Chaînes de référence et structure textuelle dans les Essais sur la peinture de Diderot

Céline Guillot-Barbance, Matthieu Quignard

Chaînes de référence et structure textuelle dans les *Essais sur la peinture* de Diderot

Céline Guillot-Barbance

UMR 5371 IRHIM, ENS de Lyon

Matthieu Quignard

UMR 5191 ICAR, CNRS

Résumé

Nous étudions le rapport qu'entretiennent les chaînes de référence avec la structure textuelle des trois premiers chapitres des *Essais sur la peinture* de Diderot. Le corpus est issu de projet ANR DEMOCRAT au sein duquel il a été préparé et annoté. Dans une première partie, nous analysons la distribution des chaînes au sein des unités structurelles (chapitres et paragraphes), afin d'observer leur variabilité en termes de longueur et de couverture (densité référentielle et empan). Dans une seconde partie, nous nous intéressons plus précisément au rapport qu'on peut établir entre les chaînes et le thème développé dans les unités structurelles. Les analyses quantitatives multi-niveaux menées à l'aide de l'extension « annotation URS » du logiciel TXM indiquent que les chaînes de référence jouent un rôle dans le marquage de la continuité et de la discontinuité textuelle. Le thème des chapitres est plutôt associé aux chaînes longues, tandis que le thème des paragraphes s'exprime par des chaînes courtes et souvent discontinues par rapport à celles des autres paragraphes (nouvelles et spécifiques au paragraphe d'occurrence).

Mots-clés : chaînes de référence, marques configurationnelles, structures textuelles, thème de discours, corpus annoté, textométrie

Abstract

This paper addresses the relation between reference chains and structural units, in the three first chapters of the *Essais sur la peinture* by Diderot. This corpus is taken from the DEMOCRAT project, during which it has been prepared and annotated. In a first part, a close analysis addresses the distribution of referents and chains across the three chapters and five comparable paragraphs, in order to observe the variability in lengths, coverage and density. A second part focusses on reference chains with respect to units theme (chapter titles, paragraph themes). Multilevel quantitative analyses have been operated with help of the URS extension of TXM. It appears that reference chains do affect textual continuity and discontinuity. More precisely, chapter themes are rather marked by long and continuous chains whereas paragraph ones are rather marked by short and discontinuous chains (new, local and specific to that unit).

Keywords : reference chains, textual structures, discourse theme, annotated corpus, textometry

Introduction

L'analyse des chaînes de référence a connu ces dernières années un essor important en France grâce à la constitution et à la diffusion de ressources annotées, notamment le corpus ANCOR pour le français parlé et le corpus DEMOCRAT pour le français écrit[1], ainsi qu'au développement d'outils adaptés à leur exploitation[2]. Si l'exploitation de vastes corpus numériques semble particulièrement appropriée pour l'étude de phénomènes textuels de ce type, elle pose cependant des défis qui sont difficiles à relever sur les plans technique, méthodologique et scientifique.

Les recherches menées sur les corpus écrits ont déjà permis de dégager de grandes tendances. La plupart des travaux ont porté sur des textes entiers, souvent courts, ou sur des extraits de taille assez importante pour rendre possible l'analyse des chaînes (désormais CR) en contexte. Ces études ont mis en évidence les spécificités de certains textes ou de certains types de textes relativement à d'autres, en prenant le texte ou l'extrait comme une seule unité. Ce sont les variations diachroniques, inter-linguistiques et génériques qui ont été principalement analysées (Landragin et Schnedecker 2014, Schnedecker 2017, Obry *et al*. 2017).

D'autres études, parfois plus anciennes et portant plus généralement sur la référence et l'anaphore, ont relevé l'importance des frontières textuelles dans le choix des expressions référentielles (notamment Ariel 1990). On s'intéresse en général à ce qui se passe au début ou à la fin des unités, en faisant l'hypothèse que le passage d'une unité à l'autre a une influence sur les CR ou que les CR marquent, ou concourent à marquer, ce passage. Ce sont les catégories grammaticales des expressions référentielles qui sont alors examinées, certaines expressions marquant la continuité référentielle et textuelle, tandis que d'autres indiquent au contraire une forme de discontinuité.

On a donc pris jusqu'ici le texte comme une unité relativement homogène dans les types de CR qu'il instancie, même si l'on admet que certains lieux comme les frontières d'unités structurelles ont un impact, qu'on suppose récurrent à l'échelle de l'unité-texte, sur ces chaînes. Or, même si l'on s'en tient à des caractéristiques assez générales comme le nombre, la longueur ou la composition des CR, rien ne dit qu'il ne puisse y avoir des variations assez importantes selon les parties internes d'un texte. Ce sera notre premier angle d'approche des interférences entre les chaînes et la structure textuelle.

On s'intéressera ensuite aux rapports complexes qu'il est possible d'établir entre thèmes de discours, chaînes de référence et structure du texte. L'étude d'un texte organisé en deux niveaux hiérarchiques d'unités structurelles (chapitres et paragraphes), nous permettra d'étudier les chaînes dans leur relation avec l'identification du thème et avec la diversité thématique des niveaux textuels où ces chaînes se trouvent.

[1] http://tln.li.univ-tours.fr/Tln_Corpus_Ancor.html et http://www.lattice.cnrs.fr/democrat.
[2] TXM : http://textometrie.ens-lyon.fr.

Marques configurationnelles, chaînes de référence et thèmes de discours

Dans ses travaux fondateurs sur les plans d'organisation du texte, Charolles (1988 et 1995) définit un certain nombre d'outils relationnels de nature sémantico-pragmatique qui créent des connexions et participent à la construction de la cohérence textuelle. Ces éléments sont de différents types et agissent à différents niveaux du texte. On distingue ainsi les connecteurs, les anaphores et chaînes de référence, les expressions introductrices de cadres de discours et les marques configurationnelles, comme le paragraphe, « qui délimitent au sein du continuum textuel des ensembles présentés par le locuteur comme constituant une ou plusieurs unités en regard d'un certain critère dispositionnel » (Charolles 195 : 128). Parmi tous ces éléments, les CR semblent jouer un rôle de premier plan dans la mesure où elles sont intimement liées au thème (ou topique) de discours.

Bien qu'elles paraissent de prime abord jouer au niveau (typo)graphique, les marques configurationnelles sont en réalité elles aussi en relation étroite avec le thème et la progression thématique. Les nombreuses études portant sur le paragraphe (notamment Longacre 1979, Mitterand 1985, Bessonnat 1988, Adam 2018), qu'elles reposent sur une approche strictement textuelle ou plus cognitive, insistent en général sur l'adéquation entre unité paragraphique et unité thématique.

Il semble donc pertinent de s'interroger sur la façon dont les CR et paragraphes interagissent pour déterminer l'organisation thématique du texte. L'un des intérêts de cette recherche est de mettre en regard des marques opérant sur un plan micro-textuel, les chaînes de référence, avec d'autres marques caractéristiques du niveau méso-textuel, les paragraphes (Adam 2018). On complètera l'analyse en prenant en compte un troisième organisateur textuel, la division en chapitres, qui fait également partie des marques configurationnelles mais se situe au niveau de la macro structure. L'autre apport des chapitres pour notre étude sera qu'ils sont introduits par des titres, dont la fonction est également thématique.

L'étude des CR dans chaque palier ou unité structurelle des *Essais sur la peinture* de Diderot (texte au *niveau macro* > chapitre au *niveau macro* > paragraphe au *niveau méso*) permettra de répondre à deux objectifs principaux : (i) mesurer le degré d'homogénéité du texte à travers l'homogénéité des CR de ses unités structurelles (section 3), (ii) vérifier dans quelle mesure les CR et leurs propriétés permettent d'identifier les thèmes de ces unités (section 4). On s'appuiera pour cela sur l'annotation des chaînes de référence d'une partie des *Essais* (environ 9000 mots). L'annotation et les outils d'analyses mis en œuvre pour cette recherche ont été produits par le projet ANR DEMOCRAT[3] et sont librement accessibles[4].

[3] DEMOCRAT – ANR-15-CE38-0008.

[4] Le corpus et les outils DEMOCRAT sont entreposés sur la plateforme Ortolang : https://hdl.handle.net/11403/democrat. L'outil d'annotation et d'exploitation est accessible sous la forme d'un module d'extension appelé « annotation URS (Unité-Relation-Schéma) » (Heiden 2019) du logiciel TXM (Heiden *et al.* 2010).

2. Méthodologie de linguistique de corpus

2.1. Choix du corpus

Le texte choisi pour cette étude, les *Essais sur la peinture* de Diderot, a été composé pour l'essentiel en 1766. Il est adressé à Melchior Grimm, qui le publie dans la *Correspondance littéraire* (revue manuscrite à diffusion très restreinte) en août, novembre et décembre 1766. C'est au même destinataire que sont adressés les *Salons* de 1759, 1761, 1763, 1765 et 1766 et tous ces textes paraissent dans la même revue.

Les *Essais* se présentent comme une sorte de complément au *Salon de 1765*. Diderot annonce à la fin du texte « un petit traité de peinture » qui doit « exposer franchement les motifs de confiance qu'on peut avoir dans nos jugements ». Ce ne sera finalement pas un traité mais des essais qui paraîtront peu après. On verra que le titre et le genre de l'œuvre peuvent faire débat mais le lien avec les *Salons*, sur le fond comme sur la forme, est évident.

Le texte des *Essais* comporte au départ cinq chapitres, qui ne sont pas numérotés mais qui sont précédés d'un titre. S'y ajoutent rapidement deux chapitres portant sur l'architecture puis un supplément au troisième chapitre, intitulé *Examen du clair-obscur*. Ce supplément a sans doute été rédigé par Diderot entre 1766 et 1773. Il est absent de la première édition, qui paraît en 1795, mais inclus dans la seconde, publiée en 1798. Les quatre chapitres annotés dans le cadre du projet Democrat sont les trois premiers chapitres initiaux et cet ajout sur le clair-obscur.

La version du texte exploitée dans cette étude a été éditée par G. May. Elle est tirée de l'édition de référence des œuvres complètes de Diderot, dite « édition Dieckmann-Varloot », et se fonde sur les meilleures copies des *Essais*, principalement les copies de Stockholm. L'éditrice indique que les graphies des noms propres et la ponctuation du texte original ont été respectées tandis que l'orthographe a été modernisée. Les limites des structures textuelles (chapitres et paragraphes) varient très peu entre cette version et la première édition des *Essais*. Seul le paragraphe 14 de l'édition moderne regroupe deux paragraphes du texte imprimé en 1795. Puisque les structures textuelles qui nous intéressent sont très stables entre ces deux éditions, nous pouvons nous appuyer sur l'édition G. May pour notre étude. Nous écartons cependant le supplément sur le clair-obscur, parce que c'est un ajout postérieur et parce que ce passage n'a pas pu être comparé avec l'édition initiale de 1795.

Du point de vue générique, le texte des *Essais* occupe une place très singulière. On sait que Diderot renouvelle le genre naissant de la critique d'art grâce à la « polygénéricité » de ses textes et à ce que l'on a parfois appelé « l'absence d'œuvre » (Benrekassa 1992). Textes de commande de son ami Melchior Grimm, les *Salons* prennent dès le départ la forme de lettres adressées à cet ami. C'est donc sur le mode de la conversation épistolaire que sont construits à la fois les *Salons* et les *Essais*. Et la revue qui les accueille, la *Correspondance littéraire*, a pour vocation de rendre compte de l'actualité intellectuelle et artistique parisienne à un public très ciblé de princes européens (Fernandes 2014).

L'influence du genre épistolaire explique quelques-uns des traits saillants des *Essais* : fréquence des adresses au destinataire[5], des interrogations, des verbes à l'impératif et au présent de l'indicatif, et plus généralement de tout ce qui donne l'illusion de la communication directe et de l'échange privé. Le registre familier, le ton très libre et parfois décousu qui caractérisent le texte vont dans le même sens.

L'autre trait distinctif des *Essais* est qu'ils mettent en scène une forme d'échange entre l'auteur et le ou les destinataire(s). De même qu'il lui arrive parfois de faire parler les tableaux dans les *Salons*, Diderot tend à mettre en scène une situation de débat où des avis contradictoires s'opposent et se répondent les uns aux autres. C'est bien entendu son avis et son sentiment personnel qui priment et la première personne est omniprésente dans tout le texte. La visée argumentative de l'œuvre, qui doit convaincre le lecteur du bien-fondé de ses jugements sur la peinture, est supportée par cette représentation de la parole de l'autre. Le texte prolonge ainsi à sa façon la longue tradition du dialogue didactique.

Enfin, sur le fond comme sur la forme, le but affiché de Diderot semble être de déconstruire les canons littéraires et l'académisme sclérosé. Cette déconstruction passe par exemple par la forme des titres de ses chapitres, qui cadre mal avec l'image habituelle du traité ou de l'essai : *Mes pensées bizarres sur le dessin, mes petites idées sur la couleur, tout ce que j'ai compris de ma vie du clair-obscur*, etc. De même qu'il valorise la « peinture de genre », qui délaisse les grandes figures antiques et historiques au profit des thèmes de la vie quotidienne, Diderot s'écarte du grand style. En multipliant digressions et écarts, il brise la séquentialité du texte, ce qui autorise le rapprochement avec l'art pictural : « on pourrait évoquer le genre du *coq-à-l'âne* pour le passage sans souci de liaison ou de cohérence d'un sujet à l'autre, d'un registre à l'autre, un peu, somme toute, sur le modèle plan et simultané du tableau ». (Vasak 2007 : 23). Tous ces éléments font des *Essais* un texte difficile à classer, au genre mal défini et à l'allure polymorphe. L'étude des chaînes de référence, dans ses rapports à la structure et à la linéarité du texte, nous permettra de voir en quoi cette impression d'ensemble est confirmée.

2.2. Encodage numérique du corpus

Le fichier numérique du texte provient de l'UMR ATILF sous le code R029. Les structures textuelles – chapitres et paragraphes – ont été encodées en XML selon les recommandations de la TEI[6]. Le corpus TXM réalisé pour l'étude est formé des 3 premiers chapitres et est appelé « DIDEROTESSAIS », l'ensemble du corpus d'étude étant désigné par « bloc » dans l'article.

2.3. Annotation du corpus

Le corpus a d'abord été annoté automatiquement au moment de son import dans le logiciel TXM. Il s'agit d'un étiquetage des mots en morphosyntaxe et en lemmes à l'aide du logiciel TreeTagger[7] pour le français moderne.

Le corpus a ensuite fait l'objet d'une annotation manuelle en CR dans le cadre du projet DEMOCRAT[8]. Cette annotation se base sur la définition traditionnelle de la

[5] Il peut s'agir de Grimm ou du lecteur générique.
[6] Text Encoding Initiative, http://www.tei-c.org.
[7] https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/

chaîne : font partie d'une chaîne de référence toutes les expressions coréférentes d'un texte. Les limites d'unités structurelles n'entraînent pas de rupture de chaîne, ce qui nous permettra d'étudier de manière à la fois distincte et corrélée la structure textuelle et l'évolution des chaînes au fil du texte.

Les maillons des CR sont annotés comme *mentions*. Les mentions sont chacune associées à un *référent* et regroupées dans une même CR lorsqu'elles ont le même référent. La seule restriction à la constitution d'une chaîne est que le nombre de mentions coréférentes doit être supérieur à deux. En deçà de ce seuil, on distingue les *singletons* (mentions isolées) et les *paires* (successions de deux mentions, que les « notions d'anaphore et de coréférence suffisent amplement à décrire », Schnedecker et Landragin 2014 : 4). Les singletons et les paires seront parfois pris en compte dans nos analyses, mais uniquement par comparaison avec les chaînes proprement dites.

Une annotation complémentaire des mentions est ensuite réalisée en mode semi-automatique, en deux passes. Il s'agit d'une annotation en catégorie des mentions selon qu'elles réfèrent par pronoms personnels, déterminants possessifs, syntagmes nominaux définis, indéfinis, etc. Une macro attribue automatiquement une catégorie aux mentions (sur la base des étiquettes morphosyntaxiques des mots qui la composent), qu'un annotateur humain vérifie dans un second temps. Les annotations en référence et en catégorie sont ensuite vérifiées par un second annotateur[9].

Lorsque l'annotation en mentions est terminée, on réalise l'annotation en CR en rassemblant automatiquement dans un même ensemble (un schéma dans le modèle URS, voir plus loin) les mentions qui partagent la même propriété référent. Le corpus TXM annoté résultant fait partie du corpus DEMOCRAT sous le nom de fichier « DIDEROTESSAIS.txm ».

Une annotation complémentaire est ensuite réalisée pour les besoins de la présente étude par le calcul de la propriété « accessibilité » des mentions et des propriétés « cibles » et « empan » des CR.

2.4. L'extension « annotation URS » de TXM : un outil pour annoter, vérifier et exploiter

L'extension « Annotation URS » de TXM ajoute des fonctionnalités et des interfaces pour l'annotation dynamique de modèles d'annotation URS (Unité-Relation-Schéma) pour tous les corpus gérés par TXM. Le modèle URS a été défini et implémenté à l'origine dans le logiciel Glozz (Widlöcher *et al*. 2009). Il a également été implémenté dans le logiciel Analec (Landragin *et al*. 2012).

L'extension TXM reprend l'implémentation du logiciel Analec en la rendant compatible avec l'environnement de la plateforme TXM (architecture des corpus textuels, outils d'exploitation, interface utilisateur intégrée, outils d'import et d'export de textes, d'annotations et de résultats) tout en développant de nouvelles fonctionnalités basées sur ce modèle, comme par exemple l'interrogation croisée des

[8] Les principes d'annotation sont précisés dans le manuel d'annotation DEMOCRAT : http://www.lattice.cnrs.fr/democrat/files/ANR-15-CE38-0008-DEMOCRAT_livrable_methodo.pdf.

[9] Tout le travail d'annotation et de vérification a été réalisé à l'aide des nouvelles interfaces d'annotation interactive offertes par l'extension « annotation URS » de TXM développée dans le cadre du projet ANR DEMOCRAT.

structures textuelles (issues de l'encodage TEI à l'import) et des annotations URS des chaînes de référence pour les extractions et les décomptes de cette étude.

L'extension permet d'annoter interactivement les unités au sein des éditions de texte de TXM, d'enrichir les annotations d'unités, de schémas et de relations par commandes ou par macros, de vérifier leur cohérence et de procéder à diverses extractions pour affichage ou décomptes.

3. Chaînes de référence et homogénéité des unités structurelles

La double annotation des *Essais* en CR et en unités structurelles permet de se faire une idée assez précise des caractéristiques des CR de chaque unité. Nos observations porteront avant tout sur les chapitres et quelques sondages réalisés sur les paragraphes seront évoqués en fin de section.

Les trois chapitres étudiés sont de taille comparable (autour de 3000 tokens, ponctuation comprise). Les mesures qui suivent sont réalisées sur chacun d'entre eux et les CR se définissent relativement à cette unité : une CR du chapitre 1 est une chaîne qui a au moins trois mentions dans le chapitre 1, indépendamment du fait qu'elle soit présente ou non dans une autre partie du texte.

Les premières mesures ne détaillent pas les chaînes mais les analysent en bloc à l'intérieur de chaque chapitre. Les mesures suivantes permettront de distinguer des types de chaînes dans les chapitres.

3.1. Caractéristiques générales des chaînes des chapitres

On s'intéresse tout d'abord à la *densité référentielle* ou proportion d'expressions référentielles relativement à tous les mots du chapitre, que ces expressions soient ou non intégrées à des chaînes (Tableau 1).

Unité	Densité référentielle
Chapitre 1	27,04%
Chapitre 2	29,66%
Chapitre 3	29,39%
Bloc	28,64%

Tableau 1 : Densité référentielle des chapitres

Les chiffres sont proches pour les trois chapitres et pour le bloc dans son ensemble. Où qu'on se situe dans le texte, un peu moins d'un tiers des mots sont des expressions référentielles. Cette donnée peut être complétée par le calcul de la *couverture du texte par les chaînes*. Pour évaluer cette couverture, on commence par établir le rapport entre le nombre de mots des chaînes et le nombre de mots du texte (Tableau 2).

Unité	Densité des chaînes (en nombre de mots)
Chapitre 1	12,66%
Chapitre 2	12,89%
Chapitre 3	9,57%
Bloc	12,62%

Tableau 2 : Densité en chaînes des chapitres (en nombre de mots)

On compare ensuite le nombre de maillons des chaînes au nombre d'expressions référentielles (ou mentions) des chapitres (Tableau 3).

Unité	Maillons de chaînes	Mentions	Densité
Chapitre 1	381	814	46,8%
Chapitre 2	379	872	43,5%
Chapitre 3	270	829	32,6%
Bloc[10]	1108	2515	44%

Tableau 3 : Densité en chaînes des chapitres (en nombre de mentions)

Les tableaux 2 et 3 signalent tous deux que les CR occupent une part plus faible du chapitre 3. Le tableau suivant, qui détaille le nombre de singletons, de paires et de CR, permet de comprendre que cette singularité tient au fait que le nombre de CR est moins élevé dans ce chapitre. Le déficit en CR y est compensé par un nombre plus important de singletons et de paires. Il y a donc au total beaucoup de référents dans le chapitre 3 (522 au total), mais un grand nombre d'entre eux disparaît très vite, avant même de former une chaîne. Il peut arriver qu'un référent ne soit pas mentionné plus de deux fois dans le chapitre 3 mais qu'il l'ait été suffisamment dans les chapitres précédents pour construire une CR à l'échelle du bloc. Ces cas sont rares (voir note 11) et ne changent pas la situation d'ensemble.

Unités	Singletons	Paires	CR	Total
Chapitre 1	325	54	56	435
Chapitre 2	355	69	51	475
Chapitre 3	389	85	48	522
Bloc	1069	208	155	1432

Tableau 4 : Nombre de singletons, de paires et de CR des chapitres[11]

La section suivante apportera une indication supplémentaire : les chaînes du chapitre 3 sont moins longues que celles des deux autres chapitres.

3.2. Types de chaînes des chapitres

La mesure de la *longueur des chaînes en nombre de mentions* permet d'établir une première typologie. On peut ainsi comparer les chaînes courtes (3 et 4 mentions), intermédiaires (entre 5 et 10 mentions) et longues (plus de 10 mentions) à l'intérieur des chapitres.

[10] La somme des mentions des CR de chaque chapitre n'est pas strictement égale au nombre de mentions des CR du bloc dans la mesure où 78 CR ont moins de 3 mentions dans un chapitre tout en ayant au moins 3 mentions dans le bloc.

[11] Il faut signaler en outre 5 CR qui ont moins de 3 mentions dans chaque chapitre mais qui en ont plus de 3 dans le bloc et 15 paires dont les 2 mentions ne sont pas dans le même chapitre.

	CR courtes			CR intermédiaires	CR longues	
Unité	3 maillons	4 maillons	Total	5 à 10 maillons	plus de 10 maillons	Total
Chapitre 1	22	9	31	20	5	56
Chapitre 2	24	7	31	10	10	51
Chapitre 3	21	10	31	12	5	48
Bloc	67	26	93	42	20	155

Tableau 5 : Longueur des CR des chapitres

Les chaînes courtes sont en nombre très constant. Le chapitre 1 est plus fourni en CR intermédiaires, le chapitre 2 en CR longues. Le chapitre 3 se distingue en ce qu'il n'a ni beaucoup de CR intermédiaires, ni beaucoup de CR longues. Il présente donc peu de CR au total, beaucoup de singletons et de paires et des CR moins longues. L'examen de la taille de la CR la plus longue des chapitres confirme cette dernière particularité :

Unité	Chaîne la plus longue
Chapitre 1	42 mentions
Chapitre 2	50 mentions
Chapitre 3	29 mentions

Tableau 6 : Taille de la CR la plus longue des chapitres (en mentions)

Le deuxième critère qui permet de définir une typologie des chaînes est l'*empan des chaînes*. On dénombre trois types de chaînes : les CR qui ne dépassent pas les limites d'un paragraphe, celles qui se limitent au chapitre et celles qui vont au-delà :

Unité	CR de paragraphe	CR de chapitre	CR de bloc	Total
Chapitre 1	28	21	7	56
Chapitre 2	28	4	19	51
Chapitre 3	28	10	10	48

Tableau 7 : Empan des CR des chapitres

Les trois types de CR du tableau 7 ne se recouvrent jamais. Dans le chapitre 1 par exemple, on dénombre 28 CR dont toutes les mentions sont dans un seul paragraphe, 21 CR qui courent toujours sur plus d'un paragraphe mais sans dépasser les limites du chapitre, et seulement 7 CR qui sont instanciées au moins une fois dans un autre chapitre du bloc.

Le nombre de CR qui se limitent à un seul paragraphe reste remarquablement stable quel que soit le chapitre et ce nombre est toujours plus élevé que les autres. Pour le reste, chaque chapitre a un profil particulier. Les CR du premier chapitre restent pour la plupart cantonnées dans cette unité. La situation est inversée dans le chapitre deux et les deux types de CR sont parfaitement équilibrés dans le chapitre 3. La fréquence des CR dépassant le chapitre 2 n'est pas très surprenante, dans la mesure où les titres des chapitres 2 et 3 (*mes petites idées sur la couleur* et *tout ce que j'ai compris de ma vie du clair-obscur*) laissent penser que leurs thématiques sont liées. On pouvait donc s'attendre à ce qu'un certain nombre de référents présents dans le chapitre 2 soient repris dans le suivant. La position intermédiaire de ce chapitre peut également introduire un biais dans l'analyse, puisqu'il a plus de

chances d'avoir des chaînes qui se prolongent soit dans le chapitre qui précède soit dans celui qui suit.

3.3. Bilan sur l'homogénéité des unités structurelles

L'analyse comparée des CR des trois chapitres des *Essais* révèle une situation contrastée. Si la densité référentielle est assez stable au fil du texte, la couverture des chapitres par les chaînes et les caractéristiques de ces chaînes sont plus hétérogènes.

Le chapitre 1 se distingue par le fait que la plupart de ses référents ne sont pas réinstanciés au-delà du chapitre. On note toutefois un nombre important de CR qui dépassent les limites d'un paragraphe et se poursuivent dans le chapitre, ce qui est peut-être à mettre en relation avec le nombre important de CR de longueur intermédiaire (entre 5 et 10 mentions).

Le chapitre 2 comporte beaucoup de CR qui sont présentes ailleurs dans le bloc. C'est également dans ce chapitre qu'on rencontre le nombre le plus important de CR longues (dans la limite du chapitre).

Le chapitre 3 se distingue par la rapidité avec laquelle ses référents disparaissent : il comporte moins de CR, des CR qui sont globalement moins longues et qui occupent une moindre part du texte.

Les trois chapitres partagent une même particularité : la moitié au moins de leurs chaînes se cantonne à un seul paragraphe. Ce résultat n'a rien d'étonnant. De même que les chaînes courtes sont toujours les plus nombreuses (dans le texte, dans le chapitre, dans le paragraphe), de même, une grande partie des chaînes a un empan très local. Il y a donc dans le texte – et ce constat vaut probablement pour tous les textes – un grand nombre de référents éphémères dont les mentions sont peu nombreuses et très rapprochées (dans le même paragraphe).

Cinq paragraphes de taille comparable (autour de 300 mots) extraits des trois livres ont été analysés les uns indépendamment des autres selon le même protocole. Les résultats de cette analyse révèlent de grandes disparités. La densité référentielle, la couverture et la densité en chaînes, la longueur et l'empan des chaînes varient dans des proportions importantes, ce qui montre qu'à l'intérieur d'un texte comme les *Essais* l'étude des CR peut donner des résultats assez différents selon les paragraphes qu'on examine. Notre étude permet en revanche de repérer deux constantes : le nombre de CR varie peu dans des unités de taille comparable (4 ou 6 chaînes pour 300 mots) et lorsque ce nombre est plus élevé (6 chaînes) ce sont en réalité les chaînes courtes qui augmentent. Ce constat renforce le poids des CR courtes dans la cartographie générale.

L'étude confirme ainsi le rôle micro-textuel des CR. La plupart des chaînes sont très courtes, plus de la moitié d'entre elles ne dépassent pas le niveau méso-textuel du paragraphe et leur nombre dans le paragraphe semble très dépendant de sa taille. Le fait que l'auteur des *Essais* organise ses unités de paragraphe autour d'un petit nombre de CR pose par ailleurs la question de leur relation à l'unité thématique. Nous tenterons de préciser ce lien dans la section qui suit.

4. Chaînes de référence et thèmes des unités structurelles

Le thème est entendu ici au sens « classique » de *thème de discours* (Marandin 1988 : 68) et défini grâce à la relation d'« à propos » : c'est ce dont le discours « parle », ce à propos de quoi il dit quelque chose. Nous sommes conscients qu'il s'agit là d'une définition parmi d'autres, qui tend à calquer de manière un peu simpliste la notion de thème de discours sur celle de thème de phrase en supposant que le thème est porté par un constituant de l'énoncé. Mais parce qu'elle établit un lien direct entre le thème et les référents mentionnés à l'intérieur des chaînes, cette définition a le mérite d'offrir un point de départ qui paraît opératoire (on verra que les moyens mis en œuvre pour l'étude deviennent vite complexes, en particulier lorsqu'il s'agit de traiter une grande masse de données).

Pour réaliser cette étude, nous nous inspirons du modèle de Givón (1983) et de la manière dont il articule les thèmes et les référents discursifs (qu'il appelle topiques) en les situant sur des niveaux d'organisation différents. Givón distingue en réalité non pas deux mais trois niveaux qui s'emboitent les uns dans les autres et qui s'organisent autour de trois types de continuité/discontinuité discursive : la continuité thématique (*thematic continuity*) > la continuité d'action (*action continuity*) > la continuité topicale (*topics/participants continuity*). L'intrication de ces trois niveaux explique qu'il y ait une corrélation assez forte entre la continuité thématique et la continuité topicale : ce dont traite une unité thématique, que Givón appelle « paragraphe thématique » (indépendamment du découpage graphique), ce sont généralement un ou plusieurs référents (ou topiques) proéminents. Ces référents les plus centraux sont le plus souvent les topiques les plus continus de cette unité. Il y a donc *a priori* une association étroite entre les référents les plus continus, les plus centraux/proéminents et ceux qui définissent le thème du discours :

> « Within the thematic paragraph it is most common for one topic to be the continuity marker, the *leitmotif*, so that it is the participant *most crucially involved* in the action sequence running through the paragraph ; **it is the participant most closely associated with the higher-level 'theme' of the paragraph**[12]; and finally, it is the participant most likely to be coded as the *primary topic* – or *grammatical subject* – of the vast majority of sequentially-ordered clauses/sentences comprising the thematic paragraph. It is thus, obviously, the most *continuous* of all the topics mentioned in the various clauses in the paragraph » (Givón 1983 : 8).

La continuité topicale pourrait ainsi permettre d'approcher ce qui fait l'unité d'un paragraphe thématique et donner un moyen d'identifier son thème. Elle se mesure notamment à travers la persistance du référent : on s'attend en effet à ce qu'un référent continu soit réinstancié un grand nombre de fois et soit récurrent dans l'unité thématique. La longueur des chaînes de référence sera pour nous une manière de mesurer la persistance des référents à l'intérieur des unités structurelles. La fréquence des catégories grammaticales permettra de compléter l'analyse croisée de la continuité topicale et thématique, certaines catégories (les marques de haute

[12] C'est nous qui soulignons.

accessibilité) marquant *a priori* la continuité maximale, d'autres au contraire une continuité plus faible (les marques d'accessibilité moyenne ou faible).

4.1. Application du cadre d'analyse au corpus

Dans cette section, nous envisageons les unités structurelles (bloc, chapitres et paragraphes) comme des unités thématiques. En ce sens, nous faisons l'hypothèse que l'auteur développe dans chaque unité un thème qui lui est propre et que ce thème est sensiblement différent de celui de l'unité qui précède et qui suit. Nous n'irons pas jusqu'à prétendre que chaque paragraphe procède d'un saut thématique radical mais qu'il existe d'un paragraphe à l'autre une évolution sensible. Puisque selon la même hypothèse un chapitre est aussi en soi une unité thématique, les différents paragraphes qui le composent devront à la fois participer d'un thème commun (celui du chapitre) tout en ayant une déclinaison qui leur est propre au niveau du paragraphe. Dans ce cadre, nous cherchons à savoir de manière empirique si les CR qui jalonnent ces unités structurelles sont à même d'aider à identifier, par le biais de leurs propriétés, le thème de ces unités. Nous étudions pour cela trois propriétés en particulier : (1) *la longueur des chaînes* (on suppose que le thème d'une unité est porté par les référents les plus souvent mentionnés, les plus persistants), (2) *l'empan des chaînes* (le thème serait porté par les référents qu'on ne mentionne pas ailleurs dans une autre unité) et (3) *l'accessibilité des référents* (le thème serait porté par les référents le plus accessibles, les plus présents en mémoire).

4.2. Méthode : définition des thèmes et des chaînes cibles

Le bloc que nous étudions est composé de 3 chapitres eux-mêmes divisés en 71 paragraphes au total. Si les chapitres sont de taille comparable, les paragraphes sont au contraire de taille très variable. Dans cette section, nous nous focaliserons sur les 3 paragraphes les plus longs de chaque partie :

Chapitre	1			2			3		
Paragraphe	14	17	22	29	44	46	53	63	66
Taille (mots)	529	268	298	287	394	267	308	287	296

Tableau 8 : Taille des paragraphes retenus pour l'étude « thématique »

Du point de vue des thèmes des unités structurelles, la situation est à nouveau inégale. Les chapitres ont un nom donné par l'auteur, que l'on peut prendre pour thème, tandis que les paragraphes n'en ont aucun. Les titres de chapitre sont les suivants :

Chapitre 1 : « Mes pensées bizarres sur le dessin »
Chapitre 2 : « Mes petites idées sur la couleur »
Chapitre 3 : « Tout ce que j'ai compris de ma vie sur le clair-obscur »

Pour établir un lien entre les chaînes et le thème des paragraphes, nous recourons à une méthode manuelle qui consiste à annoter chacun d'eux pour lui attribuer une sorte de titre et déterminer ainsi ses éléments thématiques. Cette méthode implique une certaine subjectivité du lecteur, que nous allons modérer en procédant à une double annotation. Chaque annotateur lit et propose un thème pour chaque

paragraphe. Les deux annotateurs confrontent ensuite leurs propositions et s'accordent sur une description commune. En procédant ainsi, nous disposons de descriptions thématiques pour les deux niveaux de structure, le chapitre et le paragraphe. Nous concédons que ces descriptions ne sont pas de la même nature ou du moins de la même origine, puisque les premières sont fournies par l'auteur dans un certain dessein tandis que les autres sont construites par un binôme de lecteurs. Il s'agit d'une première approche d'un problème difficile (on a mentionné plus haut le flou conceptuel qui entoure en général la notion de thème de discours, cf. Marandin, 1988).

Après annotation des unités structurelles, le tableau des thèmes est le suivant (les guillemets signalent qu'il s'agit du titre original) :

Unité	**Thème ou titre**
Chapitre 1	« Mes pensées bizarres sur le dessin »
Paragraphe 14	Le dessin académique d'après modèle empêche de peindre la nature en vérité.
Paragraphe 17	Nécessité de peindre une action véritable plutôt qu'un modèle cherchant à imiter
Paragraphe 22	L'école de dessin idéale
Chapitre 2	« Mes petites idées sur la couleur »
Paragraphe 29	La variété des couleurs est liée à celle des états d'âme du peintre
Paragraphe 44	La palette des couleurs d'un artiste est plus limitée que celle de la nature
Paragraphe 46	Les couleurs de la chair sont très fugaces et difficiles à fixer
Chapitre 3	« Tout ce que j'ai compris de ma vie sur le clair-obscur »
Paragraphe 53	Par leur grande beauté, certains paysages d'ombre et de lumière nous paraissent tels des œuvres d'art
Paragraphe 63	L'éclat de la lumière influence la couleur des objets et celle de leur ombre
Paragraphe 66	Comme la lumière influe sur la couleur, elle joue aussi sur la teinte des ombres

Tableau 9 : Les thèmes des unités structurelles

La phase ultime de la préparation des données consiste à relever parmi toutes les CR celles dont les référents sont sémantiquement proches du thème annoté dans l'unité. Pour ce faire, nous considérons chaque unité comme un sous-corpus et dressons la liste des chaînes candidates (il faut notamment qu'elles aient au moins trois mentions à l'intérieur de l'unité). Ces chaînes candidates à exprimer le thème de l'unité sont appelées chaînes cibles. On ne tient aucun compte dans cette phase des propriétés des chaînes et on se base uniquement sur des rapports sémantiques.

Dans un second temps, nous étudierons les propriétés de longueur et d'empan des chaînes cibles ainsi que l'accessibilité de leurs référents (mesurée grâce aux catégories grammaticales des mentions), en les comparant aux autres chaînes des unités.

Pour le premier chapitre, nous retiendrons les CR de l'auteur, ses pensées ou réflexions et le dessin pris au sens large (peinture, tableau, œuvre, esquisse, croquis, etc.). Parmi les 56 chaînes qui figurent dans le chapitre, nous ne retiendrons donc que 3 chaînes cibles : l'auteur (42 mentions), le dessin (4 mentions) et l'écorché (le dessin anatomique, 8 mentions).

Pour le paragraphe 14 et ses 11 chaînes, nous retiendrons « l'Académie » (7 mentions) ainsi que « les actions, les positions et les figures fausses, apprêtées et froides » (7 mentions) pour exprimer la notion de vérité.

Le paragraphe 17 ne comporte que 3 chaînes. Les trois référents que l'auteur mentionne à plus de trois reprises dans le paragraphe sont : « l'auteur », « les deux camarades » et « les jeunes élèves ». Aucun de ces référents n'exprime même en partie le thème annoncé. Ce paragraphe sera pour cette raison exclu de la suite des analyses.

Le reste des unités a été traité selon le même principe. La liste des chaînes cibles est donnée dans le tableau suivant (le nombre entre parenthèses indique la longueur c'est-à-dire le nombre de fois que le référent est mentionné dans l'unité considérée).

Unité	**Thème ou titre**	**Chaînes cibles (référents)**
Chapitre 1	« Mes pensées bizarres sur le dessin »	l'auteur (40), l'écorché (8), le dessin (4)
Paragraphe 14	Le dessin académique d'après modèle empêche de peindre la nature en vérité.	l'Académie (7), les figures fausses (7)
Paragraphe 17	Nécessité de peindre une action véritable plutôt qu'un modèle cherchant à imiter	(aucun)
Paragraphe 22	L'école de dessin idéale	l'élève de l'école de dessin (10), le modèle académique (3)
Chapitre 2	« Mes petites idées sur la couleur »	l'auteur (34), la couleur (26), les couleurs (7)
Paragraphe 29	La variété des couleurs est liée à celle des états d'âme du peintre	le même voile jaune (3), l'homme qui peint (29)
Paragraphe 44	La palette des couleurs d'un artiste est plus limitée que celle de la nature	les couleurs (6), l'arc-en-ciel (3)
Paragraphe 46	Les couleurs de la chair sont très fugaces et difficiles à fixer	le coloriste vrai (5), la chair (3), ce qui achève de rendre fou le grand coloriste (4)
Chapitre 3	« Tout ce que j'ai compris de ma vie sur le clair-obscur »	l'auteur (29), le clair-obscur (3), la lumière (25), les ombres (12), la lumière générale (3), les lumières (7)
Paragraphe 53	Par leur grande beauté, certains paysages d'ombre et de lumière nous paraissent tels des œuvres d'art	les vieux arbres épargnés (3), le soleil (4), les rayons obliques du soleil (6)
Paragraphe 63	L'éclat de la lumière influence la couleur des objets et celle de leur ombre	la lumière (4), la lumière générale (3)
Paragraphe 66	Comme la lumière influe sur la couleur, elle joue aussi sur la teinte des ombres	la lumière (4), les ombres (5)

Tableau 10 : Liste des chaînes cibles par unité structurelle

On notera que chaque chapitre est relativement bien thématisé et que le thème ressort partiellement à partir des référents des chaînes : le premier porte plutôt sur le dessin académique (et pourrait sans doute avoir un autre titre que celui donné par l'auteur), le deuxième sur la couleur et le troisième sur la lumière. Chaque paragraphe reprend en général une partie du thème du chapitre et développe une ou

deux chaînes particulières. Mais le lien qui existe entre les chaînes cibles et le thème est plus ou moins direct selon les paragraphes (voir par exemple les paragraphes 29, 44 et 53). En effet, le thème est parfois davantage porté par des exemples variés et spécifiques (« les vieux arbres épargnés » pour « les paysages d'ombre ») et par ailleurs par des référents qui ne sont pas suffisamment répétés pour constituer des chaînes[13].

Nous étudions à présent si les propriétés énoncées plus haut – longueur, empan des chaînes et accessibilité des référents – permettent ou non d'identifier le thème des unités structurelles.

4.3. Longueur des chaînes

L'hypothèse que nous testons ici est que l'expression du thème se fait par la multiplication des mentions de certains référents. Selon cette hypothèse, le thème serait porté par les chaînes les plus longues de l'unité. D'après les analyses faites dans la section précédente, nous considérons les chaînes longues de chapitre à partir de 10 maillons et seulement à partir de 5 pour les chaînes de paragraphe.

À titre d'exemple, nous reproduisons le paragraphe 22 en mettant en évidence en italique les chaînes longues (mentionnés au moins 5 fois). Les mentions en gras appartiennent aux chaînes cibles, c'est-à-dire celles qui qui constituent d'après notre méthode exploratoire de bons candidats à la définition du thème. Lorsque les mentions sont à la fois en gras et en italique, c'est qu'elles appartiennent à une chaîne cible qui a aussi la propriété d'être longue.

> Voici donc comment [*je*]$_{a}$ désirerais qu'une école de dessin fût conduite. Lorsque [***l'élève***]$_{e}$ sait dessiner facilement d'après l'estampe et la bosse, [*je*]$_{a}$ [***le***]$_{e}$ tiens pendant deux ans devant [**le modèle académique de l'homme et de la femme**]$_{m}$. Puis [*je*]$_{a}$ [***lui***]$_{e}$ expose des enfants, des adultes, des hommes faits, des vieillards, [*des sujets*]$_{s}$ de tout âge, de tout sexe, pris dans toutes les conditions de la société, toutes sortes de natures, en un mot. [*Les sujets*]$_{s}$ se présenteront en foule à la porte de [*mon*]$_{a}$ académie, si [*je*]$_{a}$ [*les*]$_{s}$ paie bien ; si [*je*]$_{a}$ suis dans un pays d'esclaves, [*je*]$_{a}$ [*les*]$_{s}$ y ferai venir. Dans [*ces différents modèles*]$_{s}$ le professeur aura soin de [***lui***]$_{e}$ faire remarquer les accidents que les fonctions journalières, la manière de vivre, la condition et l'âge ont introduits dans les formes. [[*Mon*]$_{a}$ ***élève***]$_{e}$ ne reverra plus [**le modèle académique**]$_{m}$ qu'une fois tous les quinze jours ; et le professeur abandonnera [**au modèle**]$_{m}$ le soin de se poser [**lui-même**]$_{m}$. Après la séance de dessin un habile anatomiste expliquera à [[*mon*]$_{a}$ ***élève***]$_{e}$ l'écorché, et [***lui***]$_{e}$ fera l'application de ses leçons sur le nu animé et vivant ; et [***il***]$_{e}$ ne dessinera d'après l'écorché que douze fois au plus dans une année. C'en sera assez pour qu'[***il***]$_{e}$ sente que les chairs sur les os et les chairs non appuyées ne se dessinent pas de la même manière, qu'ici le trait est rond, là comme anguleux ; et que s'[***il***]$_{e}$ néglige ces finesses, le tout aura l'air d'une vessie soufflée ou d'une balle de coton.

On note qu'il y a dans ce paragraphe deux chaînes cibles : celle de l'élève de l'école de dessin, indexée par la lettre *e*, et celle du modèle académique, indexée par *m*. L'une est longue (l'élève, 10 mentions), l'autre est courte (le modèle, 4 mentions). Il y a dans le paragraphe deux autres chaînes longues : la chaîne de l'auteur (9 mentions, indexée par *a*), qui comporte une majorité de pronoms

[13] On trouve en général 3 à 7 chaînes par paragraphe (400 mots).

déictiques dont on peut se demander s'ils forment réellement une chaîne, et la chaîne des sujets (5 mentions, indexée par *s*).

La méthode propose donc 3 chaînes longues (l'élève de l'école de dessin, l'auteur, les sujets), dont malheureusement une seule est une chaîne cible (l'élève de l'école de dessin). Cela lui vaut un score de précision de 1/3, soit 33%. Elle ne rappelle qu'une chaîne cible sur les deux à trouver (elle manque le modèle académique). Son taux de rappel est de 1/2, soit 50%. La longueur des chaînes n'est donc pas un critère suffisamment précis pour cibler le thème de ce paragraphe.

4.4. Empan des chaînes

L'hypothèse à tester est que la spécificité d'un thème local est portée principalement par des chaînes qui ne sont pas développées ailleurs dans le texte, dans une autre unité de même niveau. Nous attribuons donc une propriété particulière à chaque chaîne, son empan, qui qualifie la chaîne selon la position des mentions dans les unités. Nous aurons ainsi des chaînes de paragraphe (dont les limites ne dépassent pas le paragraphe), de chapitre ou de bloc.

Comme précédemment, nous illustrons l'application de ce critère grâce à l'exemple du paragraphe 22. Nous mettons en italique les chaînes spécifiques à ce paragraphe. Nous indiquerons toujours en gras les chaînes cibles.

> Voici donc comment je désirerais qu'une école de dessin fût conduite. Lorsque [**l'élève**]$_e$ sait dessiner facilement d'après l'estampe et la bosse, je [**le**]$_e$ tiens pendant deux ans devant [***le modèle académique de l'homme et de la femme***]$_m$. Puis je [**lui**]$_e$ expose des enfants, des adultes, des hommes faits, des vieillards, [*des sujets*]$_s$ de tout âge, de tout sexe, pris dans toutes les conditions de la société, toutes sortes de natures, en un mot. [*Les sujets*]$_s$ se présenteront en foule à la porte de mon académie, si je [*les*]$_s$ paie bien ; si je suis dans un pays d'esclaves, je [*les*]$_s$ y ferai venir. Dans [*ces différents modèles*]$_s$ le professeur aura soin de [**lui**]$_e$ faire remarquer les accidents que les fonctions journalières, la manière de vivre, la condition et l'âge ont introduits dans les formes. [**Mon élève**]$_e$ ne reverra plus [***le modèle académique***]$_m$ qu'une fois tous les quinze jours ; et le professeur abandonnera [***au modèle***]$_m$ le soin de se poser [***lui-même***]$_m$. Après la séance de dessin un habile anatomiste expliquera à [**mon élève**]$_e$ l'écorché, et [**lui**]$_e$ fera l'application de ses leçons sur le nu animé et vivant ; et [**il**]$_e$ ne dessinera d'après l'écorché que douze fois au plus dans une année. C'en sera assez pour qu'[**il**]$_e$ sente que les chairs sur les os et les chairs non appuyées ne se dessinent pas de la même manière, qu'ici le trait est rond, là comme anguleux ; et que s'[**il**]$_e$ néglige ces finesses, le tout aura l'air d'une vessie soufflée ou d'une balle de coton.

Le paragraphe comporte toujours les mêmes deux chaînes cibles : celle de l'élève de l'école de dessin et celle du modèle académique. L'une est spécifique à ce paragraphe (le modèle académique), l'autre ne l'est pas (l'élève de l'école de dessin est aussi mentionné dans d'autres paragraphes). On détecte une autre chaîne spécifique à ce paragraphe, celle des sujets. La méthode identifie donc deux chaînes spécifiques à ce paragraphe (le modèle académique et les sujets), dont une seule chaîne candidate à l'explicitation du thème. Son score de précision n'est que de 50%. On note que la chaîne cible repérée par cette méthode (le modèle académique) n'avait pas été détectée par la méthode précédente, parce que trop courte. La chaîne des sujets, non candidate d'après nous à la définition du thème du paragraphe, se

distingue pourtant par sa longueur et par le fait qu'elle est spécifique à ce paragraphe.

4.5. Accessibilité des référents des chaînes

L'hypothèse que nous souhaitons tester ici est que le thème, toujours fortement présent en mémoire, apparaît plutôt sous la forme d'une expression qui marque la haute accessibilité du référent (pronom personnel anaphorique, pronom relatif, déterminant possessif et sujet zéro dans le cas de l'ellipse du sujet). Nous repérons donc les chaînes qui ont un fort taux de mentions de ce type.

Comme précédemment, nous illustrons l'exploitation du troisième critère à l'aide du paragraphe 22. Nous mettons en italique les chaînes qui ont un fort taux d'expressions référentielles marquant la haute accessibilité (c'est-à-dire au moins 50%). Nous indiquerons toujours en gras les mentions des chaînes cibles.

> Voici donc comment [*je*]$_{a}$ désirerais qu'une école de dessin fût conduite. Lorsque [***l'élève***]$_{e}$ sait dessiner facilement d'après l'estampe et la bosse, [*je*]$_{a}$ [***le***]$_{e}$ tiens pendant deux ans devant [**le modèle académique de l'homme et de la femme**]$_{m}$. Puis [*je*]$_{a}$ [***lui***]$_{e}$ expose des enfants, des adultes, des hommes faits, des vieillards, des sujets de tout âge, de tout sexe, pris dans toutes les conditions de la société, toutes sortes de natures, en un mot. Les sujets se présenteront en foule à la porte de [*mon*]$_{a}$ académie, si [*je*]$_{a}$ les paie bien ; si [*je*]$_{a}$ suis dans un pays d'esclaves, [*je*]$_{a}$ les y ferai venir. Dans ces différents modèles le professeur aura soin de [***lui***]$_{e}$ faire remarquer les accidents que les fonctions journalières, la manière de vivre, la condition et l'âge ont introduits dans les formes. [[*Mon*]$_{a}$ ***élève***]$_{e}$ ne reverra plus [**le modèle académique**]$_{m}$ qu'une fois tous les quinze jours ; et le professeur abandonnera [**au modèle**]$_{m}$ le soin de se poser [**lui-même**]$_{m}$. Après la séance de dessin un habile anatomiste expliquera à [[*mon*]$_{a}$ ***élève***]$_{e}$ l'écorché, et [***lui***]$_{e}$ fera l'application de ses leçons sur le nu animé et vivant ; et [***il***]$_{e}$ ne dessinera d'après l'écorché que douze fois au plus dans une année. C'en sera assez pour qu'[***il***]$_{e}$ sente que les chairs sur les os et les chairs non appuyées ne se dessinent pas de la même manière, qu'ici le trait est rond, là comme anguleux ; et que s'[***il***]$_{e}$ néglige ces finesses, le tout aura l'air d'une vessie soufflée ou d'une balle de coton.

Les deux chaînes cibles sont toujours l'élève de l'école de dessin et le modèle académique. L'une contient une forte proportion de marques de haute accessibilité (l'élève de l'école de dessin ; 70%), l'autre non (33%). Une seule autre chaîne se caractérise par son taux élevé de marques de haute accessibilité : il s'agit de la chaîne de l'auteur, sur laquelle on peut émettre de nouvelles réserves étant donné ses fortes particularités (présence exclusive de pronoms personnels de 1ère personne et de possessifs). La méthode identifie donc deux chaînes (l'élève de l'école de dessin et l'auteur), dont une seule participe à la définition du thème. Son score de précision n'est que de 50%.

4.6. Synthèse des résultats sur les chapitres et les paragraphes

Après avoir exposé la mise en œuvre des trois méthodes, nous présentons ci-dessous deux tableaux de synthèse, l'un traitant de l'identification du thème des chapitres, l'autre du thème des paragraphes. S'agissant d'un problème d'identification d'une propriété, nous nous intéressons au score de précision des

méthodes, qui mesure *a posteriori* la pertinence des chaînes proposées[14]. Pour chaque unité (chapitre ou paragraphe), nous indiquons combien de chaînes cibles étaient à trouver, combien de chaînes ont été proposées par chaque méthode et combien étaient correctes. Le rapport entre ces deux derniers chiffres nous donne la précision.

Chapitre	Chaînes cibles	Longueur	Empan	Accessibilité
1	3	1/5 (20%)	1/6 (17%)	< 1/5 (20%)
2	3	2/4 (50%)	0/4 (0%)	2/3 (67%)
3	6	3/5 (60%)	4/10 (40%)	< 2/11 (18%)
Moyenne	**4**	**43%**	**19%**	**< 35%**

Tableau 11 : Précision des trois méthodes quant à l'identification des thèmes des chapitres.

À la lecture du tableau 11, nous constatons qu'aucune des trois méthodes n'est en mesure d'identifier avec précision (ou fiabilité) les référents en rapport avec le thème des chapitres. Sachant qu'il y a en moyenne 4 chaînes cibles à trouver, on pourrait considérer qu'une bonne méthode obtiendrait un score moyen de 75 à 80% (3 chaînes correctes sur 4 proposées ou 4 chaînes correctes sur 5 proposées).

La méthode qui s'appuie sur l'empan des chaînes (c'est-à-dire sur la spécificité des référents à un chapitre donné) échoue complètement par rapport au thème du chapitre 2. La méthode qui s'en sort le mieux est celle qui s'appuie sur la propriété des longueurs des chaînes. En retenant les référents qui apparaissent au moins dix fois dans le chapitre courant, elle propose entre 4 à 5 référents, dont la moitié relèvent du thème du chapitre.

Paragraphe	Chaînes cibles	Longueur	Empan	Accessibilité
14	2	2/5	1/3	1/5
22	2	1/3	1/2	1/2
29	2	1/1	1/1	2/2
44	2	1/3	1/2	1/4
46	3	1/2	1/1	2/3
53	3	1/1	2/2	1/1
63	2	0/1	1/2	1/3
66	2	1/2	0/2	0/3
Moyenne	**2,25**	**51 %**	**60 %**	**49 %**

Tableau 12. Précision des trois méthodes quant à l'identification des thèmes des paragraphes.

Pour le problème de l'identification du thème des paragraphes (tableau 12), il faut identifier en moyenne 2,25 chaînes cibles dans chaque paragraphe. Pour obtenir un bon score, il ne faudrait faire qu'une erreur sur trois chaînes proposées, soit un score de 66%. D'après ce tableau, nous constatons que chacune des trois méthodes progresse. Elles parviennent même à atteindre le seuil des 50% avec 60% pour la méthode de l'empan. Il semblerait donc que la relation entre chaîne de référence et

[14] Les scores en rappel sont moyens et varient peu selon les méthodes : 40% sur le thème des chapitres ; 45% sur le thème des paragraphes.

thème soit plus intime au niveau des paragraphes, à une échelle plus locale, qu'au niveau plus macroscopique du chapitre.

On constate également que dans le cas des paragraphes les trois méthodes connaissent des taux de réussite relativement comparables. Chacune d'entre elles connaît une situation d'échec, mais aussi des situations de succès relatif. Contrairement à ce qui se produit avec les chapitres, les méthodes de l'empan et de l'accessibilité sont en mesure de faire des propositions pertinentes. La méthode de l'empan est d'ailleurs la plus efficace et avoisine le seuil recherché des 66%, alors qu'elle était la moins précise pour l'identification du thème des chapitres. Selon cette méthode, l'unité thématique du paragraphe serait plus particulièrement liée à la rupture qui s'établit entre cette unité et les unités voisines.

4.7. Bilan sur les thèmes des unités structurelles

Les résultats exposés ci-dessus démontre que le rapport entre les chaînes de référence et le thème des unités de discours est complexe et délicat à mettre en évidence. Chacune des trois méthodes explorées, certes simplistes (mais c'est une première approche), peine à fournir avec précision et satisfaction des référents en lien avec le thème. L'étude apporte néanmoins quelques éclairages intéressants.

L'efficacité des méthodes est variable selon qu'on traite d'une unité large comme le chapitre ou plus étroite comme le paragraphe (même si nous avons sélectionné les paragraphes les plus longs du texte). Dans le cas des chapitres, le thème semble davantage porté par le maintien répété de certains référents tout au long de l'unité, d'où un meilleur score du critère de la longueur des chaînes. Dans le cas des paragraphes, c'est plutôt le caractère discriminant, spécifique des référents qui semble jouer, comme si le thème d'un paragraphe devait se définir par opposition aux paragraphes qui l'entourent.

On notera que le critère de longueur est sans doute dépendant de la longueur du chapitre lui-même ou de sa densité référentielle. Le seuil à partir duquel on détermine qu'une chaîne est longue doit donc être apprécié relativement à d'autres propriétés de l'unité. En revanche le critère de l'empan, qui s'appuie uniquement sur la présence ou l'absence des référents dans le restant du texte, est indépendant de la taille des unités considérées.

On remarquera enfin – au moins pour le paragraphe 22 – que les chaînes proposées par les trois méthodes sont sensiblement les mêmes. Il serait sans doute intéressant d'étudier qualitativement dans quelle mesure ces méthodes se recoupent et envisager le cas échéant si une certaine combinaison de celles-ci n'apporterait pas un gain significatif.

5. Discussion et perspectives

Nous avons mené une large étude sur les propriétés des chaînes de référence en relation avec les unités structurelles imbriquées des *Essais sur la peinture* de Diderot. Nous nous sommes en particulier intéressés aux deux premiers niveaux de structure (chapitres et paragraphes) et aux propriétés de longueur, d'empan et à l'accessibilité des référents. Après un premier travail d'analyse sur la distribution des mentions dans ces structures, nous avons conduit une seconde étude sur la question du rapport entre chaînes de référence et thème des unités structurelles. Pour

ce faire, nous nous sommes appuyés tantôt sur les titres fournis par l'auteur (pour les chapitres), tantôt sur des thèmes produits par une annotation manuelle (pour les paragraphes). D'un point de vue méthodologique, il conviendrait d'homogénéiser l'approche par une annotation systématique des thèmes de toutes les unités. Cette étude multi-niveaux a néanmoins pu être réalisée grâce à l'outillage très précis offert par l'extension « annotation URS » de TXM et aux ressources produites par le projet DEMOCRAT. Les facilités apportées par cet environnement pour définir les sous-corpus par chapitre et par paragraphe, pour croiser ainsi les propriétés de chaînes et les contraintes structurelles ont été déterminantes.

Il ressort de cette recherche que les chaînes de référence sont pour l'essentiel cantonnées à une dimension très locale : une majorité d'entre elles ne dépasse pas les limites du paragraphe et toutes les unités structurelles des *Essais* comportent une forte proportion de chaînes très courtes. Pour le reste, les paragraphes se caractérisent par des variations importantes (du point de vue de la densité des chaînes, etc.) et les chapitres sont plus homogènes. Selon l'échelle à laquelle on mène l'analyse (niveau de la macrostructure ou niveau méso), l'étude des propriétés des chaînes donne ainsi des résultats bien différents, ce qui montre la difficulté à dégager des tendances stables et la complexité des rapports entre structures et chaînes.

La méthode mise en œuvre pour dégager les thèmes des unités structurelles à partir des chaînes et de leurs propriétés a montré des limites importantes, en particulier pour les unités de niveau supérieur. Notre définition du thème et le postulat qu'il est porté par les référents du texte peut naturellement être mis en cause. L'étude montre aussi que le modèle de Givón, sur lequel nous nous sommes basés, ne donne pas de résultats très satisfaisants sur les *Essais*. Il est possible, sinon probable, que ce modèle fonctionne mieux sur des textes narratifs, moins « digressifs » et centrés de manière constante sur quelques référents de discours. On peut aussi supposer qu'il est plus adapté à la progression à thème constant qu'aux progressions à thème linéaire ou éclaté (Combettes 1983). La prise en compte des progressions thématiques, mais aussi des relations de discours, apporterait sans nul doute des données complémentaires, mais complexifierait d'autant la recherche, dont on a vu que même avec une approche limitée elle pose d'assez grandes difficultés méthodologiques.

Malgré toutes ses limites, notre étude donne une illustration de la manière dont les chaînes de référence, en participant de la continuité comme de la discontinuité textuelle, créent un maillage local qui n'est ni totalement dépendant, ni totalement indépendant de la structure textuelle. Pour compléter la recherche, on pourrait combiner les différentes approches que nous avons suivies les unes indépendamment des autres. On pourrait aussi prendre en compte l'ordre d'apparition des chaînes à l'intérieur des unités structurelles, les positions initiales et finales ayant *a priori* un statut particulier. Enfin, la cadence des chaînes à l'intérieur des unités mériterait d'être étudiée à son tour. Autant de pistes pour de futurs travaux.

Remerciements

Nous tenons en tout premier lieu à remercier les relecteurs de la revue Discours pour la qualité et la précision de leurs remarques qui nous ont permis d'améliorer le

présent article. Notre gratitude va également à nos collègues du projet Democrat, notamment Matthieu Decorde, Serge Heiden et Bénédicte Pincemin pour toute l'aide apportée à notre recherche et aux annotatrices Zeina Tmart et Audrey Arpin-Pont pour leur travail si précieux.

Références

DIDEROT, D. *Essais sur la peinture*, éd. par Gita MAY, Paris, Hermann, 1984 (revue en 2007).

ADAM, J-M. 2018. *Le paragraphe : entre phrases et texte*, Paris : Armand Colin.

ARIEL, M. 1990. *Accessing Noun-Phrase Antecedents.* London & New York : Routledge (Croom Helm Linguistics Series)

BENREKASSA, G. 1992. Diderot, l'absence d'œuvre. In G. BENREKASSA, M. BUFFAT & P. CHARTIER (éd.), *Études sur le Neveu de Rameau et le Paradoxe sur le comédien de Denis Diderot*, *Cahiers textuel*, n° 11, p. 133-140.

BESSONNAT, D. 1988. Le découpage en paragraphes et ses fonctions, *Pratiques* 57, 81-106.

CHAROLLES, M. 1988. Les plans d'organisation textuelle : périodes, chaînes, portées et séquences, *Pratiques* 57, 3-13.

CHAROLLES, M. 1995. Cohésion, cohérence et pertinence du discours, *Travaux de linguistique* 29, 125-151.

COMBETTES, B. 1983. *Pour une grammaire textuelle : la progression thématique*. Bruxelles : De Boeck / Paris : Duculot.

FERNANDES, A. 2014. Les Salons de Diderot : une chronique de la création artistique, *Carnets* [En ligne], Deuxième série - 2 | 2014, mis en ligne le 30 novembre 2014, consulté le 25 juin 2019. URL : http://journals.openedition.org/carnets/1295 ; DOI : 10.4000/carnets.1295.

GIVON, T. 1983. An introduction, dans T. GIVON (éd), *Topic continuity in discourse. A quantitative cross-language study*, Amsterdam/Philadelphia, John Benjamins Publishing Company (Typological Studies in Language, vol. 3).

HEIDEN, S., MAGUE, J.-P., PINCEMIN, B. 2010. TXM : Une plateforme logicielle open-source pour la textométrie – conception et développement. In I. C. Sergio Bolasco (Ed.), Proc. of *10th International Conference on the Statistical Analysis of Textual Data - JADT 2010*) (Vol. 2, p. 1021-1032). Edizioni Universitarie di Lettere Economia Diritto, Roma, Italy. Online.

HEIDEN, S. 2019. Manuel de TXM, Extension Annotation URS (Unité-Relation-Schéma) version 1.0 [En ligne], DOI : 10.5281/zenodo.3267345

LANDRAGIN, F., POIBEAU, T., VICTORRI, B. ANALEC: a New Tool for the Dynamic Annotation of Textual Data. *International Conference on Language Resources and Evaluation* (LREC2012), May 2012, Istanbul, Turkey. pp.357-362. halshs-00698971

LANDRAGIN, F., SCHNEDECKER, C. (éd.) 2014. *Langages* 195, *Les chaînes de référence*, Paris : Larousse/Armand Colin.

LONGACRE, R. 1979. The Paragraph as a Grammatical Unit, dans T. GIVÓN (éd.), *Syntax and semantics. Discourse and Syntax*, vol 12, New York : Academic Press, 115-134.

MARANDIN, J.-M. 1988. À propos de la notion de thème de discours. Éléments d'analyse dans le récit. *Langue française* 78, 67-87.

MITTERAND, H. 1985. Le paragraphe est-il une unité linguistique ?, dans R. LAUFER (éd.), *La notion de paragraphe*, Paris, Editions du CNRS, 85-95.

OBRY, V., GLIKMAN, J., GUILLOT-BARBANCE, C. et PINCEMIN, B. 2017. Les chaînes de référence dans les récits brefs français : étude diachronique (XIII^{e}-XVI^{e} s.), *Langue française* 195, 91-110.

SCHNEDECKER, C. 2017. Chaînes de référence et variations selon le genre, *Langages* 195, 23-42.

SCHNEDECKER, C., LANDRAGIN, F. 2014. Les chaînes de référence : présentation, *Langages* 195, 3-22.

VASAK, A. 2007. La question du genre dans les *Salons*. In : G. CAMMAGRE & C. TALON-HUGON (éd.), *Diderot, l'expérience de l'art. Salons de 1759, 1761, 1763 et Essais sur la peinture*, Paris, PUF, 11-25.

WIDLÖCHER, A., MATHET, Y. 2009. La plate-forme Glozz: environnement d'annotation et d'exploration de corpus. In *Actes de la 16e Conférence Traitement Automatique des Langues Naturelles* (TALN'09), session posters (p. 10). Senlis, France, France. Retrieved from https://hal.archives-ouvertes.fr/hal-0101196