessous. Plus de détails concernant la structure xml et les informations pour chaque élément sont disponibles dans la documentation mise à disposition (fichier *Democrat_RC_19-21_schema.pdf*).

Pour assurer l'interopérabilité avec le corpus DEMOCRAT d'origine les `xml:id` des chaînes et des maillons correspondent aux identifiants attribués à ces éléments dans les fichiers DEMOCRAT ; les `id` des tokens correspondent aux identifiants du corpus DEMOCRAT et permettent de récupérer les informations relatives aux tokens dans les fichiers source .xml tagués par TreeTagger.

Dans les 29 fichiers DEMOCRAT pris en compte, la structure textuelle est signalée de façon minimale. Nous avons reproduit les marques de saut de ligne `<lb>` parce que cet élément pourrait aider à identifier des titres et/ou des paragraphes.

Ci-dessous la liste des éléments que nous avons utilisés et leurs attributs :

- `<teiHeader>` : contient les métadonnées relatives au texte (titre, auteur, date de publication), aux contributeurs, à la licence et aux informations relatives au nombre d'éléments présents dans le fichier.
- `<div type="Reference_chain" xml:id="" from="" id_firstMention="" to="" id_lastMention="" extent="">` : délimite les chaînes de référence. Ses attributs :
 - `xml:id` : indique l'identifiant DEMOCRAT de la CR
 - `from` : correspond à l'identifiant DEMOCRAT du premier token du premier maillon
 - `id_firstMention` : correspond à l'identifiant DEMOCRAT du premier maillon de la chaîne
 - `to` : correspond à l'identifiant DEMOCRAT du dernier token du dernier maillon
 - `id_lastMention` : correspond à l'identifiant DEMOCRAT du dernier maillon de la chaîne
 - `extent` : correspond au nombre de maillons qui constituent la CR
- `<rs type="mention" xml:id="" from="" to="">` : délimite les maillons de la chaîne
 - `xml:id` : correspond à l'identifiant DEMOCRAT du maillon
 - `from` : indique l'identifiant DEMOCRAT du token de début du maillon
 - `to` : indique l'identifiant DEMOCRAT du token de fin du maillon
- `<s n="" nbToks="" nbWords="" from="" to="">` : délimite les phrases du texte. Si des maillons sont présents dans la phrase, alors le texte est reproduit. Si la phrase ne contient pas de maillons, le texte n'est pas reproduit. Ses attributs :
 - `n` : correspond à l'identifiant DEMOCRAT de la phrase ; si deux phrases ont été fusionnées (cf. traitement 2), les deux identifiants sont présents, séparés par une virgule

- **nbToks** : indique le nombre de tokens de la phrase, y compris la ponctuation
- **nbWords** : indique le nombre de mots de la phrase sans prendre en compte la ponctuation
- **from** : correspond à l'identifiant DEMOCRAT du premier token de la phrase
- **to** : correspond à l'identifiant DEMOCRAT du dernier token de la phrase
- **nbMentions** : si la phrase ne contient pas de maillons cet attribut est ajouté à la balise avec la valeur 0
- **<lb n=" ">** : indique les sauts de ligne présents dans les fichiers d'origine. L'attribut **n** est l'identifiant DEMOCRAT de ces sauts de ligne.

4 Fichiers CoNLL

Les fichiers CoNLL sont des fichiers au format CoNLL-U résultant de l'analyse des textes avec Stanza (Qi et al., 2020). Chaque fichier correspond à une chaîne de référence. Le nom des fichiers indique le texte auquel la chaîne appartient et l'identifiant DEMOCRAT de la chaîne.

Nous avons utilisé la boîte à outil Stanza pour tokeniser et lemmatiser les textes, obtenir l'analyse morpho-syntaxique, les dépendances syntaxiques et les entités nommées. Les fichiers mis à disposition ont été obtenus en utilisant les modèles pour le français ⁶. Le modèle ⁷ utilisé pour la reconnaissance des entités nommées prend en charge quatre types : PER (Person), LOC (Location), ORG (Organization) et MISC (Miscellaneous).

Tâche	Modèle
tokenize	gsd
mwt (multi-word token)	gsd
pos	gsd
lemma	gsd
depparse	gsd
ner (named entity recognition)	wikiner

L'inclusion d'un token dans un maillon a été indiquée dans des colonnes supplémentaires en suivant le codage utilisé pour les entités nommées :

- **S-mention** : indique que le token correspond à un maillon complet. Lorsque le maillon correspond à la forme contractée du pronom relatif *lequel*, deux S-mention se succèdent.
- **B-mention** : indique que le token est le début d'un maillon. Lorsque le maillon commence par un déterminant dont la forme est contractée, deux B-mention se succèdent.
- **I-mention** : indique que le token est dans un maillon contenant plusieurs tokens (sans en être le premier ou le dernier).
- **E-mention** : indique que le token est le dernier d'un maillon contenant plusieurs tokens.

À titre d'exemple nous proposons ci-dessous un extrait du fichier Stanza correspondant à la chaîne 285 du fichier *articleswiki* (*articleswiki_s-CHAINE-285*).

Le colonnes correspondent dans l'ordre à :

- la forme
- le lemme
- la catégorie grammaticale (*Universal Dependencies tagset*)
- les informations morpho-syntaxiques
- l'index du gouverneur
- la forme du gouverneur
- la fonction syntaxique (*Universal Dependencies tagset*)
- le type d'entité nommée le cas échéant

6. Pour plus de détails sur les modèles : https://universaldependencies.org/treebanks/fr_gsd/index.html

7. Pour plus de détails sur le modèle : https://stanfordnlp.github.io/stanza/available_models.html

- i) le offset de début du token
- j) le offset de fin du token
- k) l'annotation des maillons le cas échéant
- l) l'identifiant DEMOCRAT du maillon le cas échéant

Les couches utilisées pour ajouter les informations sont différentes de celles utilisées pour le format CoNLL. Plus précisément nous ajoutons des dépendances syntaxiques et les entités nommées ont seulement 4 types (18 dans les format CoNLL d'origine).

1	Elle	il	PRON	Gender=Fem Number=Sing Person=3 PronType=Prs	2	peut	nsubj	O	115	119		
2	peut	pouvoir	VERB	Mood=Ind Number=Sing Person=3 Tense=Pres VerbForm=Fin	0	root	root	O	120	124		
3	l'	le	PRON	Number=Sing Person=3 PronType=Prs	4	allonger	obj	O	125	127		
4	allonger	allonger	VERB	VerbForm=Inf	2	peut	xcomp:obj	O	127	135		
5	pour	pour	ADP		6	atteindre	mark	O	136	140		
6	atteindre	atteindre	VERB	VerbForm=Inf	4	allonger	advcl	O	141	150		
7	les	le	DET	Definite=Def Number=Plur PronType=Art	8	pousses	det	O	151	154	B-mention	u-MENTION-406
8	pousses	pousse	NOUN	Gender=Fem Number=Plur	6	atteindre	obj	O	155	162	I-mention	u-MENTION-406
9	les	le	DET	Definite=Def Number=Plur PronType=Art	11	tendres	advmod	O	163	166	I-mention	u-MENTION-406
10	plus	plus	ADV		9	les	fixed	O	167	171	I-mention	u-MENTION-406
11	tendres	tendre	ADJ	Gender=Fem Number=Plur	8	pousses	amod	O	172	179	E-mention	u-MENTION-406
12	entre	entre	ADP		14	barrières	case	O	180	185		
13	les	le	DET	Definite=Def Number=Plur PronType=Art	14	barrières	det	O	186	189		
14	barrières	barrière	NOUN	Gender=Fem Number=Plur	8	pousses	nmod	O	190	199		
15	d'	de	ADP		16	épinas	case	O	200	202		
16	épinas	épine	NOUN	Gender=Fem Number=Plur	14	barrières	nmod	O	202	208		
17	d'	de	ADP		18	acacias	case	O	209	211		
18	acacias	acacia	NOUN	Gender=Fem Number=Plur	16	épinas	nmod	O	211	218		
19	..	..	PUNCT		2	peut	punct	O	218	220		
1	La	le	DET	Definite=Def Gender=Fem Number=Sing PronType=Art	2	girafe	det	B-MISC	221	223		
2	girafe	girafe	NOUN	Gender=Fem Number=Sing	4	a	nsubj	E-MISC	224	230		
3	n'	ne	ADV	Polarity=Neg	4	a	advmod	O	231	233		
4	a	avoir	VERB	Mood=Ind Number=Sing Person=3 Tense=Pres VerbForm=Fin	0	root	root	O	233	234		
5	pas	pas	ADV	Polarity=Neg	4	a	advmod	O	235	238		
6	d'	un	DET	Definite=Ind Number=Plur PronType=Art	7	incisives	det	O	239	241		
7	incisives	incisive	NOUN	Gender=Fem Number=Plur	4	a	obj	O	241	250		
8	à	à	ADP		10	mâchoire	case	O	251	252		
9	la	le	DET	Definite=Def Gender=Fem Number=Sing PronType=Art	10	mâchoire	det	O	253	255		
10	mâchoire	mâchoire	NOUN	Gender=Fem Number=Sing	4	a	obl:arg	O	256	264		
11	supérieure	supérieur	ADJ	Gender=Fem Number=Sing	10	mâchoire	amod	O	265	275		
12	.	.	PUNCT		4	a	punct	O	275	276		
1	Elle	il	PRON	Gender=Fem Number=Sing Person=3 PronType=Prs	2	saisit	nsubj	O	277	281		
2	saisit	saisir	VERB	Mood=Ind Number=Sing Person=3 Tense=Pres VerbForm=Fin	0	root	root	O	282	288		
3	donc	donc	ADV		2	saisit	advmod	O	289	293		
4	les	le	DET	Definite=Def Number=Plur PronType=Art	5	pousses	det	O	294	297	B-mention	u-MENTION-414
5	pousses	pousse	NOUN	Gender=Fem Number=Plur	2	saisit	obj	O	298	305	I-mention	u-MENTION-414
6	d'	de	ADP		7	acacias	case	O	306	308	I-mention	u-MENTION-414
7	acacias	acacia	NOUN	Gender=Fem Number=Plur	5	pousses	nmod	O	308	315	E-mention	u-MENTION-414
8	avec	avec	ADP		10	langue	case	O	316	320		
9	sa	son	DET	Gender=Fem Number=Sing PossNumber=Sing PossPerson=3 PronType=Prs	10	langue	det	O	321	323		
10	langue	langue	NOUN	Gender=Fem Number=Sing	2	saisit	obl:mod	O	324	330		
11	,	,	PUNCT		19	referme	punct	O	330	331		
12	puis	puis	ADV		19	referme	cc	O	332	336		
13	les	le	DET	Definite=Def Number=Plur PronType=Art	14	guide	det	O	337	340	S-mention	u-MENTION-418
14	guide	guide	NOUN	Gender=Masc Number=Sing	19	referme	nsubj	O	341	346		
15	entre	entre	ADP		17	lèvres	case	O	347	352		
16	ses	son	DET	Number=Plur PossNumber=Sing PossPerson=3 PronType=Prs	17	lèvres	det	O	353	356		

FIGURE 2 – Extrait du fichier *articleswiki_RC_s-CHAINE-285_3_maillons_Stanza.csv*

Remarque : Un même type de traitement a été réalisé par Wilkens, Oberle, Landragin, et Todirascu (2020), ce qui a donné lieu à la diffusion de fichiers au format CoNLL similaires avec ajout d'informations sur les chaînes et les mentions. Les fichiers sont disponibles à l'adresse suivante : https://github.com/boberle/coreference_databases

Références

- Landragin, F. (2016). Description, modélisation et détection automatique des chaînes de référence (DEMOCRAT). *Bulletin de l'Association Française pour l'Intelligence Artificielle*(92), 11-15. Consulté sur <https://hal.archives-ouvertes.fr/hal-01347949>
- Lattice, LiLPa, ICAR, & IHRIM. (2019). *Democrat*. Consulté sur <https://hdl.handle.net/11403/democrat/v1.1> (ORTOLANG (Open Resources and TOols for LANGuage) –www.ortolang.fr)
- Péry-Woodley, M.-P., Afantenos, S., Ho-Dac, L.-M., & Asher, N. (2011). La ressource ANNODIS, un corpus enrichi d'annotations discursives. *Traitement Automatique des Langues*, 52(3), 71-101.
- Qi, P., Zhang, Y., Zhang, Y., Bolton, J., & Manning, C. D. (2020). Stanza : A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th annual meeting of the association for computational linguistics : System demonstrations*. Consulté sur <https://nlp.stanford.edu/pubs/qi2020stanza.pdf>
- Quignard, M., Le Mené, M., & Landragin, F. (2021, décembre). Élaboration du corpus DEMOCRAT : procédures d'annotation et d'évaluation :. *Langages*, N° 224(4), 25–46. Consulté le 2022-03-08, sur <https://www.cairn.info/revue-langages-2021-4-page-25.htm?ref=doi> doi: 10.3917/lang.224.0025
- Wilkins, R., Oberle, B., Landragin, F., & Todirascu, A. (2020). French coreference for spoken and written language. In *Language resources and evaluation conference (lrec 2020)* (pp. 80–89).

Annexe D

Guide d'annotation de la nature du référent

1 Présentation de la tâche d'annotation

Ce guide décrit la tâche d'annotation qui consiste à annoter la nature du référent dans des textes du 19^{ème}, 20^{ème} et 21^{ème} siècle, relevant de plusieurs genres textuels, constituant une partie du corpus DEMOCRAT (Landragin, 2021).

Les textes dont vous allez annoter la nature des référents ont été annotés en expressions référentielles. Ce sont des expressions linguistiques (noms propres, syntagmes nominaux, pronoms) qui signalent le référent, à savoir l'entité extra-linguistique dont on parle (personnes, objets, idées, concepts, événements etc.).

Annoter la nature de ces référents signifie ici indiquer

- s'il s'agit d'un référent concret (un objet ou une personne), abstrait (un concept) ou autre (cf. 2.1) ;
- si le référent est concret, s'il s'agit ou pas d'un humain (cf. 2.2).

L'annotation porte sur le référent, c'est-à-dire que vous devez penser à l'entité à laquelle l'expression référentielle réfère pour savoir si elle est concrète ou abstraite, humaine ou non humaine. Les référents à annoter sont présentés au sein d'un contexte qui peut vous aider à bien interpréter le référent et à désambiguïser les cas polysémiques.

Les sections qui suivent décrivent les fichiers d'annotation et la procédure à suivre.

1.1 Description des fichiers d'annotation

Vous allez effectuer l'annotation à partir de fichiers `.html` qui figurent dans un répertoire.

Chaque fichier contient 10 segments de texte regroupant plusieurs expressions référentielles qui désignent un même référent (on parle alors de *mentions*). Ces expressions référentielles ont été annotées manuellement dans le cadre du projet DEMOCRAT (Landragin, 2021).

Chaque segment est séparé par des astérisques et organisé en 5 parties, comme illustré dans l'exemple ci-dessous. Vous trouverez dans l'ordre :

- l'identifiant du segment du texte : cet élément n'est pas important pour l'annotation ;
- un premier bloc appelé *Têtes lexicales et/ou verbes en cas de sujet zéro* : il s'agit de la tête lexicale de chaque expression référentielle annotée dans le segment de texte. Il se peut que la mention annotée corresponde à un verbe. Ces cas indiquent la présence d'un sujet zéro, c'est-à-dire que le sujet est instancié avant le premier verbe de la phrase et n'est pas répété avant les verbes qui suivent (***Tou** se glisse entre eux dans l'espace libre et **marche** sagement* vs ***Tou** se glisse entre eux dans l'espace libre et **il** marche sagement*). Donc, pour annoter le sujet de tous les verbes la solution adoptée a consisté à annoter le groupe verbal ;
- le segment de texte dans lequel les mentions apparaissent. Dans chaque segment les mentions sont signalées en gras et en couleur ;
- un deuxième bloc appelé *Concret / Abstrait* où vous allez cocher la case correspondant à la nature du référent (cf. section 2) ;
- Si vous cochez la case *Concret* un troisième bloc appelé *Humain / Non humain* apparaîtra à droite (cf. 1.2).

d'identifier les mentions du référent en contexte. Appuyez-vous sur le contexte pour déterminer la nature du référent, surtout dans les cas polysémiques.

Les segments de texte peuvent être plus ou moins longs et présenter un nombre de mentions plus ou moins élevé. Si les premières mentions du référent en contexte suffisent pour en indiquer la nature, il n'est pas nécessaire de lire le segment en entier.

L'étape suivante consiste donc à cocher la case correspondant à la nature du référent. Dans un premier temps il vous est demandé d'indiquer si le référent est concret ou abstrait et d'en indiquer le degré (cf. 2.1). Si la nature du référent est abstraite, vous pouvez passer au segment de texte suivant. Si vous avez coché *concret*, il vous sera demandé d'annoter la propriété humain - non humain, comme vous pouvez le voir dans l'exemple ci-dessous.

estrepublikain-2_RC_19-21_s-CHAINE-1738
Têtes lexicales et/ou verbes en cas de sujets zéro
Polochon ; Polochon ; Polochon ; Sa ; le ; chat

Polochon a disparu Polochon a pris la fuite après l'explosion.
Effrayé par l'explosion, Polochon, un petit chat appartenant à Roxane Herbrecht a disparu.
Sa propriétaire le recherche activement.
Si vous apercevez dans le secteur de la rue du Tramway, le chat au pelage de couleur noire et pourvu d'un collier avec une petite plaque, prière de contacter la jeune femme au 06.

Concret / Abstrait
☒ Concret
Tangible — Non Tangible
☐ Abstrait
☐ Terme à facettes

Humain / Non humain
Humain
☐ Individuel
☐ Collectif
☒ Non humain
☐ Doute

Une fois l'annotation terminée vous devez cliquer sur *Enregistrer les annotations*.

Concret / Abstrait
☒ Concret
Tangible — Non Tangible
☐ Abstrait
☐ Terme à facettes

Humain / Non humain
Humain
☒ Individuel
☐ Collectif
☐ Non humain
☐ Doute

Enregistrer les annotations

Un fichier au format `.txt` sera automatiquement téléchargé sur votre ordinateur. Ce fichier devra être déplacé dans le dossier contenant les fichiers `.html`.

Lorsque vous aurez terminé toutes les annotations, vous devez compresser le dossier, déposer l'archive sur l'espace scout prévu à cet effet et notifier le dépôt à l'adresse mail suivante : silvia.federzoni@univ-tlse2.fr

2 Nature des référents : propriétés à annoter

Cette section décrit et illustre les différentes propriétés à annoter pour les référents mentionnés dans les textes. Pour rappel, l'annotation porte sur le référent, c'est-à-dire que vous devez penser à l'entité à laquelle la mention réfère pour savoir si elle est concrète ou abstraite, humaine

ou non humaine. Le contexte peut vous aider à désambiguïser les mots polysémiques.

Les définitions données dans cette section vous guident dans l'annotation sachant qu'il n'y a pas de bonne ou de mauvaise réponse : vous devez annoter la nature du référent en fonction de votre interprétation du référent en contexte. Fiez-vous à la représentation que vous vous êtes faite du référent en lisant le segment de texte et sur la base des expressions référentielles utilisées par l'auteur.

Les illustrations proposées sont issues des segments des textes du corpus DEMOCRAT que vous allez annoter. Les exemples sont présentés sous forme de capture d'écran du fichier d'annotation que vous allez utiliser (cf. 1.1).

2.1 Concret - abstrait

Tout d'abord vous allez annoter le trait *concret / abstrait*. Pour annoter ce trait, deux choix sont possibles : *concret* et *abstrait*. Vous cocherez une seule réponse.

2.1.1 Concret

Sont à considérer comme *concrets* les référents qui sont facilement perceptibles par nos sens : « certains référents sont évidemment perceptibles par la vue et le toucher (*un cheval, une pomme, une table...*), d'autres par l'ouïe (*un bruit, un son, une mélodie...*) ; d'autres à l'odorat (*parfums et senteurs diverses*), d'autres encore de manière privilégiée au toucher (*l'air ou le vent*) » (Martin, 1996, 42).

Dans le cadre de cette tâche d'annotation, vous allez distinguer deux sous-catégories de ce trait : *tangible* vs *non tangible*. Pour annoter ces deux sous-catégories vous devez positionner le curseur à gauche ou à droite de la barre, selon la sous-catégorie choisie, comme le montrent les exemples ci-dessous.

- **Tangible** : sont à considérer comme *tangibles* tous les référents qui sont facilement perceptibles par le toucher (*une table, un chien, une personne*). Les référents *tangibles* possèdent une forme, une couleur, un poids, etc. Ils peuvent être touchés, vus ou manipulés et ils peuvent être localisés dans le temps et dans l'espace (Galmiche & Kleiber, 1996).

The screenshot shows a web-based annotation interface. At the top, the text 'ventredeparis_RC_19-21_s-CHAINE-1212' is displayed in red. Below it, a box contains the text 'Têtes lexicales et/ou verbes en cas de sujets zéro' with the example 'l'usage / tasses - que'. The main text area contains a paragraph: 'La femme, armée d'une cuiller à pot, prenant de minces tranches de pain au fond d'une corbeille garnie d'un linge, trempait la soupe dans des tasses jaunes.. [...] Florent tournait la tête, gêné par ces tasses pleines, que les consommateurs vidaient sans mot dire, avec un regard de côté d'animaux méfiants..'. Below the text, there are two panels. The left panel, titled 'Concret / Abstrait', has a 'Concret' checkbox checked and a slider bar positioned towards the 'Tangible' end. The right panel, titled 'Humain / Non humain', has a 'Non humain' checkbox checked. Other options like 'Individuel', 'Collectif', 'Doute', and 'Terme à facettes' are unchecked.

- **Non tangible** : sont à considérer comme *non tangibles* les référents qui ne possèdent pas une forme, une couleur ou un poids. Ces référents ne peuvent pas être manipulés mais sont perceptibles par certains de nos sens, notamment par l'ouïe (*un son, un bruit, une mélodie*), par l'odorat (*un parfum*) ou « de manière privilégiée au toucher (*l'air ou le vent*) » (Martin, 1996, 42).

jeanchristophe-1_RC_19-21_s-CHAINE-1143

Têtes lexicales et/ou verbes en cas de sujets zéro

chanson ; dont ; musique ; et

Mais le suprême bonheur est lorsque maman vient, qu' elle prend la main de Christophe assoupi, et que, penchée sur lui, à sa demande, elle chante à mi-voix **une vieille chanson, dont** les mots ne veulent rien dire..
Le père trouve **cette musique** stupide ; mais Christophe ne s' **en** lasse pas..

Concret / Abstrait

☒ Concret

Tangible Non Tangible

☐ Abstrait

☐ Terme à facettes

Humain / Non humain

Humain

☐ Individuel

☐ Collectif

☒ Non humain

☐ Doute

Remarque : si vous ne savez vraiment pas où positionner le curseur vous pouvez le laisser au milieu, ce qui vous permettra de signaler le fait que vous doutez. Utilisez cette dernière solution seulement dans les cas où vous n'arrivez vraiment pas à trancher. Sinon, fiez-vous à votre interprétation : rappelez-vous qu'il n'y a pas de bonne ou de mauvaise réponse.

2.1.2 Abstrait

Sont à considérer *abstrait* les référents qui « échapp[ent] aux cinq sens de la vue, de l'ouïe, de l'odorat, du goût et du toucher » (Wilmet, 1996), comme par exemple un concept, une idée, des propriétés, les émotions ou les sentiments.

Dans le cadre de cette annotation vous allez classer les référents abstraits sur une échelle, en distinguant *non perceptible* vs *perceptible indirectement*. Pour annoter ces deux sous-catégories vous devez positionner le curseur à gauche ou à droite de la barre, selon la sous-catégorie choisie, comme le montrent les exemples ci-dessous.

- **Non perceptible** : il s'agit de référents qui ne sont pas perceptibles par nos sens, comme par exemple un concept, une idée, un sentiment ou une émotion.

elisabethseton_RC_19-21_s-CHAINE-341

Têtes lexicales et/ou verbes en cas de sujets zéro

loutellexicales

Actions de grâces, louanges à Dieu, **inexprimable joie** pour n'avoir pas commis cet acte horrible.. [...] Mme Seton avait trouvé dans le mariage plus de **joie** que les plus illusionnés n'en osent espérer.. [...] « **Joie**, étonnement, ravissement pour moi, de voir que ce secours sur lequel j'avais compté si tendrement, avec une foi si affermie, une espérance si abandonnée, l'heure de l'épreuve étant venue, me soutenait, me consolait au delà de tout ce que j'avais pu espérer, même concevoir!! »

Concret / Abstrait

☐ Concret

☒ Abstrait

Non

perceptible

Perceptible

indirectement

☐ Terme à facettes

- **Perceptible indirectement** : les émotions ou les sentiments peuvent avoir une manifestation extérieure perceptible par un de nos sens. Dans ces cas le référent est à considérer comme *perceptible indirectement* : « s'il est vrai que l'on peut interpréter certaines manifestations extérieures comme des indices de tel ou tel sentiment, il n'en reste pas moins que les sentiments ne font jamais l'objet d'une perception directe » (Anscombe, 1996, 257).

jeanchristophe-1_RC_19-21_s-CHAINE-406

Têtes lexicales et/ou verbes en cas de sujets zéro

joie ; me ; joie ; joie

— Et, parmi ce chaos, la lumière des yeux amis qui lui sourient, le flot de **joie** qui, du corps maternel, du sein gonflé de lait, se répand dans sa chair, la force qui est en lui et qui s' amasse énorme, inconsciente, l' océan bouillonnant qui gronde dans l' étroite prison de ce petit corps d' enfant.. [...] Le cheval filait vite, et Christophe riait de **joie**, à moins qu' on ne vint à croiser d' autres promeneurs : alors, il prenait un air grave et dégagé, comme quelqu' un qui est habitué à aller en voiture ; mais son cœur était inondé d' orgueil.. Joie de manger, bonheur incomparable, enthousiasme religieux, trépignements de **joie**!! ... Quelle surabondance de force, de **joie**, d' orgueil, en ce petit être!!

Concret / Abstrait

☐ Concret

☒ Abstrait

Non

perceptible

Perceptible

indirectement

☐ Terme à facettes

Dans l'exemple ci-dessus on trouve dans le contexte des indices qui nous permettent de comprendre que la *joie* est perceptible au travers de manifestations physiques, telles le rire (*Christophe riait de joie*) ou des trépignements (*trépignements de joie*).

Remarque : si vous ne savez vraiment pas où positionner le curseur vous pouvez le laisser au milieu, ce qui vous permettra de signaler le fait que vous doutez. Utilisez cette dernière solution seulement dans les cas où vous n'arrivez vraiment pas à trancher. Sinon, fiez-vous à votre interprétation : rappelez-vous qu'il n'y a pas de bonne ou de mauvaise réponse.

2.2 Humain - Non humain

Le trait *humain* / *non humain* est à annoter lorsque le référent est concret. Pour annoter ce trait, quatre choix sont possibles : *humain collectif* ou *humain individuel*, *non humain* et *doute*.

2.3 Collectif

Sont à considérer comme *collectifs* les « groupes constitués d'êtres humains (*peuple, magistrature, comité, parti, public, foule, équipe, chambrée, clientèle, racaille, guilde, rassemblement, collectif, noblesse, réseau, armée, secte*, par exemple) » (Lecolle, 2019, 17). Ces groupes présentent une certaine « homogénéité des éléments entre eux. [...] les membres d'une mouvance, d'une association ou d'un parti, par exemple, sont posés comme similaires entre eux de par leur interdépendance, leur but commun, leur caractère interchangeable tant que l'un d'entre eux n'est pas spécifiquement distingué. Dans un orchestre, les musiciens sont, à certains égards, équivalents tant qu'on ne s'intéresse pas à l'un d'entre eux dans sa spécificité (le violoniste, le clarinettiste) » (Lecolle, 2019, 26).

The screenshot shows a linguistic annotation tool interface. At the top, the text 'estrepublicain-2_RC_19-21_s-CHAINE-1363' is displayed in red. Below it, a box contains the phrase 'Têtes lexicales et/ou verbes en cas de sujets zéro' with the word 'tribunal' highlighted in green. The main text area contains a paragraph: 'Ce jour là, le tribunal condamne Franck à deux mois de prison avec sursis et mise à l'épreuve consistant en une obligation de soins. Devant le tribunal correctionnel, le prévenu s'était voulu rassurant. [...] C'était le 13 novembre dernier devant le tribunal correctionnel de Belfort. [...] Titulaire d'une maîtrise en sciences, le jeune homme est en arrêt maladie depuis de longs mois lorsqu'il se présente devant le tribunal.' The word 'tribunal' is highlighted in green in the text. Below the text, there are two main panels. The left panel is titled 'Concret / Abstrait' and contains a slider between 'Tangible' and 'Non Tangible', with 'Concret' selected. Below the slider are checkboxes for 'Abstrait' and 'Terme à facettes'. The right panel is titled 'Humain / Non humain' and contains checkboxes for 'Humain' (with sub-options 'Individuel' and 'Collectif', where 'Collectif' is selected), 'Non humain', and 'Doute'.

Les toponymes sont considérés comme des référents humains collectifs lorsqu'ils font référence à l'ensemble des personnes, comme par exemple dans

La France manifeste contre la réforme des retraites.

Lorsque les toponymes, ou d'autres mentions collectives (comme par exemple *pays, état*), ne réfèrent pas aux êtres humains, ils sont à considérer comme non humains (cf. 2.3.2). Par exemple dans

La France est située à l'extrémité occidentale de l'Europe

Le pays s'étend sur une superficie totale de 672 051 km²

On peut aussi trouver des cas où l'occurrence du toponyme peut être interprétée comme un référent abstrait, comme dans *La France est le pays des droits de l'homme*. Dans ces cas vous aurez annoté le référent comme abstrait et vous n'aurez pas besoin d'annoter le caractère *humain* / *non humain*.

Il peut être parfois difficile de dire avec précision si la mention fait référence aux êtres humains, aux entités non humaines ou si le référent est abstrait. Par exemple, dans le segment suivant *la Cour internationale de justice* peut être interprétée comme l'ensemble des juges qui la composent ou bien comme instance judiciaire.

conventionmilieumarin_RC_19-21_s-CHAINE-971

Têtes lexicales et/ou verbes en cas de sujets zéro

<COUR / COUR / COUR>

a) Si, dans un délai de deux mois après la nomination du deuxième arbitre, le président du tribunal arbitral n'est pas désigné, le président de la **Cour internationale de justice** procède, à la requête de la partie la plus diligente, à sa désignation dans un nouveau délai de deux mois. . b) Si, dans un délai de deux mois après la réception de la requête, l'une des parties au différend ne procède pas à la nomination d'un arbitre, l'autre partie peut saisir le président de la **Cour internationale de justice**, qui désigne le président du tribunal arbitral dans un nouveau délai de deux mois. [...] Passé ce délai, il saisit le président de la **Cour internationale de justice**, qui procède à cette nomination dans un nouveau délai de deux mois.

Concret / Abstrait

☒ Concret

Tangible

Non

Tangible

☐ Abstrait

☐ Terme à facettes

Humain / Non humain

Humain

☐ Individuel

☒ Collectif

☐ Non humain

☒ Doute

Le contexte devrait vous permettre de résoudre l'ambiguïté. Si cela n'est pas le cas, vous choisirez la réponse qui vous semble correspondre le plus à votre interprétation du texte. Vous trancherez donc selon votre propre interprétation.

2.3.1 Individuel

Sont à considérer comme humain *individuel* les individus, qu'ils soient seuls ou plusieurs.

madamedehautefort_RC_19-21_s-CHAINE-1071

Têtes lexicales et/ou verbes en cas de sujets zéro

<dauphin / roi / dauphin>

Après les couches de la reine, Mme de La Flotte, qui n' avait pas l' humeur aussi désintéressée que sa petite-fille, désira vivement monter de sa place de dame d' atours à celle de gouvernante **du petit dauphin**. [...] Richelieu n' était pas homme à remettre **le futur roi** entre les mains de ses ennemis, et il avait déjà fait nommer à cet emploi important Mme de Lansac, qui lui était toute dévouée. [...] Un jour, elle reçut de sa part le portrait **du petit dauphin** comme un présage de jours meilleurs.

Concret / Abstrait

☒ Concret

Tangible

Non

Tangible

☐ Abstrait

☐ Terme à facettes

Humain / Non humain

Humain

☒ Individuel

☐ Collectif

☐ Non humain

☐ Doute

Remarque :

L'exemple ci-dessus porte sur un référent humain, *roi*, signalé par les expressions référentielles *le petit dauphin* (2 mentions) et *le futur roi*. Dans l'exemple les expressions référentielles sont utilisées en emploi spécifique, c'est-à-dire qu'elles renvoient à un référent spécifique. Il est possible que vous rencontriez des expressions référentielles en emploi générique, c'est-à-dire que le référent visé n'est pas un individu spécifique mais il est question d'un individu quelconque, de l'espèce, de la classe.

La tâche d'annotation se focalisant sur la nature du référent, et non pas sur le type d'emploi des expressions référentielles, il est important de ne pas confondre les deux niveaux : quel que soit le type d'emploi des expressions référentielles, votre annotation doit porter sur le référent.

codecivildesfrenches-2_RC_19-21_s-CHAINE-869

Têtes lexicales et/ou verbes en cas de sujets zéro

individu : il ; individu : il ; l'enfant : son ; son : il ; lui : son ; enfant : réclamant : il ; il

La possession d'état s'établit par une réunion suffisante de faits qui indiquent le rapport de filiation et de parenté entre **un individu** et la famille à laquelle **il** prétend appartenir. .

Les principaux de ces faits sont, Que **l'individu** a toujours porté le nom du père auquel **il** prétend appartenir ; Que le père l'a traité comme **son enfant**, et a pourvu, en cette qualité, à **son** éducation, à **son** entretien et à **son** établissement ; Qu'il a été reconnu constamment pour tel dans la société ; Qu'il a été reconnu pour tel par la famille. [...]

Nul ne peut réclamer un état contraire à celui que **lui** donnent **son** titre de naissance et la possession conforme à ce titre ; Et réciproquement, nul ne peut contester l'état de celui qui a une possession conforme à **son** titre de naissance. . [...]

À défaut de titre et de possession constante, ou si **l'enfant** a été inscrit, soit sous de faux noms, soit comme né de père et mère inconnus, la preuve de filiation peut se faire par témoins. . [...]

La preuve contraire pourra se faire par tous les moyens propres à établir que **le réclamant** n'est pas l'enfant de la mère qu'il prétend avoir, ou même, la maternité prouvée, qu'il n'est pas l'enfant du mari de la mère.

Concret / Abstrait

☒ Concret

☐ Abstrait

☐ Terme à facettes

Humain / Non humain

Humain

☒ Individuel

☐ Collectif

☐ Non humain

☐ Doute

codexvildesfrançais-I_RC_19-21_s-CHAINE-562

Têtes lexicales et/ou verbes en cas de sujets zéro

(têtes lexicales : (des) pères / (des) mères ; mots : adjectifs : public / de l'état / les / pleins / pleins / tout / même / pare / père / plus / pour / tous / autres / parents / autres :

Ces publications, et l'acte qui en sera dressé, énonceront les prénoms, noms, professions et domiciles, des futurs époux, leur qualité de majeurs ou de mineurs, et les prénoms, noms, professions et domiciles de **leurs pères et mères**. [...]

L'acte de notoriété contiendra la déclaration faite par sept témoins de l'un ou de l'autre sexe, parents ou non parents, des prénoms, nom, profession et domicile du futur époux, et de ceux de **ses père et mère**, s'ils sont connus : le lieu, et, autant que possible, l'époque de sa naissance, et les causes qui empêchent d'en rapporter l'acte... [...]

L'acte authentique du consentement **des pères et mères** ou aïeuls et aïeules, ou, à leur défaut, celui de la famille, contiendra les prénoms, noms, professions et domiciles du futur époux et de tous ceux qui auront concouru à l'acte, ainsi que leur degré de parenté. [...]

Les prénoms, noms, professions et domiciles **des pères et mères** : [...]

Le consentement **des pères et mères**, aïeuls et aïeules, et celui de la famille, dans les cas où ils sont requis ; [...]

Le fils qui n'a pas atteint l'âge de vingt-cinq ans accomplis, la fille qui n'a pas atteint l'âge de vingt-un ans accomplis, ne peuvent contracter mariage sans le consentement de **leurs père et mère** : en cas de dissentiment, le consentement du père suffit. [...]

Si l'un **des deux** est mort, ou s'il est dans l'impossibilité de manifester sa volonté, le consentement de l'autre suffit. [...]

Si **le père et la mère** sont morts, ou s'ils sont dans l'impossibilité de manifester leur volonté, les aïeuls et aïeules les remplacent : s'il y a dissentiment entre l'aïeul et l'aïeule de la même ligne, il suffit du consentement de l'aïeul. [...]

Les enfants de famille ayant atteint la majorité fixée par l'article 148, sont tenus, avant de contracter mariage, de demander, par un acte respectueux et formel, le conseil de **leur père et de leur mère**, ou celui de leurs aïeuls et aïeules, lorsque **leur père et leur mère** sont décédés, ou dans l'impossibilité de manifester leur volonté. [...]

Les officiers de l'état civil qui auraient procédé à la célébration des mariages contractés par des fils n'ayant pas atteint l'âge de vingt-cinq ans accomplis, ou par des filles n'ayant pas atteint l'âge de vingt-un ans accomplis, sans que le consentement **des pères et mères**, celui des aïeuls et aïeules, et celui de la famille, dans les cas où ils sont requis, soient énoncés dans l'acte de mariage, seront, à la diligence des parties intéressées et du commissaire du Gouvernement près le tribunal de première instance du lieu où le mariage aura été célébré, condamnés à l'amende portée par l'article 192, et, en outre, à un emprisonnement dont la durée ne pourra être moindre de six mois. [...]

Les dispositions contenues aux articles 148 et 149, et les dispositions des articles 151, 152, 153, 154 et 155, relatives à l'acte respectueux qui doit être fait **aux père et mère** dans le cas prévu par ces articles, sont applicables aux enfants naturels légalement reconnus. [...]

L'enfant naturel qui n'a point été reconnu, et celui qui après l'avoir été a perdu **ses père et mère**, ou dont **les père et mère** ne peuvent manifester leur volonté, ne pourra, avant l'âge de vingt-un ans révolus, se marier qu'après avoir obtenu le consentement d'un tuteur ad hoc qui lui sera nommé. [...]

Le père, et à défaut du père, la mère, et à défaut de **père et mère**, les aïeuls et aïeules, peuvent former opposition au mariage de **leurs** enfants et descendants, encore que ceux-ci aient vingt-cinq ans accomplis. [...]

Le mariage contracté sans le consentement **des père et mère**, des ascendants, ou du conseil de famille, dans les cas où ce consentement était nécessaire, ne peut être attaqué que par ceux dont le consentement était requis, ou par celui des deux époux qui avait besoin de ce consentement. [...]

L'action en nullité ne peut plus être intentée ni par les époux, ni par les **parents** dont le consentement était requis, toutes les fois que le mariage a été approuvé expressément ou tacitement par ceux dont le consentement était nécessaire, ou lorsqu'il s'est écoulé une année sans réclamation de leur part, depuis qu'ils ont eu connaissance du mariage. [...]

Tout mariage qui n'a point été contracté publiquement, et qui n'a point été célébré devant l'officier public compétent, peut être attaqué par les époux eux-mêmes, par **les père et mère**, par les ascendants, et par tous ceux qui y ont un intérêt né et actuel, ainsi que par le ministère public.

Concret / Abstrait

☒ Concret

☐ Abstrait

☐ Terme à facettes

Humain / Non humain

Humain

☒ Individuel

☐ Collectif

☐ Non humain

☐ Doute

Les référents comme *Dieu*, *Diable* sont à annoter comme humains (et donc concrets) seulement dans les cas où ils sont décrits de façon concrète, comme ayant une apparence humaine et agissant comme des humains. Dans tous les autres cas ces référents sont considérées comme abstraits et donc vous ne devrez pas annoter le trait humain / non humain.

2.3.2 Non humain

Tout référent concret qui n'est pas un être humain, par exemple *les objets* ou *les animaux*. Les parties du corps sont également à considérer comme des référents non humains.

estrepublikain-2_RC_19-21_s-CHAINE-1738

Têtes lexicales et/ou verbes en cas de sujets zéro

Polochon ; Polochon ; Polochon ; Sa ; le ; chat

Polochon a disparu. Polochon a pris la fuite après l'explosion.
Effrayé par l'explosion, Polochon, un petit chat appartenant à Roxane Herbrecht a disparu.
Sa propriétaire le recherche activement.
Si vous apercevez dans le secteur de la rue du Tramway, le chat au pelage de couleur noire et pourvu d'un collier avec une petite plaque, prière de contacter la jeune femme au 06.

Concret / Abstrait

☒ Concret

Tangible Non Tangible

☐ Abstrait

☐ Terme à facettes

Humain / Non humain

Humain

☐ Individuel

☐ Collectif

☒ Non humain

☐ Doute

2.3.3 Doute

Cette catégorie est à utiliser exclusivement s'il vous est impossible de classer le référent selon les trois catégories présentées précédemment. L'idéal serait d'indiquer dans tous les cas une catégorie et si vous n'êtes pas convaincus, vous pouvez quand même signaler le doute. Vous avez la possibilité de cocher deux cases dans ce cas là, comme illustré ci-dessous.

conventionmilieumarin_RC_19-21_s-CHAINE-971

Têtes lexicales et/ou verbes en cas de sujets zéro

COUR ; COUR ; COUR

a) Si, dans un délai de deux mois après la nomination du deuxième arbitre, le président du tribunal arbitral n'est pas désigné, le président de la Cour internationale de justice procède, à la requête de la partie la plus diligente, à sa désignation dans un nouveau délai de deux mois. . b) Si, dans un délai de deux mois après la réception de la requête, l'une des parties au différend ne procède pas à la nomination d'un arbitre, l'autre partie peut saisir le président de la Cour internationale de justice, qui désigne le président du tribunal arbitral dans un nouveau délai de deux mois. [...] Passé ce délai, il saisit le président de la Cour internationale de justice, qui procède à cette nomination dans un nouveau délai de deux mois.

Concret / Abstrait

☒ Concret

Tangible Non Tangible

☐ Abstrait

☐ Terme à facettes

Humain / Non humain

Humain

☐ Individuel

☒ Collectif

☐ Non humain

☒ Doute

2.4 Cas particuliers

Certains référents peuvent être signalés par des termes à facettes de sens, comme par exemple *roman*, *livre*, *déclaration* (Kleiber, 1997). Ces noms ont une facette qui fait référence à l'aspect matériel, à l'objet concret et une autre facette qui fait référence au contenu informationnel. Les différents sens sont tous pertinents pour l'interprétation, mais généralement une seule facette est mise en avant dans le discours. En vous appuyant sur le contexte, vous annoterez le référent comme *concret* si la facette en focus dans le discours est celle de l'objet concret ; le référent est à considérer comme *abstrait* si l'aspect informationnel (contenu intellectuel) est visé dans le discours.

Par exemple, dans l'exemple suivant la facette qui est visée pour les occurrences de *la collection* et *le livre* est celle de l'aspect matériel. Les verbes *porter* et *apporter* fournissent des indications pour interpréter le référent dans ce sens.

diablaucorps_RC_19-21_s-CHAINE-863
Têtes lexicales et/ou verbes en cas de sujets zéro
collection, livre, livre

Je promis à Marthe de lui porter, le jeudi suivant, la collection du journal *Le Mot et Une saison en enfer*. [...] Pourtant, le mardi soir, ne pouvant attendre, je sus trouver à ma faiblesse de bonnes excuses qui me permirent de porter après le dîner le livre et les journaux. [...] Rougissant, je priai Mme Grangier de m'excuser de la déranger à pareille heure, comme s'il eût été une heure du matin : ne pouvant venir jeudi, j'apportais le livre et les journaux à sa fille.

Concret / Abstrait
☒ Concret
Tangible Non Tangible
☐ Abstrait
☒ Terme à facettes

Humain / Non humain
Humain
☐ Individuel
☐ Collectif
☒ Non humain
☐ Doute

Néanmoins, si dans le discours les deux facettes sont successivement mises en avant (certaines mentions visent l'aspect matériel et d'autres visent l'aspect informationnel dans le même segment de texte) vous allez annoter la nature du référent en fonction de la première expression référentielle rencontrée dans la texte. Si cette mention vise le contenu matériel, vous allez donc considérer que le référent est concret, si elle vise le contenu informationnel, vous considérerez le référent comme abstrait.

Dans tous les cas, en plus de l'annotation *concret* / *abstrait* vous devez cocher la case *terme à facette*.

Références

- Anscombre, J.-C. (1996). Noms de sentiment, noms d'attitude et noms abstraits. In N. Flaux, M. Glatigny, & D. Samain (Eds.), *Les noms abstraits* (pp. 257–273). Presses universitaires du Septentrion. doi: 10.4000/books.septentrion.115975
- Galmiche, M., & Kleiber, G. (1996). Sur les noms abstraits. In N. Flaux, M. Glatigny, & D. Samain (Eds.), *Les noms abstraits* (pp. 23–40). Presses universitaires du Septentrion. doi: 10.4000/books.septentrion.115750
- Kleiber, G. (1997, juin). Cognition, sémantique et facettes : Une « histoire » de livres et de... romans. In G. Kleiber & M. Riegel (Eds.), *Les formes du sens. Études de linguistique française médiévale et générale offertes à Robert Martin à l'occasion de ses 60 ans* (pp. 219–231). Louvain-la-Neuve : Duculot. doi: 10.3917/dbu.kleib.1996.01
- Landragin, F. (Ed.). (2021). *Un corpus annoté en chaines de référence et son exploitation : le projet DEMOCRAT* (N° 224 (décembre 2021)). Malakoff : Armand Colin. doi: 10.3917/lang.224.0011
- Lecolle, M. (2019). *Les noms collectifs humains en français : Enjeux sémantiques, lexicaux et discursifs*. Limoges : Lambert-Lucas.
- Martin, R. (1996). Le fantôme du nom abstrait. In N. Flaux, M. Glatigny, & D. Samain (Eds.), *Les noms abstraits* (pp. 41–50). Presses universitaires du Septentrion. doi: 10.4000/books.septentrion.115760
- Wilmet, M. (1996). À la recherche du nom abstrait. In N. Flaux, M. Glatigny, & D. Samain (Eds.), *Les noms abstraits* (pp. 67–76). Presses universitaires du Septentrion. doi: 10.4000/books.septentrion.115775

Annexe E

Séquences représentatives : exemples

E.1 Cluster C2a-11c DemRC

- (54) a. ▼ Le comité exerce ses fonctions conformément aux règles de **procédures**, directives et instructions adoptées par la Commission. [...] La Commission nomme, conformément aux règles de **procédure** adoptées et en tenant compte de tous les critères qui y sont établis, un directeur dont les compétences dans le domaine de la présente convention sont établies et généralement reconnues, en particulier en ce qui concerne ses aspects scientifiques, techniques et administratifs, qui rend compte à la Commission et que la Commission peut révoquer à sa discrétion. [...]

d) nommer, révoquer et diriger le personnel administratif, scientifique, technique et autre, nécessaire à l'exercice des fonctions de la Commission, conformément aux règles de **procédure** adoptées par la Commission; [...]

Ces participants doivent avoir accès en temps opportun aux informations pertinentes, sous réserve des règles de **procédure** et de confidentialité adoptées par la Commission en ce qui concerne l'accès à ce type d'informations. [...]

Chaque partie fournit à la Commission toutes les informations nécessaires à la réalisation de l'objectif de la présente convention, y compris les informations statistiques et biologiques et celles relatives à ses activités de pêche dans la zone de la convention, et met à disposition de la Commission les informations concernant les actions entreprises pour appliquer les mesures adoptées conformément à la présente convention, lorsque la Commission le requiert et en tant que de besoin, sous réserve des dispositions de l'article XXII de la présente convention et conformément aux règles de **procédure** élaborées et adoptées par la Commission. [...]

Chaque partie :

- a) autorise l'utilisation et la diffusion, sous réserve des règles de confidentialité applicables, des informations pertinentes collectées par des observateurs embarqués de la Commission ou d'un programme national;

- b) veille à ce que les propriétaires et / ou les capitaines des navires autorisent la Commission, conformément aux règles de **procédure** adoptées par celle-ci dans ce domaine, à collecter et analyser les informations nécessaires à l'exercice des fonctions du comité chargé de l'examen de l'application des mesures adoptées par la Commission ; [...] [TJ_{CTT} ; ⚡]
- b. ▲ Les actes d' **appel** seront réciproquement signifiés tant à l'autre époux qu'au commissaire du Gouvernement près du tribunal de première instance. [...] Dans les dix jours à compter de la signification qui lui aura été faite du second acte d' **appel**, le commissaire du Gouvernement près le tribunal de première instance fera passer au commissaire du Gouvernement près du tribunal d' **appel**, l'expédition du jugement, et les pièces sur lesquelles il est intervenu.
Le commissaire près du tribunal d' **appel** donnera ses conclusions par écrit, dans les dix jours qui suivront la réception des pièces ; le président, ou le juge qui le suppléera, fera son rapport au tribunal d' **appel**, en la chambre du conseil, et il sera statué définitivement dans les dix jours qui suivront la remise des conclusions du commissaire. [TJ_{CCF} ; ⚡]
- c. ◆ Les tribunaux civils seront seuls compétents pour statuer sur les réclamations d' **état** [...] L'action criminelle contre un délit de suppression d' **état** , ne pourra commencer qu'après le jugement définitif sur la question d' **état** [...] L'action en réclamation d' **état** est imprescriptible à l'égard de l'enfant. [TJ_{CCF} ; ⚡]
- d. ■ Les registres seront cotés par première et dernière, et paraphés sur chaque feuille, par le président du tribunal de **première instance** , ou par le juge qui le remplacera. [...] Les registres seront clos et arrêtés par l'officier de l'état civil, à la fin de chaque année ; et dans le mois ; l'un des doubles sera déposé aux archives de la commune, l'autre au greffe du tribunal de **première instance** [...] Les extraits délivrés conformes aux registres, et légalisés par le président du tribunal de **première instance** , ou par le juge qui le remplacera, feront foi jusqu'à inscription de faux. [...] Dans tous les cas où la mention d'un acte relatif à l'état civil devra avoir lieu en marge, d'un autre acte déjà inscrit, elle sera faite à la requête des parties intéressées, par l'officier de l'état civil, sur les registres courants ou sur ceux qui auront été déposés aux archives de la commune, et par le greffier du tribunal de **première instance** sur les registres déposés au greffe à l'effet de quoi l'officier de l'état civil en donnera avis dans les trois jours au commissaire du Gouvernement près ledit tribunal, qui veillera à ce que la mention soit faite d'une manière uniforme sur les deux registres. [...] Toute contravention aux articles précédents de la part des fonctionnaires y dénommés sera poursuivie devant le tribunal de **première instance** , et punie d'une amende qui ne pourra excéder cent francs. [...] Le commissaire du Gouvernement près le tribunal de **première instance** sera tenu de vérifier l'état des registres lors du dépôt qui en sera fait au greffe il dressera un procès verbal sommaire de la vérification dénoncera les contraventions ou délits commis par les officiers de l'état civil et requerra contre eux la condamnation aux amendes. [...] Dans tous les cas où un tribunal de **première instance** connaîtra des actes relatifs à l'état civil, les parties intéressées pourront se pourvoir contre le jugement. [...] [TJ_{CCF} ; ⚡]

- e.  Article R15-23 [...]
 7° Les brigades de recherches, les brigades de renseignements et d'investigations judiciaires, les brigades de surveillance du littoral, les unités navigantes de la gendarmerie maritime, les brigades ou postes de la gendarmerie maritime placés auprès des directions interrégionales de la mer, les pelotons de sûreté maritime et portuaire, le peloton de gendarmerie maritime de Paris ; [...]
 Article R15-24 [...]
 2° Les brigades départementales de renseignements et d'investigations judiciaires de la gendarmerie nationale et les brigades de renseignements et d'investigations judiciaires dans les collectivités d'outre-mer et en Nouvelle-Calédonie ; 3° Les brigades motorisées de la gendarmerie départementale ; [...]

[TJ_{CPP} ; 

E.2 Cluster C2b-11c DemRC

- (55) a.  Paragraphe 2 : Habilitation des militaires de la gendarmerie et des fonctionnaires de la police nationale, ayant la qualité d'officier de police judiciaire, à exercer effectivement les attributions attachées à cette qualité A : Habilitation des officiers de police judiciaire de la gendarmerie
 Article R13 Les militaires de la gendarmerie mentionnés à l'article 16 (2°) ne peuvent être habilités à exercer effectivement les attributions attachées à leur qualité d'officier de police judiciaire que lorsqu'ils sont affectés à un emploi comportant l'exercice de ces attributions. [...] Cette requête contient toutes indications utiles sur la décision de suspension ou de retrait de l'habilitation. Article R15-9 Dès réception de la requête, le secrétaire de la commission en transmet copie au procureur général près la Cour de cassation et au procureur général qui a pris la décision de suspension ou de retrait d'habilitation. [TJ_{CPP} ; 

b.  Sauvé de la destitution, le président russe reprend le combat contre la Douma pour imposer son candidat au poste de Premier ministre, Sergueï Stépachine. [...] Boris Eltsine s'apprête à livrer cette semaine une nouvelle bataille à la Douma, chambre basse du parlement russe, pour imposer son candidat au poste de Premier ministre, Sergueï Stépachine, après avoir échappé samedi à la procédure de destitution lancée contre lui par les communistes. [...] Affaibli physiquement, crédité de 2% d'opinions favorables dans les sondages, Eltsine est paradoxalement en position de force face à la majorité communiste de la Douma, qui n'a pas réussi samedi à faire voter son projet de destitution du président. [AP_{ER} ; 

c.  Les parties contractantes font en sorte : a) qu'aucun déchet ou autre matière mentionné au paragraphe 2 de l'article 3 de la présente annexe ne soit immergé sans autorisation de leurs autorités compétentes ou sans réglementation ; b) que cette autorisation ou cette réglementation soit conforme aux critères, aux lignes directrices et aux procédures pertinents et applicables, adoptés par la Commission conformément à l'article 6 de la présente annexe ; [...] Cette autorisation ou cette réglementation est conforme aux critères, aux lignes directrices et aux procédures pertinents et applicables, adoptés par la Commission conformément à l'article 6 de

- la présente annexe [...] Article 6 Aux fins de la présente annexe, il incombe à la Commission notamment d'élaborer et d'adopter des critères, des lignes directrices et des procédures pour l'immersion de déchets ou d'autres matières énumérés au paragraphe 2 de l'article 3 et pour le dépôt des matières visées à l'article 5 de la présente annexe, dans le but de prévenir et de supprimer la pollution. [TJ_{CMM} ; ⚡]
- d. ■ – Il l'appelait : scélérat, bête féroce, homme sans moralité. Et si ce langage avait pour objet de rétablir dans l'esprit de l'enfant le sens de la justice, il faut avouer qu'il manquait son but ; car la logique enfantine risquait fort de conclure : « Si un grand homme comme celui-là n'avait pas de moralité, c'est donc que la moralité n'est pas grand-chose, et que la première affaire, c'est d'être un grand homme. [R_{JC} ; ⚡]
- e. ○ Article 18 Dommage causé à la marchandise 1 [...] Le transport aérien, au sens du paragraphe 1 du présent article, comprend la période pendant laquelle la marchandise se trouve sous la garde du transporteur. [...] Dans le cas de transport intermodal effectué en partie par air et en partie par tout autre moyen de transport, les dispositions de la présente convention ne s'appliquent, sous réserve du paragraphe 4 de l'article 18, qu'au transport aérien et si celui-ci répond aux conditions de l'article 1. [TJ_{CTA} ; ⚡]

E.3 Cluster C3a-11c DemRC

- (56) a. ▼ [...] Convention pour la protection du milieu marin de l'Atlantique du nord-est [...] Rappelant les résultats positifs obtenus dans le contexte de la convention pour la prévention de la pollution marine par les opérations d'immersion effectuées par les navires et aéronefs, signée à Oslo le 15 février 1972, telle qu'amendée par les protocoles du 2 mars 1983 et du 5 décembre 1989, ainsi que de la convention pour la prévention de la pollution marine d'origine tellurique, signée à Paris le 4 juin 1974, telle qu'amendée par le protocole du 26 mars 1986 ; [...]
- q) On entend par « convention d'Oslo » : la convention pour la prévention de la pollution marine par les opérations d'immersion effectuées par les navires et aéronefs, signée à Oslo le 15 février 1972, telle qu'amendée par les protocoles du 2 mars 1983 et du 5 décembre 1989. [...]
- Article 25
- Signature
- La convention est ouverte à la signature à Paris, du 22 septembre 1992 au 30 juin 1993 par :
- a) les parties contractantes à la convention d'Oslo ou à la convention de Paris ; [...]
- La convention entrera en vigueur le trentième jour suivant la date à laquelle toutes les parties contractantes à la convention d'Oslo et toutes les parties contractantes à la convention de Paris auront déposé leur instrument de ratification, d'acceptation, d'approbation ou d'adhésion. [...]
- Sans préjudice des dispositions du paragraphe 1, les décisions, les recommandations et les autres accords adoptés en application de la convention d'Oslo ou de la convention de Paris continuent d'être applicables

et conservent le même caractère juridique, dans la mesure où ils sont compatibles avec la convention ou ne sont pas explicitement abrogés par celle-ci, par toute décision ou, dans le cas des recommandations existantes, par toute recommandation adoptée en application de celle-ci. [

TJ_{CMM} ; 

- b.  Il avait déjà passé quinze jours dans l'état d'extase qui saisit toutes les jeunes imaginations à l'aspect de la reine des ruines, quand, un soir, il entra **au théâtre d'Argentina**, devant **lequel** se pressait une grande foule. [...] Elle s'avança par coquetterie sur le devant **du théâtre**, et salua le public avec une grâce infinie. [...] Sarrasine voulait s'élancer sur **le théâtre** et s'emparer de cette femme. [...] Bientôt, il fut obligé de quitter **le théâtre** [...] En sortant **du théâtre**, un inconnu l'arrêta par le bras. [...] Il reconnut les chanteurs et les cantatrices **du théâtre**, mêlés à des femmes charmantes, tous prêts à commencer une orgie d'artistes qui n'attendait plus que lui. [...] [N_{s} ; 
- c.  Un livre sur **la fièvre jaune** avait porté au loin le nom de Richard Bayley et fortement intéressé le monde scientifique d'alors. [...] Elle y était au mois d'août 1801, quand **la fièvre jaune** éclata à New-York. C'était surtout parmi les pauvres émigrants irlandais que **le fléau** sévissait. [...] Il y prit **la fièvre** : et, après six jours de cruelles souffrances, expira entre les bras de sa fille, en implorant la miséricorde de son Sauveur. [...] Mais, au départ, on avait négligé certaines formalités, et, **la fièvre jaune** sévissant à New-York, l'équipage et les passagers du brick américain allaient être traités comme des pestiférés. [...] On venait d'apprendre, à Livourne, que **la fièvre jaune** sévissait à New-York, et comme nous n'avions pas de certificat de la Santé, il n'y avait rien à répondre. [B_{ES} ; 
- d.  d) On entend par « pollution » : l'introduction par l'homme, directement ou indirectement, de substances ou d'énergie dans la zone maritime, créant ou susceptibles de créer des risques pour la santé de l'homme, des dommages aux ressources biologiques et aux écosystèmes marins, des atteintes **aux valeurs d'agrément** ou des entraves aux autres utilisations légitimes de la mer. [...] Les parties contractantes appliquent :
- a) le principe de précaution, selon lequel des mesures de prévention doivent être prises lorsqu'il y a des motifs raisonnables de s'inquiéter du fait que des substances ou de l'énergie introduites, directement ou indirectement, dans le milieu marin, puissent entraîner des risques pour la santé de l'homme, nuire aux ressources biologiques et aux écosystèmes marins, porter atteinte **aux valeurs d'agrément** ou entraver d'autres utilisations légitimes de la mer, même s'il n'y a pas de preuves concluantes d'un rapport de causalité entre les apports et les effets ; [...]
- Les parties contractantes visées au point b), qui souhaitent encore conserver la possibilité prévue au point b) rendront compte aux réunions de la Commission au niveau ministériel tous les deux ans à compter de 1999 des progrès réalisés en vue de mettre en place des options à terre et des résultats des études scientifiques montrant que toutes les opérations d'immersion éventuelles n'entraîneraient pas de risques pour la santé de l'homme, ne nuiraient pas aux ressources biologiques et aux écosystèmes marins, ne porteraient pas atteinte **aux valeurs d'agrément** et ne gêneraient pas d'autres utilisations légitimes de la mer. [...] Toute autorisation

ou réglementation visée au paragraphe 1 ne permet pas l'immersion de navires ou d'aéronefs contenant des substances qui créent ou sont susceptibles de créer des risques pour la santé de l'homme, des dommages aux ressources vivantes et aux écosystèmes marins, des atteintes **aux valeurs d'agrément**, ou des entraves aux autres utilisations légitimes de la mer. [...] Aucun permis de ce type n'est délivré si les installations offshore désaffectées ou les pipelines offshore désaffectés contiennent des substances qui créent ou sont susceptibles de créer des risques pour la santé de l'homme, des dommages aux ressources vivantes et aux écosystèmes marins, des atteintes **aux valeurs d'agrément** ou une entrave aux autres utilisations légitimes de la mer. [TJ_{CMM} ; 

- e.  Convention relative au renforcement de la Commission interaméricaine du thon tropical établie par la convention de 1949 entre les États-Unis d'Amérique et la République du Costa Rica (" convention d'Antigua ") Les parties à la présente convention : Conscientes que, conformément aux dispositions pertinentes du droit international telles que reflétées dans **la convention des Nations unies sur le droit de la mer de 1982**, tous les États ont l'obligation de prendre les mesures nécessaires à la conservation et à la gestion des ressources biologiques marines, y compris des poissons grands migrateurs, et de coopérer avec d'autres États pour prendre de telles mesures ; Rappelant les droits souverains des États côtiers aux fins de l'exploration et de l'exploitation, de la conservation et de la gestion des ressources biologiques marines dans les zones relevant de la juridiction nationale, tels qu'établis par **la convention sur le droit de la mer**, et le droit qu'ont tous les États à ce que leurs ressortissants pêchent en haute mer conformément à **la convention sur le droit de la mer** ; [...] Prenant note que la cinquantième assemblée générale des Nations unies, conformément à la résolution A / RES / 50 / 24, a adopté l'accord aux fins de l'application des dispositions de **la convention des Nations unies sur le droit de la mer du 10 décembre 1982** relatives à la conservation et à la gestion des stocks de poissons dont les déplacements s'effectuent tant à l'intérieur qu'au-delà des zones économiques exclusives (stocks chevauchants) et des stocks de poissons grands migrateurs (" accord de New York de 1995 ") ; [...]
- 11) " convention sur le droit de la mer " : **la convention des Nations unies sur le droit de la mer du 10 décembre 1982** ;
- 12) " accord de New York de 1995 " : l'accord de 1995 aux fins de l'application des dispositions de **la convention des Nations unies sur le droit de la mer du 10 décembre 1982** relatives à la conservation et à la gestion des stocks de poissons dont les déplacements s'effectuent tant à l'intérieur qu'au-delà de zones économiques exclusives (stocks chevauchants) et des stocks de poissons grands migrateurs ; [...]
- Aucune disposition de la présente convention ne doit porter atteinte ou nuire à la souveraineté ou aux droits souverains des États côtiers liés à l'exploration et à l'exploitation, à la conservation et à la gestion des ressources biologiques marines dans les zones relevant de leur souveraineté ou de leur juridiction nationale tels qu'établis dans **la convention sur le droit de la mer** ou au droit qu'ont tous les États à ce que leurs ressortissants pêchent en haute mer conformément à **la convention sur le droit de**

la mer

[TJ_{CTT} ; ⚡]

E.4 Cluster C4e-11c DemRC

(57) a. ▼ MADAME BAPTISTE

Quand j'entrai dans la salle des voyageurs de la gare de Loubain, mon premier regard fut pour l'horloge. J'avais à attendre deux heures dix minutes l'express de Paris. Je me sentis las soudain comme après dix lieues à pieds ; puis je regardai autour de moi comme si j'allais découvrir sur les murs un moyen de tuer le temps ; puis je ressortis et m'arrêtai devant la porte de la gare, l'esprit travaillé par le désir d'inventer quelque chose à faire. [...]

[N_{MF} ; ⚡]

- b. ▲ Les majeurs qui servent ou travaillent habituellement chez autrui auront le même domicile que la personne qu'ils servent ou chez laquelle ils travaillent, lorsqu'ils demeureront avec elle dans la même maison.

[TJ_{CCF} ; ⚡]

- c. ◆ La vue de cette noce amena Bouvard et Pécuchet à parler des femmes, qu'ils déclarèrent frivoles, acariâtres, têtues. Malgré cela, elles étaient souvent meilleures que les hommes ; d'autres fois elles étaient pires. Bref, il valait mieux vivre sans elles ; aussi Pécuchet était resté célibataire.

[R_{BeP} ; ⚡]

- d. □ Les officiers de l'état civil qui auraient procédé à la célébration des mariages contractés par des fils n'ayant pas atteint l'âge de vingt-cinq ans accomplis, ou par des filles n'ayant pas atteint l'âge de vingt-un ans accomplis, sans que le consentement des pères et mères, celui des aïeuls et aïeules, et celui de la famille, dans le cas où ils sont requis, soient énoncés dans l'acte de mariage, seront, à la diligence des parties intéressées et du commissaire du Gouvernement près le tribunal de première instance du lieu où le mariage aura été célébré, condamnés à l'amende portée par l'article 192, et, en outre, à un emprisonnement dont la durée ne pourra être moindre de six mois. [...] Lorsqu'il n'y aura pas eu d'actes respectueux, dans les cas où ils sont prescrits, l'officier de l'état civil qui aurait célébré le mariage, sera condamné à la même amende, et à un emprisonnement qui ne pourra être moindre d'un mois.

[TJ_{CCF} ; ⚡]

- e. ● Suis -je fou ? [...] Je n'en sais rien, mais j'ai souffert horriblement. J'ai accompli un acte de folie, de folie furieuse, c'est vrai ; mais la jalousie haletante, mais l'amour exalté, trahi, condamné, mais la douleur abominable que j'endure, tout cela ne suffit-il pas pour nous faire commettre des crimes et des folies sans être vraiment criminel par le cœur ou par le cerveau ? [...] j'ai souffert, souffert, souffert d'une façon continue, aiguë, épouvantable. J'ai aimé cette femme d'un élan frénétique ... [...]

[N_{MF} ; ⚡]

E.5 Cluster C4f-11c DemRC

- (58) a. ▼ Titre IX : Dispositions finales
Article XXVII : Signature

1 [...]

En ce qui concerne les organisations régionales d'intégration économique visées **au paragraphe 1 du présent article**, aucun État membre d'une telle organisation ne peut signer la présente convention à moins qu'il ne représente un territoire situé en dehors du champ d'application territorial du traité établissant l'organisation et sous réserve que la participation de cet État membre soit limitée exclusivement à la représentation des intérêts de ce territoire. [...]

a) en signant, pendant la période visée à l'article XXVII, **paragraphe 1**, de la présente convention, un instrument rédigé à cette fin conformément à une résolution que la Commission doit adopter en vertu de la convention de 1949 [...]

[TJ_{CTT} ; ⚡]

- b. ▲ Articles 152, **153**, 154, 155, 156 et 157, décrétés le 21 Ventôse an XII. [...]

Article 153 [...]

Les dispositions contenues aux articles 148 et 149, et les dispositions des articles 151, 152, **153**, 154 et 155, relatives à l'acte respectueux qui doit être fait aux père et mère dans le cas prévu par ces articles, sont applicables aux enfants naturels légalement reconnus.

[TJ_{CCF} ; ⚡]

- c. ◆ [...] Adulte et en bonne santé, la girafe n'a à craindre **que le lion**, bien qu'elle **puisse le** tuer d'un coup de patte.

Par contre, les jeunes qui échappent à la surveillance de leur mère ou isolés du troupeau et les sujets affaiblis par la vieillesse, la maladie ou une blessure, peuvent être la proie de prédateurs **comme les lions**, les hyènes rayées, les hyènes tachetées, les léopards ou les lycaons.

[TJ_{CTT} ; ⚡]

- d. ■ CHAPITRE II

DOCUMENTS ET OBLIGATIONS DES PARTIES RELATIFS AU TRANSPORT DES PASSAGERS, DES BAGAGES ET DES MARCHANDISES

Article 3

Passagers et bagages **1** [...]

L'emploi de tout autre moyen constatant les indications qui figurent **au paragraphe 1** peut se substituer à la délivrance du titre de transport mentionné dans ce paragraphe. [...] Sauf au paragraphe 2 de l'article 1, **au paragraphe 1, alinéa b)**, de l'article 3 [...]

[TJ_{CTA} ; ⚡]

- e. ● Article R7 [...]

3° Le chef de l'inspection générale de la gendarmerie nationale ou son représentant ; [...] Les membres du jury mentionnés aux 1°, 2°, **3°**, 4° et 5° peuvent être différents des membres mentionnés aux 1°, 2°, 4°, 7° et 8° de l'article R. 3.

Le nombre total des magistrats prévus aux 1° et 5° est égal au nombre total des membres prévus aux 2°, **3°**, 4° et 6°.

[TJ_{CPP} ; ⚡]

Annexe F

Analyse des patrons : lien avec les facteurs de variation

F.1 Les patrons AnnoDis et le type de référent

Patron	AnnoDis _{Hum-}		AnnoDis _{Hum+}		AnnoDis		p-value	Significativité
	#	%	#	%	#	%		
P1	24	8,4	42	14,2	66	11,4	4,81e-02	*
P2	24	8,4	22	7,4	46	7,9	7,68e-01	ns
P3	32	11,2	19	6,4	51	8,8	6,76e-02	ns
P4	30	10,5	3	1,0	33	5,7	0,00e+00	***
P5	0	0,0	10	3,4	10	1,8	1,98e-03	**
P6	0	0,0	15	5,1	15	2,6	6,00e-05	***
P7	3	1,1	2	0,7	5	0,9	6,82e-01	ns

TABLEAU F.1 – P-value et degré de significativité de la distribution des patrons AnnoDis selon la nature du référent

F.2 Projection des patrons AnnoDis sur DEMRC : comparaison entre AnnoDis et DEMRC_{Nar-}

Patron	AnnoDis		DemRC _{Nar-}		DemRC _{Nar+}		p-value	Significativité
	#	%	#	%	#	%		
P1	66	11,4	74	3,7	109	4,9	8,21E-36	***
P2	46	7,9	256	12,6	126	5,7	2,41E-13	***
P3	51	8,8	79	3,9	77	3,5	9,20E-18	***
P4	33	5,7	38	1,9	21	1	3,00E-17	***
P5	10	1,7	0	0	1	0	8,56E-16	***
P6	15	2,6	8	0,4	5	0,2	3,77E-14	***
P7	5	0,9	32	1,6	18	0,8	1,52E-03	**

TABLEAU F.2 – P-value et degré de significativité de la distribution des patrons AnnoDis et DEMRC_{Nar-}

F.3 Projection des patrons AnnoDis sur DEMRC en fonction du type de texte

Patron Significativité	DemRC _{Nar+}		DemRC _{Nar-}		p-value
	#	%	#	%	
P1	109	4,9	74	3,7	6.50e-03 **
P2	126	5,7	256	12,6	1.04e-31 ***
P3	77	3,5	79	3,9	0,30261 ns
P4	21	1,0	38	1,9	1.24e-04 ***
P5	1	0,0	0	0,0	1 ns
P6	5	0,2	8	0,4	0,15017 ns
P7	18	0,8	32	1,6	6.95e-04 ***

TABLEAU F.3 – Résultats du test binomial entre DEMRC_{Nar+} et DEMRC_{Nar-} par patron

F.4 Projection des patrons AnnoDis sur DEMRC : comparaison selon la nature du référent

Patron Significativité	DemRC _{Hum+}		DemRC _{Hum-}		p-value
	#	%	#	%	
P1	55	1,8	73	2,4	1,85e-01
P2	88	2,9	147	4,8	4,04e-04
P3	37	1,2	68	2,2	4,54e-03
P4	11	0,4	23	0,7	5,89e-02
P5	0	0,0	1	0,0	1,00e+00
P6	10	0,3	2	0,1	2,04e-02
P7	10	0,3	36	1,2	1,73e-04

TABLEAU F.4 – Résultats du test binomial entre DEMRC_{Hum+} et DEMRC_{Hum-} par patron

Annexe G

Les patrons dans DEMRC

G.1 Distribution des patrons selon le type de texte

Patron	DemRC _{Nar+}		DemRC _{Nar-}	
	#	%	#	%
P1	126	5,70	256	12,64
P2	143	6,47	115	5,68
P3	125	5,66	131	6,47
P4	128	5,79	38	1,88
P5	70	3,17	54	2,67
P6	7	0,32	99	4,89
P7	6	0,27	86	4,25
P8	71	3,21	13	0,64
P9	62	2,81	18	0,89
P10	65	2,94	11	0,54
P11	42	1,90	27	1,33
P12	40	1,81	14	0,69
P13	45	2,04	9	0,44
P14	8	0,36	43	2,12
P15	18	0,81	32	1,58
P16	33	1,49	16	0,79
P17	8	0,36	38	1,88
P18	19	0,86	27	1,33
P19	3	0,14	35	1,73
P20	24	1,09	13	0,64
P21	3	0,14	25	1,23
P22	10	0,45	18	0,89
P23	4	0,18	22	1,09
P24	13	0,59	13	0,64
P25	12	0,54	11	0,54
P26	22	1,00	1	0,05
P27	15	0,68	6	0,30
P28	4	0,18	17	0,84
P29	3	0,14	18	0,89
P30	2	0,09	17	0,84
P31	6	0,27	9	0,44
P32	12	0,54	3	0,15
P33	6	0,27	8	0,40
P34	11	0,50	3	0,15
P35	6	0,27	5	0,25
P36	8	0,36	3	0,15
Total	1236	55,97	2025	100,00

TABLEAU G.1 – Distribution des patrons DEMRC dans DEMRC_{Nar+}
et DEMRC_{Nar-}

G.2 Distribution des patrons selon la nature du référent

Patron	Abstrait		Concret		DemRC _{Hum-}		DemRC _{HumColl+}		DemRC _{HumInd+}	
	#	%	#	%	#	%	#	%	#	%
P1	147	12,8	235	7,6	147	9,4	47	14,3	41	3,5
P2	69	6,0	189	6,1	101	6,4	34	10,3	54	4,5
P3	67	5,8	189	6,1	109	6,9	26	7,9	54	4,5
P4	30	2,6	136	4,4	81	5,2	4	1,2	51	4,3
P5	26	2,3	98	3,2	32	2,0	12	3,6	54	4,5
P6	79	6,9	27	0,9	15	1,0	4	1,2	8	0,7
P7	51	4,4	41	1,3	38	2,4	2	0,6	1	0,1
P8	19	1,7	65	2,1	28	1,8	6	1,8	31	2,6
P9	10	0,9	70	2,3	50	3,2	2	0,6	18	1,5
P10	10	0,9	66	2,1	13	0,8	9	2,7	44	3,7
P11	12	1,0	57	1,8	47	3,0	2	0,6	8	0,7
P12	5	0,4	49	1,6	5	0,3	5	1,5	39	3,3
P13	23	2,0	31	1,0	15	1,0	1	0,3	15	1,3
P14	24	2,1	27	0,9	18	1,1	5	1,5	4	0,3
P15	4	0,3	46	1,5	36	2,3	1	0,3	9	0,8
P16	12	1,0	37	1,2	25	1,6	5	1,5	7	0,6
P17	24	2,1	22	0,7	14	0,9	5	1,5	3	0,3
P18	22	1,9	24	0,8	17	1,1	5	1,5	2	0,2
P19	20	1,7	18	0,6	14	0,9	1	0,3	3	0,3
P20	15	1,3	22	0,7	16	1,0	3	0,9	3	0,3
P21	23	2,0	5	0,2	4	0,3	1	0,3	0	0,0
P22	5	0,4	23	0,7	16	1,0	2	0,6	5	0,4
P23	9	0,8	17	0,6	11	0,7	5	1,5	1	0,1
P24	1	0,1	25	0,8	11	0,7	0	0,0	14	1,2
P25	2	0,2	21	0,7	12	0,8	4	1,2	5	0,4
P26	8	0,7	15	0,5	11	0,7	1	0,3	3	0,3
P27	1	0,1	20	0,6	6	0,4	2	0,6	12	1,0
P28	8	0,7	13	0,4	9	0,6	1	0,3	3	0,3
P29	15	1,3	6	0,2	5	0,3	0	0,0	1	0,1
P30	13	1,1	6	0,2	5	0,3	1	0,3	0	0,0
P31	9	0,8	6	0,2	4	0,3	1	0,3	1	0,1
P32	4	0,3	11	0,4	10	0,6	0	0,0	1	0,1
P33	3	0,3	11	0,4	4	0,3	3	0,9	4	0,3
P34	3	0,3	11	0,4	5	0,3	1	0,3	5	0,4
P35	2	0,2	9	0,3	4	0,3	1	0,3	4	0,3
P36	2	0,2	9	0,3	4	0,3	2	0,6	3	0,3
Total	777		1657		942		204		511	

TABLEAU G.2 – Distribution des patrons DEMRC selon la nature du référent

Titre : Vers une typologie des chaînes de référence à la lumière de corpus annotés diversifiés

Mots clés : Discours, Chaînes de référence, Corpus annotés, Continuité référentielle, TAL, Analyse des séquences

Résumé : Cette thèse s'intéresse aux chaînes de référence (CR) et à leur fonctionnement textuel.

Composées d'expressions référentielles liées entre elles, les CR assurent la continuité référentielle dans le discours et permettent ainsi au lecteur ou à l'interlocuteur de suivre les évolutions des référents dans l'espace textuel, jouant un rôle organisationnel fondamental. Suivant une perspective fonctionnelle, nous considérons que les chaînes de référence contribuent à la création de liens de cohésion.

C'est dans cette optique que nous envisageons notre objet d'étude : les enchaînements d'expressions référentielles. Ces enchaînements sont au cœur des mécanismes de référence et de cohésion, et leur observation fine permet d'accéder aux stratégies discursives mises en place pour introduire, maintenir ou réactiver les référents. L'objectif principal de cette recherche est de dégager des patrons d'enchaînement récurrents et d'interroger leur contribution à l'organisation textuelle. Nous faisons l'hypothèse que ces enchaînements obéissent à des régularités qui reflètent des stratégies discursives dépendant notamment du type ou du genre de texte et de la nature du référent.

Afin de fournir une description systématique des enchaînements des maillons, tout en prenant en compte un large volume de données, nous proposons une méthode outillée de corpus adoptant une approche linéaire — et non plus uniquement globale — des chaînes de référence. Cette méthode combine les techniques du Traitement Automatique des Langues (TAL) avec les outils de l'analyse des séquences, issus des sciences sociales.

L'étude repose sur deux corpus du français écrit annotés selon des modèles différents : AnnoDis et Democrat. L'exploitation de corpus annotés de manière hétérogène constitue un choix méthodologique assumé : plutôt que de percevoir cette hétérogénéité comme un obstacle, nous interrogeons la possibilité d'en tirer parti pour appréhender différents aspects des CR, précisément parce que les phénomènes annotés varient.

Les résultats obtenus sont croisés avec plusieurs facteurs de variation — genre textuel, type de texte, nature du référent — afin d'interroger leur impact sur les formes d'enchaînement observées. Par ailleurs, en s'appuyant sur ces deux corpus diversifiés, notre étude examine l'influence du modèle d'annotation sur la description des CR.

Title: Towards a Typology of Reference Chains in the Light of Diverse Annotated Corpora

Key words: Reference chains, NLP, Referential continuity, Annotated corpora, Discourse, Sequence analysis

Abstract: This dissertation focuses on reference chains (RCs) and their textual functioning.

Composed of referential expressions linked to one another, RCs ensure referential continuity throughout discourse and, through a critical organizational role, enable the reader to follow the evolution of referents across the textual space. From a functional perspective, we consider RCs as contributing to the creation of cohesive ties.

From this perspective, our object of study focuses on sequences of referential expressions. These sequences play a central role in referential and cohesive mechanisms, and their detailed observation sheds light on the discursive strategies used to introduce, maintain, or reactivate referents. The main objective of this research is to identify recurring sequencing patterns and to investigate their contribution to textual organization. We hypothesize that these sequences follow regularities that reflect discursive strategies, depending in particular on the text type, genre, and nature of the referent.

In order to systematically describe sequences of mentions, while also considering a large amount of data, we propose a corpus-based, tool-supported methodology that adopts a linear approach—rather than a purely global one—to the analysis of reference chains. This method combines techniques from Natural Language Processing (NLP) with tools from sequence analysis traditionally used in the social sciences.

The study is based on two written French corpora annotated using different models: AnnoDis and Democrat. The use of heterogeneously annotated corpora is a deliberate methodological choice: rather than treating heterogeneity as a constraint, we explore its potential for capturing diverse aspects of RCs, precisely because the annotated phenomena vary.

The results are analysed in relation to different variation factors—textual genre, text type, and referent type to assess their impact on the observed forms of sequencing. Furthermore, our study investigates the influence of annotation models on the description of RCs.