Chaînes de référence dans le corpus Democrat : une analyse en diachronie longue

Frédéric Landragin, Julie Glikman, Catherine Schnedecker, Amalia Todirascu

Chaînes de référence dans le corpus Democrat : une analyse en diachronie longue

Frédéric Landragin, Julie Glikman, Catherine Schnedecker, Amalia Todirascu[1]

DRAFT auteurs

Introduction

Le corpus Democrat, annoté en expressions référentielles et en chaînes de référence (désormais CR), est paru fin 2019 et a déjà fait l'objet de plusieurs études, la plupart initiées dans le cadre du projet ANR Democrat[2], souvent à un moment où le corpus n'était pas encore complet (voir Schnedecker *et al.* (dir.) 2017 ; Landragin (dir.) 2021). Trois ans plus tard, avec le recul, force est de constater qu'il manque encore des statistiques sur l'ensemble du corpus, ainsi que des analyses remettant en perspective les annotations dans leur contexte textuel, voire les croisant avec d'autres annotations.

Or le corpus avait été constitué en intégrant des textes depuis le 11e jusqu'au 21e siècle avec l'objectif de permettre des analyses en diachronie longue. Ici, notre objectif est triple : nous voulons premièrement réaliser une présentation critique du corpus, aussi bien au niveau du choix des textes (répartition diachronique[3]) qu'à celui des choix d'annotation (nature des données et indications du manuel d'annotation[4]) ; nous présentons deuxièmement quelques statistiques générales qui permettent d'apprécier la constitution du corpus dans sa dimension diachronique ; enfin, nous approfondissons les analyses amorcées précédemment en nous focalisant sur leur contexte, notamment sur les liens entre expressions référentielles et discours direct.

Nous contribuons ainsi à explorer la nature et la constitution des CR en diachronie : nature des « maillons » (expressions référentielles appartenant à une CR), nombre de maillons, longueur des CR (en nombre de maillons, voire en nombre de mots couverts dans le texte), densité référentielle (quotient du nombre de maillons divisé par le nombre de mots), distance entre deux maillons (en nombre de mots), nombre de référents dans un texte, nature des premières mentions (nom propre, pronom), etc. Outre des données relevant de l'évolution connue du français, par exemple la diminution des sujets nuls, notre objectif est de mettre au jour des arguments quant à l'évolution de la structuration textuelle. L'exploitation du corpus Democrat nous amène à une meilleure compréhension non seulement du fonctionnement des CR, mais aussi des genres textuels, et de leur évolution diachronique respective.

1 Les auteurs ont contribué de manière équivalente à cette recherche.
2 ANR-15-CE38-0008, 2016-2020, rassemblant des équipes notamment des laboratoires Lattice (Paris), LiLPa (Strasbourg), ICAR et IHRIM (Lyon). Site du projet : https://www.lattice.cnrs.fr/democrat/. Le corpus est distribué sur Ortolang, cf. www.ortolang.fr et https://hdl.handle.net/11403/democrat/v1.1.
3 Voir la composition du corpus : https://www.lattice.cnrs.fr/democrat/files/ANR-15-CE38-0008-DEMOCRAT_livrable_L1_corpus.pdf.
4 https://www.lattice.cnrs.fr/democrat/files/ANR-15-CE38-0008-DEMOCRAT_livrable_methodo.pdf.

1. Présentation critique du corpus Democrat

Début 2016, lorsque le projet Democrat s'est mis en place, il existait déjà un corpus de grande taille annoté en anaphores, le corpus ANCOR : 487 000 mots ; 116 000 mentions annotées ; 51 000 relations anaphoriques annotées[5]. Mais il n'existait pas de corpus de grande taille annoté en CR, qui permette d'apprécier leurs compositions ainsi que l'évolution historique de ces compositions. Le projet Democrat s'était donné comme objectif de concevoir, constituer et annoter manuellement un tel corpus (en tant que « *gold standard* »), pour du français écrit et non du français oral transcrit comme c'était le cas pour le corpus ANCOR.

L'objectif a été atteint, car le corpus Democrat, depuis sa parution, est considéré comme une ressource essentielle pour le français écrit (689 000 mots ; 198 000 mentions annotées, soit 20 000 CR, *cf.* Landragin (dir.), 2021), et participe à la place du français dans les ressources numériques pérennes (humanités numériques). Une partie du corpus, celle correspondant au français contemporain, a été intégrée dans une initiative internationale d'envergure, *Universal Anaphora*[6] (dir. M. Poesio), puis dans la fameuse ressource *Universal Dependencies*[7], plus précisément dans une partie nommée *CorefUD (Coreference-Universal Dependencies)*[8] (dir. A. Nedoluzhko *et al*.). Depuis 2022, CorefUD sert de données d'entraînement pour une campagne mettant en compétition des chercheurs en traitement automatique des langues, afin de détecter automatiquement les coréférences dans plusieurs langues. Cette campagne fait suite au workshop *Computational models of Reference, Anaphora and Coreference (CRAC)*[9] (dir. M. Ogrodniczuk *et al*.), qui s'est tenu en 2022 lors de l'*International Conference on Computational Linguistics (COLING).*

On peut cependant souligner quelques points faibles du corpus Democrat. Tout d'abord, les textes dans des états de langue autres que le français contemporain n'ont pas été réexploités à leur juste valeur. L'une des raisons est que la représentation des textes en ancien français ou en moyen français n'est pas aussi fournie qu'en français contemporain. Une autre raison est que les annotations diffèrent : même si elles obéissent à un même schéma d'annotation, elles mettent en avant des phénomènes de langue différents. C'est un point sur lequel nous n'avons pas assez focalisé notre attention lors de la conception du corpus, et nous y revenons ici.

De fait, la mise en œuvre du corpus Democrat s'est effectuée selon trois phases. La première a consisté à déterminer quels textes et extraits allaient constituer le corpus. Afin de rendre possibles des études variées, il a été décidé de regrouper des textes narratifs et non narratifs (par ex. techniques, encyclopédiques, journalistiques), ainsi que de regrouper des textes représentant diverses périodes du français écrit, du 11^{e} au 21^{e} s. Des textes issus de corpus existants (annotés ou non) ont été repris, afin d'offrir une nouvelle couche d'annotation à des données faisant déjà l'objet de recherches linguistiques. Pour le français médiéval, des textes issus de la Base de Français Médiéval[10] ont ainsi été retenus.

On a alors pu passer à la deuxième phase : après une série d'expérimentations, le schéma d'annotation du corpus a été conceptualisé, puis instancié dans différents outils d'annotation.

5 https://hdl.handle.net/11403/ortolang-000903
6 https://universalanaphora.github.io/UniversalAnaphora/
7 https://universaldependencies.org/
8 https://ufal.mff.cuni.cz/corefud
9 https://sites.google.com/view/crac2022/
10 http://bfm.ens-lyon.fr/

Pour compléter, un manuel d'annotation regroupant consignes, exemples prototypiques et exemples particuliers a été préparé et rédigé. Ce manuel, dont la version finale comporte 30 pages, a circulé entre les membres du projet Democrat, mais à une époque où seuls des tests d'annotation sur du français contemporain avaient été effectués. Le manuel enchaîne environ 160 exemples, tous ou presque en français contemporain. Les seuls exemples dans des états de langue plus anciens se focalisent sur des noms propres et des configurations comme le complément de nom : « Achaz roy de Juda » et « Le roy Ochosias fils de Joram roy de Juda et d'Athalie ». Ajouter des exemples supplémentaires aurait apporté une aide appréciable aux annotateurs diachroniciens.

La troisième phase a alors consisté en l'annotation proprement dite, dans les quatre laboratoires partenaires du projet (cf. note 1). Afin de vérifier au fur et à mesure la reproductibilité des annotations, des annotations doubles ont été effectuées sur 5% du corpus, ce qui a permis de calculer l'accord inter-annotateurs et de fournir des indicateurs sur la fiabilité du travail réalisé (Landragin (dir.) 2021). C'est lors de cette phase que les annotateurs diachroniciens ont été confrontés à des situations non prévues, ou prévues d'une manière quelque peu différente, pour le français contemporain.

Clairement, nous étions partis sur l'hypothèse que de nombreux exemples génériques suffiraient à couvrir 99% des situations d'annotation possibles, les cas restants requérant soit une prise de décision de la part de l'annotateur, soit un échange avec d'autres annotateurs voire l'ensemble des participants du projet – pouvant conduire à une correction du manuel. En pratique, plusieurs aspects ont soulevé des questions, notamment : 1. sujets non exprimés ; 2. non prise en compte des dialogues et des discours indirects ; 3. catégorisation des mentions ; 4. traitement du texte par la plateforme TXM[11].

Concernant le premier aspect, le manuel présente l'exemple « Pierre boit et fume », dans lequel le deuxième verbe n'a pas de sujet exprimé. Afin de rendre compte de la coréférence entre les deux actants, il a été décidé d'annoter le sujet zéro du deuxième verbe, en tant que maillon de la CR portant sur le référent Pierre. Nous considérons ainsi deux types de maillons : les maillons forts, exprimés explicitement, comme « Pierre », et les maillons faibles, invisibles mais pourtant cognitivement présents, comme le sujet zéro (Landragin, 2011). Pour ne pas annoter une espace, nous avons choisi d'annoter le verbe support, « fume ». Il y a donc coréférence entre « Pierre » et « fume », et c'est à l'expert humain de comprendre que si l'un des maillons est un verbe plutôt qu'un groupe nominal, c'est qu'on se trouve en présence d'un sujet zéro. Cette façon d'annoter n'est pas idéale. De fait, les oublis des annotateurs de français contemporain sont fréquents : quand on ne peut pas se raccrocher à la matérialité du texte, on tend à ignorer le phénomène. En ancien français, la fréquence et l'importance du phénomène empêche tout oubli. Mais celui-ci peut survenir en français moderne, et le risque d'inhomogénéité des annotations est donc réel. Avec le recul, il nous semble qu'un prétraitement des textes pour en faire ressortir les sujets non exprimés, par exemple avec le caractère « Ø », aurait permis aux annotateurs de ne pas rater le phénomène, et donc de l'annoter.

Le deuxième aspect pose la question du recours à des annotations de la structure et la composition du texte. En l'absence de telles annotations – qu'il s'agisse de délimiter les tours

11 https://txm.gitpages.huma-num.fr/textometrie/index.html

de parole dans les dialogues (et les dialogues eux-mêmes), de délimiter les passages au discours indirect, voire les subordonnées qui fonctionnent dans un plan narratif secondaire –, le texte est pris comme un tout monolithique, dans lequel les mentions sont toutes annotées comme faisant partie du plan narratif principal. Cela n'a pas posé de problème lors des premières annotations, qui se sont faites sur des textes narratifs sans dialogues. Pour garder l'homogénéité, on a décidé d'annoter les mentions appartenant à des dialogues exactement comme des mentions du plan narratif principal. Par conséquent, une CR peut tout à fait contenir des alternances de « je » et « tu ». Sans indication qu'il y a changement de locuteur, les données restent en quelque sorte incomplètes. Nous avions à l'époque considéré que cet aspect pourrait faire l'objet d'une annotation automatique *a posteriori*. Avec le recul, force est de considérer qu'un tel post-traitement pose de nombreux problèmes, la délimitation des dialogues et des discours indirects n'étant pas toujours signalée de la même manière.

Le troisième aspect pose cette fois la question du besoin d'annotations complémentaires sur les mentions elles-mêmes : dans le but d'annoter le plus rapidement possible, il a été décidé que la catégorie de la mention (nom propre, pronom, groupe nominal défini, groupe nominal indéfini, etc.) ne serait pas annotée à la main mais identifiée automatiquement par un outil de traitement automatique des langues, fondé sur la syntaxe et le repérage des déterminants. Or les amalgames viennent compliquer cette tâche, qui n'a ainsi pas pu être menée à bien : « du » dans « le fils du boulanger » n'est pas un déterminant indéfini comme dans « du beurre » mais l'amalgame de « de » et « le », seul ce deuxième terme étant utile pour identifier la catégorie du complément du nom, à savoir « groupe nominal défini ». La syntaxe elle-même, avec par exemple le recours à un parseur, pose problème car les mêmes règles ne s'appliquent pas en français médiéval ou en français contemporain (« les chevaliers le Roy » *vs* « les chevaliers du roi », par exemple). Une macro TXM a été développée pour identifier automatiquement la catégorie de chacune des mentions annotées. Malheureusement, les résultats de cette macro ne sont pas parfaits : elle fait des erreurs, qu'il faut donc corriger manuellement. Face à la quantité de travail supplémentaire requis, il a été décidé que les annotations en catégories ne seraient pas publiées dans la première version du corpus Democrat.

Avec le quatrième et dernier aspect, nous explorons un autre problème technique, qui intervient surtout quand on cherche à faire des statistiques sur l'ensemble du corpus : lorsque TXM charge un texte, il procède à une tokenisation et il applique un analyseur morphosyntaxique, TreeTagger. Or ce n'est pas la même version de TreeTagger qui sert pour l'ancien français et le français contemporain. Comme pour la syntaxe, on se retrouve tributaires de problèmes d'exploitation qui dépassent le cadre de la référence et de la coréférence. Cet aspect sera transparent pour l'utilisateur linguiste qui se sert du corpus comme d'un réservoir d'exemples attestés, mais sera plus critique pour l'utilisateur travaillant dans le cadre de la textométrie ou du traitement automatique des langues.

Soulignons cependant que ces problèmes techniques sont loin d'être insurmontables pour tout chercheur du domaine de la linguistique de corpus : bien souvent, un simple script permet de trouver des solutions à ces problèmes, y compris ceux relevant de phénomènes syntaxiques spécifiques. Autrement dit, le corpus Democrat présente certes quelques points faibles, mais il représente malgré tout un grand pas en avant dans l'étude des expressions référentielles et des CR en français.

2. Statistiques générales sur le corpus Democrat

Pour calculer les propriétés des textes en fonction du siècle, nous avons appliqué les scripts disponibles dans TXM pour chaque texte du corpus : calcul du nombre total de mentions et de CR (à partir de 3 maillons), de la densité référentielle (nombre de maillons divisé par la taille du corpus) et, pour chaque CR, de la stabilité référentielle. Cette dernière correspond à la manière dont les différents maillons d'une CR utilisent (ou non) les mêmes lexèmes. Elle se calcule à l'aide du coefficient de stabilité de Perret, c'est-à-dire du quotient du nombre d'anaphores nominales divisé par le nombre de désignations différentes. La distance inter-maillonnaire (entre 2 maillons consécutifs), la distance médiane (indiquant la proportion de CR très longues ou très courtes) et la longueur des CR sont aussi des paramètres qui varient avec le genre textuel (Schnedecker, 2014), et que nous avons comparé en diachronie. Le tableau 1 résume les principales propriétés calculées pour chaque siècle.

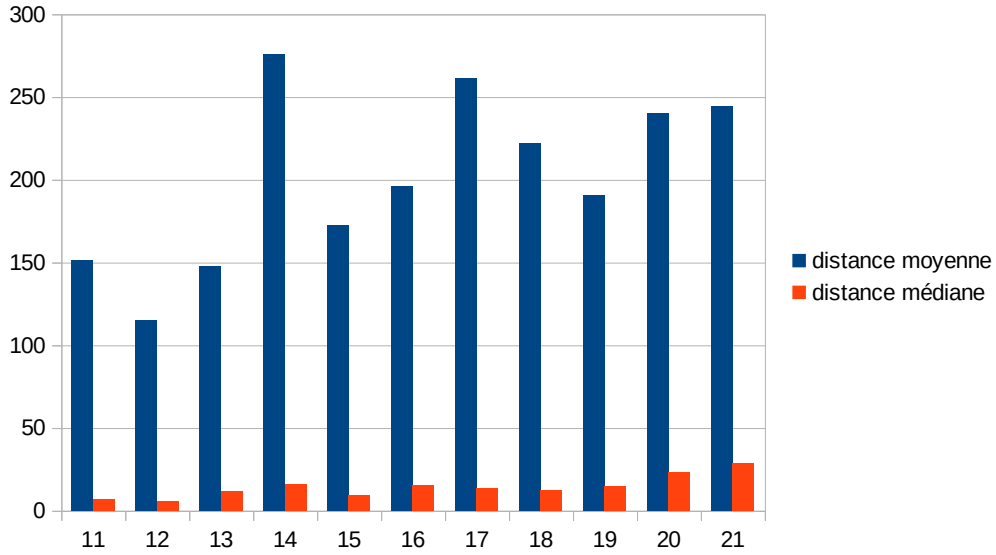
Tableau 1 : statistiques effectuées par siècle (textes narratifs et non-narratifs).

Siècle	Nb tokens	Maillons	Référents	Densité référentielle	Stabilité Ref	Distance intermaillons	Distance médiane	Longueur moyenne des chaînes	Longueur médiane
11	13059	3291	127	25,20 %	1,48 %	151,2332	7	25,91	3
12	11647	2662	104	22,86 %	1,30 %	115,0613	6	25,59	6
13	21729	4699	423	21,58 %	1,59 %	147,98	12	13,59	9,00
14	34616	5980	642	17,34 %	1,76 %	275,96	15,67	9,45	6,70
15	34416	6395	430	18,54 %	1,33 %	172,32	9,33	17,38	8,50
16	61216	9606	1453	16,05 %	1,54 %	195,95	15,40	11,21	3,25
17	86147	14613	2714	17,25 %	1,73 %	261,20	13,43	11,83	9,25
18	66889	9737	1018	14,61 %	1,64 %	222,26	12,17	10,15	4
19	161897	22524	1952	13,93 %	1,67 %	190,89	14,85	14,17	20,41
20	175434	25361	2392	14,73 %	1,76 %	240,01	23,13	11,84	10,53
21	21801	3015	320	13,81 %	2,11 %	244,50	28,50	9,89	5,00

Tout d'abord, nous comparons les paramètres entre siècles, textes narratifs et non-narratifs confondus (tableau 1). La densité référentielle est plus importante avant le 14^{e} (supérieure à 20 %). Entre le 14^{e} et le 17^{e}, le nombre de maillons présents dans les textes reste important, par rapport à la taille du corpus (entre 15 % et 20 %). A partir du 18^{e}, ce paramètre est en dessous de 15 %. Il n'y pas de différence notable pour le coefficient de stabilité, sauf pour le corpus du 21^{e} s. où on dépasse les 2 %. Cette différence s'explique par le type des deux textes du 21^{e} : une convention juridique et un article de Wikipédia (les deux non-narratifs). Dans le texte juridique, on constate la répétition de la même expression référentielle qui indique un même référent : on évite l'ambiguïté et la possibilité d'erreurs d'identification du référent, le coefficient de stabilité est alors plus important. Des variations importantes sont observées entre textes narratifs et non-narratifs : les CR concernant le personnage principal (dans les textes

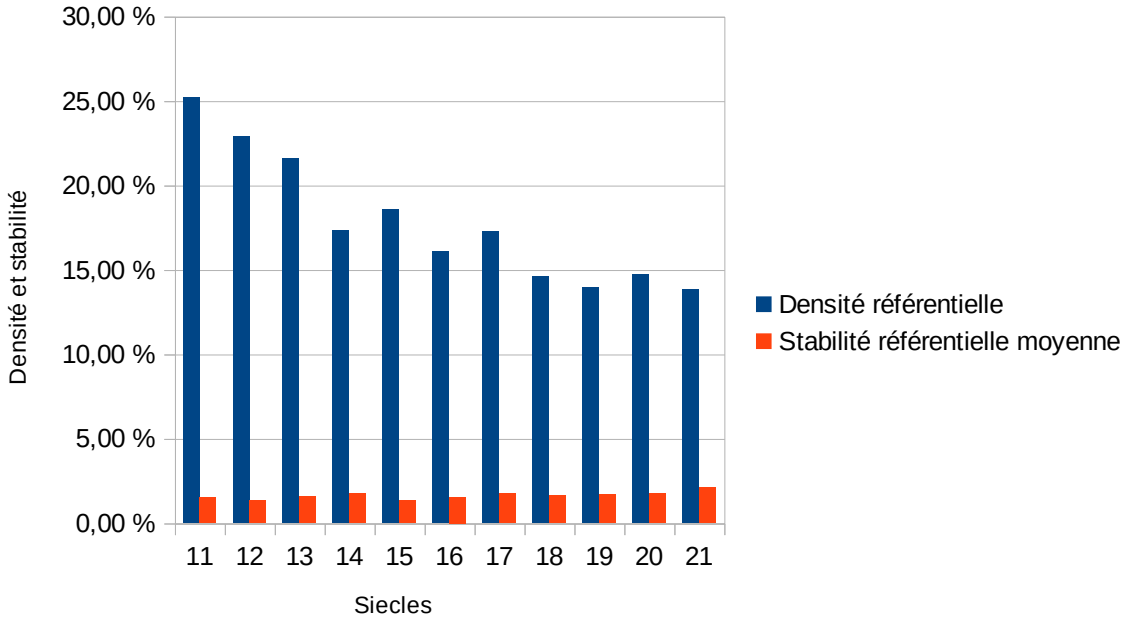
narratifs) ont un coefficient de stabilité beaucoup plus grand que les CR présentes dans les textes non-narratifs (peu de variation lexicale).

Figure 1 : distance moyenne entre deux maillons et distance médiane pour tous les textes.



D'autres paramètres étudiés sont la distance inter-maillonnaire moyenne et médiane (figure 1). En particulier, les valeurs sont très importantes pour les textes des 14^{e}, 17^{e}, 20^{e}, 21^{e} s. Dans ces 4 parties de corpus, nous constatons une présence plus importante des textes non-narratifs du domaine juridique ou encyclopédique (notamment aux 20^{e} et 21^{e} s.). Pour ces textes, les référents sont répétés plusieurs fois et repris à plusieurs moments dans le texte, mais parfois à plusieurs articles et paragraphes de distance. Ce comportement explique la différence entre les divers siècles et les distances entre 100 et presque 300 mots, alors que la distance médiane ne dépasse pas 10. La distance médiane nous indique la distribution des valeurs pour la distance inter-maillonnaire. Plus de la moitié des distances est inférieure à 10, les maillons sont en général plus rapprochés, sauf pour les cas mentionnés.

Figure 2 : densité référentielle et stabilité référentielle moyenne pour tous les textes.



Avec la longueur moyenne, on constate qu'un nombre important de CR très courtes (longueur inférieure à 10) coexistent avec des CR très longues construites sur quelques référents (en particulier pour les textes narratifs). La longueur moyenne est plus importante aux 11^{e} et 12^{e} et ce paramètre est plus réduit pour tous les autres siècles.

La composition des CR peut varier d'un genre à l'autre ou en diachronie. Ainsi, les syntagmes nominaux définis (SNDéf) entrent majoritairement dans la composition des CR et suit l'évolution en diachronie. La proportion des SNDéf augmente (de 8,64 % pour le 11^{e} à 50,20 % pour le 21^{e} s.) mais aussi des SN indéfini (de 1,62 % à 7,28 %). En règle générale, la longueur moyenne dépasse la longueur médiane. Plus la valeur de la longueur médiane est réduite, plus on constate un grand nombre de CR très courtes.

Nous comparons ensuite la densité référentielle et la stabilité référentielle dans les textes narratifs en diachronie (figures 2 et 3). Les valeurs se situent entre 10,55 % et 25,20 % pour la densité référentielle. Les textes narratifs avant le 16^{e} présentent une densité référentielle plus importante (plus de 20,55 %). L'analyse des résultats obtenus pour la densité référentielle calculée pour les textes narratifs et non-narratifs montrent que les textes narratifs ont une densité référentielle plus haute que les textes non-narratifs. Ce phénomène s'explique par la richesse de référents dans les textes narratifs (personnages, noms, lieux, objets). Les textes non-narratifs répètent le même référent plusieurs fois dans le texte.

La stabilité référentielle se situe entre 1,27 et 2,06 et ne varie pas avec le type de textes. A partir du 17^{e} on constate que toutes les valeurs dépassent le seuil de 1,6 (à une exception près, l'extrait du *Capitaine Fracasse*). Les CR ayant une stabilité référentielle plus importante sont en général celles des personnages principaux dans les textes narratifs. On constate une légère hausse des valeurs pour les textes non-narratifs du 21^{e} s.

Figure 3 : comparaison de la densité et de la stabilité référentielle entre textes narratifs et non-narratifs.

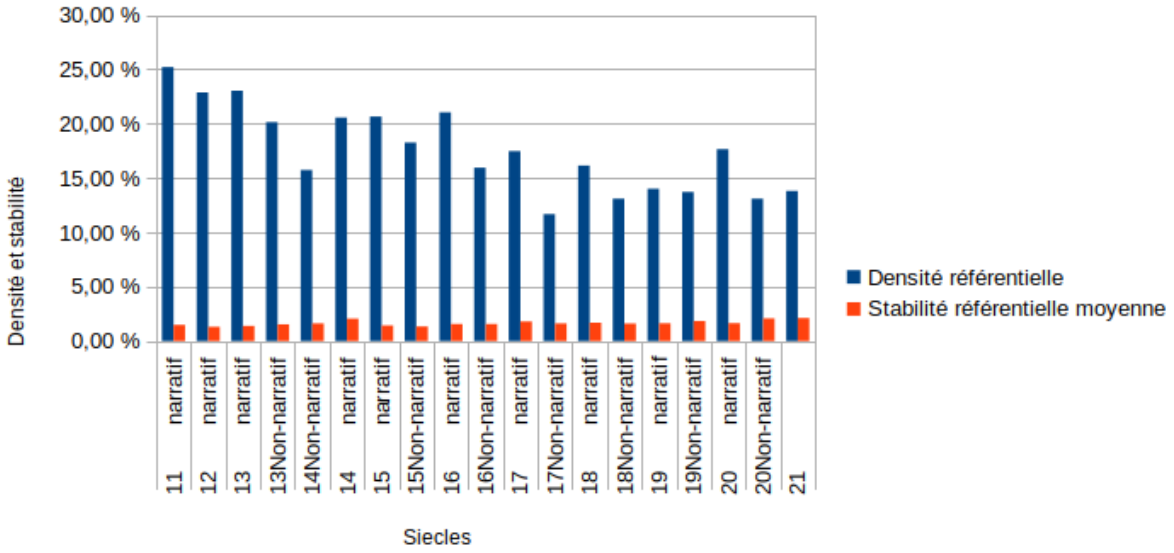


Figure 4 : distance moyenne et médiane et leur répartition entre textes narratifs et non-narratifs.

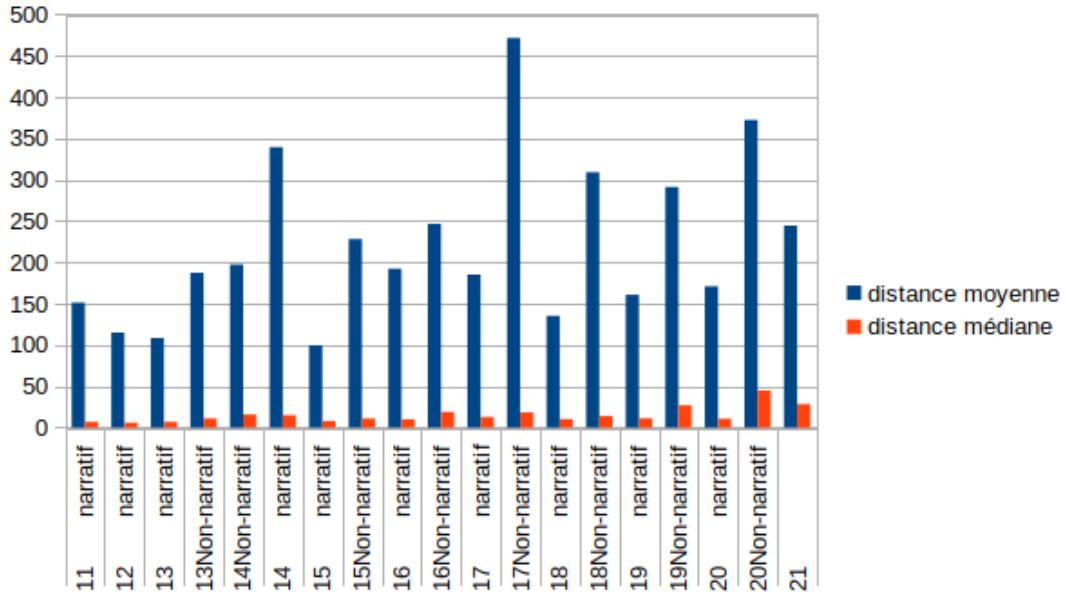
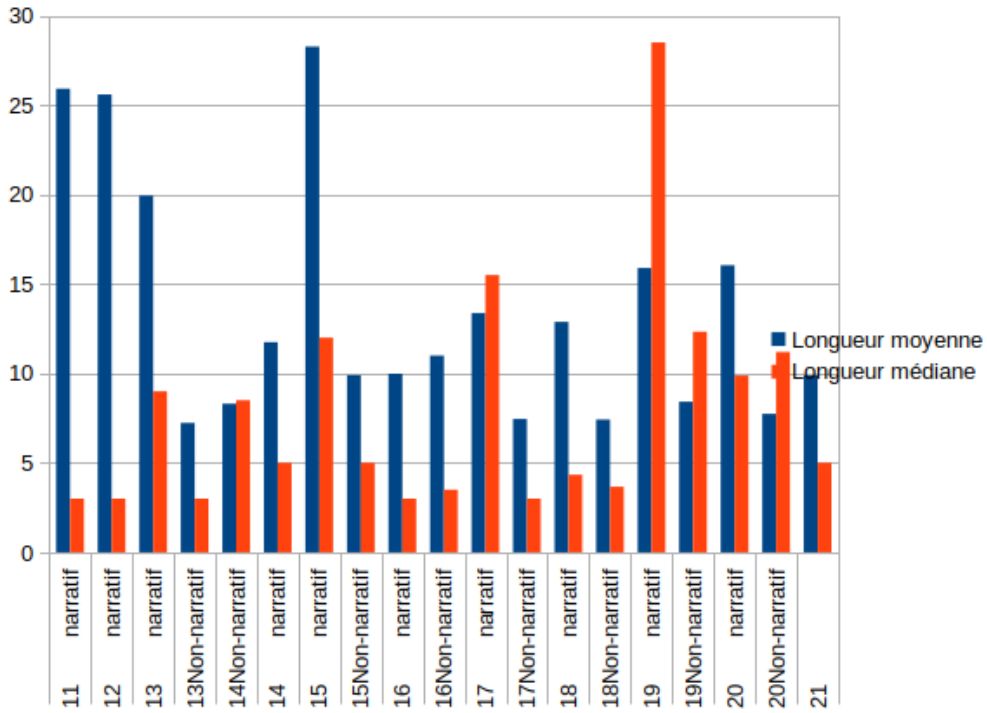


Figure 5 : longueur moyenne et médiane et leur répartition entre textes narratifs et non-narratifs.

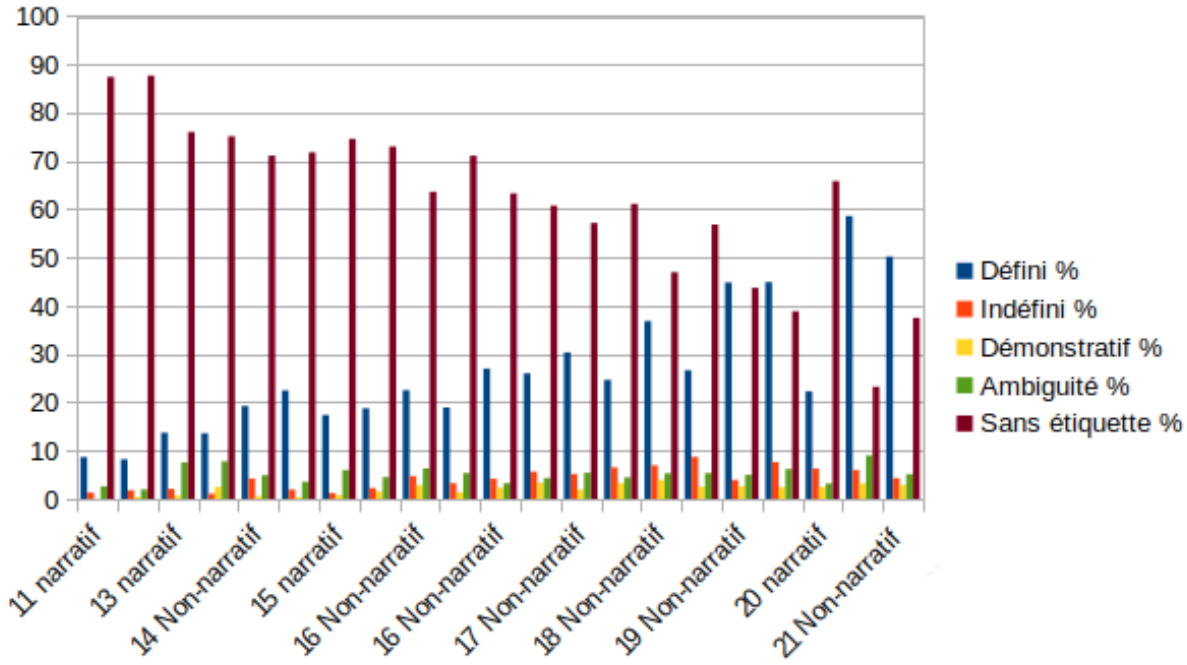


Pour tous les siècles, la distance inter-maillonaire est beaucoup plus grande dans les textes non-narratifs que dans les textes narratifs (à l'exception du 14e) : parfois on passe du simple au double (figure 4). Les distances inter-maillonaires très grandes indiquent un même référent qui est repris après plusieurs paragraphes, comme c'est le cas dans les textes juridiques, mais parfois d'un personnage principal d'un texte narratif. La distance médiane reste assez réduite (sauf pour les textes non-narratifs du 20e) ce qui se traduit par un grand nombre de CR courtes présentes dans le texte.

C'est le contraire pour la longueur moyenne de CR : elle est plus grande dans les textes non-narratifs et moins importante dans les textes narratifs (parfois moins de la moitié). La longueur médiane est plus importante que la longueur moyenne pour les textes non-narratifs des 17e, 19e et 20e s (figure 5).

Certaines catégories de maillons sont privilégiées dans les textes non-narratifs récents, par rapport aux textes non-narratifs anciens : la plupart choisissent des maillons de type SN défini, puis indéfini. Ce sont les choix les plus communs aussi bien dans les textes narratifs que non-narratifs (figure 6).

Figure 6 : catégories de maillons et leur répartition dans les textes narratifs et non-narratifs.



3. Expressions référentielles et discours rapporté

Dès l'instant où l'on conçoit les CR comme la suite des expressions coréférentielles d'un texte, deux questions émergent : cette suite est-elle, y compris dans des textes longs comme les romans, essais philosophiques, etc., conçue comme un tout ininterrompu, abstraction faite de la charge cognitive provoquée par le traitement au long cours d'une seule CR et sachant que la plupart des textes sont pluri-référentiels ? Ou, compte tenu du fait que les CR constituent l'un des plans de l'organisation des textes (Charolles 1988) et que, de ce fait, elles interagissent avec les autres plans, ceux-ci sont-ils susceptibles d'y imposer des coupures/ruptures ? On sait, en effet, que le découpage en paragraphes (participant de ce que Charolles nomme les *séquences*) est souvent corrélé à la présence de formes de basse accessibilité référentielle[12] redénommant le référent, comme pour signaler un double redémarrage typographique et référentiel[13] (cf. Capin *et al*. 2021 ; Oberlé et al. 2018 ; Schnedecker 1997 & 2021).

A ce point de vue, le découpage que constitue le discours rapporté mérite l'attention d'autant qu'il combine, si l'on peut dire, les moyens typographiques (double points et guillemets), lexico-syntaxiques (présence de *verba dicendi* et, dans le cas du discours indirect, de subordonnants, marqueurs de portée, selon Charolles, art. cit.) et énonciatifs (hétérogénéité de locuteurs/énonciateurs entre discours cité et discours citant) (voir Marnette 2006 pour cette question en français médiéval). Or, force est de constater que les études à ce propos sont peu

12 Selon la théorie de Mira Ariel (1990), les expressions référentielles se classent selon une échelle d'accessibilité, qui exprime la manière plus ou moins aisée dont le référent est cognitivement accessible, ceci en fonction de la forme prise par l'expression référentielle. La plus forte accessibilité se traduit par un pronom non exprimé, ce qui correspond à nos maillons faibles. Tous les autres degrés d'accessibilité correspondent à nos maillons forts. Parmi ceux-ci, les formes à plus faible accessibilité sont typiquement les noms propres et les SN définis.
13 Les choses sont en réalité plus complexes comme cela a été démontré dans les travaux cités ci-dessus.

nombreuses (par ordre chronologique, Schnedecker 1990[14] ; [Reichler-]Béguelin 1997 ; Perret 2008 ; Veniard 2009 et Gollut & Zufferey 2018) et, qui plus est, extrêmement éclatées tant au plan des genres discursifs étudiés (romans, Perret, Gollut & Zufferey, Schnedecker ; presse, [Reichler-]Béguelin et Veniard[15]) que des types de discours rapporté (discours direct exclusivement pour la plupart des études, hormis celle de Gollut & Zufferey qui porte sur le discours indirect libre) ou encore des expressions référentielles et de leurs configurations discursives d'accueil. C'est ainsi que Schnedecker (1990) étudie les formes de reprise, dans le discours enchâssant, d'un maillon initié dans le discours rapporté, [Reichler-]Béguelin et Gollut & Zufferey l'usage du pronom personnel en contexte énonciativement hétérogène, Veniard celui des SN démonstratifs, et Perret, les anaphores résomptives, démonstratives le plus souvent, portant sur la forme du propos (*a ces paroles, a icez moz*, …) qui marquent le retour au récit en français médiéval.

Trois points ressortent de ces études aussi disparates soient-elles. Premièrement, la présence de discours rapporté est de nature à compliquer la résolution de la coréférence, notamment quand la référence situationnelle n'est pas élucidée comme dans (1), ce que souligne [Reichler-] Béguelin dans la citation ci-après :

> 1) Ensuite le présentateur revint et dit « Les petits enfants vous pouvez **en** prendre il y **en** a de toutes les couleurs (élève, cité par Charolles, in [Reichler-] Béguelin, 1997, 42, son ex. 11)
>
> La résolution des expressions référentielles qui figurent dans la portée d'un discours cité est […] un phénomène d'une redoutable complexité : il s'agit, pour l'allocutaire de E (= discours enchâssant), non seulement d'identifier les objets-de-discours réextraits de M (= mémoire discursive), mais aussi de construire les univers de croyance assignables aux interlocuteurs de e (= discours cité) ([Reichler-] Béguelin, 1997, 39)

Deuxièmement, les configurations possibles sont tellement disparates qu'elles semblent résister à toute tentative de modélisation. Par exemple, (2) illustre un cas de démarcation extrême entre discours enchâssé et enchâssant, renforcé par l'emploi d'un SN indéfini qui marque la nouveauté référentielle, semblant ainsi faire abstraction de la présence d'un maillon pronominal dans le discours rapporté qui saisit manifestement le référent en fonction de sa saillance situationnelle, ce que corrobore le geste de monstration :

> 2) […] Selena dit : La vie, après tout, n'a pas été une horreur sans répit. Vous savez, j'ai vraiment passé quelques années merveilleuses avec **elle**.
>
> Elle *montrait du doigt* **une enfant qui paraissait jaillir d'un portrait au mur**. (G. Paley, incipit de *Amies, Plus tard le même jour*, Schnedecker, 1993, 173, son ex. 21)

Inversement, dans (3), le référent est réinstancié par des SN définis ou démonstratifs, en dépit de la variété des locuteurs-énonciateurs (le journaliste et les divers hommes politiques exprimant leur avis), ce qui assure, comme le précise Veniard, la continuité de la chaîne anaphorique et coréférentielle :

14 Cf. également 2021, 182 *et seq*.
15 Tout venant chez la première ; corpus d'article du *Monde* et du *Figaro* (entre juin 2003 et novembre 2004) pour la seconde.

3) **Le conflit des intermittents du spectacle** est central dans l'évaluation du passage de Jean-Jacques Aillagon Rue de Valois. Le constat général est sévère. La façon « catastrophique » dont le ministre de la culture a géré **l'affaire** […]

Premier faux pas, le calendrier : « Il aurait pu négocier **ce dossier** sensible de manière plus sereine […] » souligne Emmanuel Négrier […]. Hubert Astier […] estime que M. Aillagon a péché par témérité : **« Ce dossier** est une bombe à retardement […] » Le député (UMP) de Nancy évoque surtout une mauvaise appréciation **du dossier** : « […] Le problème est que l'administration du ministère connaissait mal **le dossier**. » […) (*Le Monde*, 01/04/2004, in Veniard, 2008, son ex. 18, ici abrégé)

De sorte que, troisièmement, l'on ne pourrait qu'apprécier des effets de continuité *vs* discontinuité induits par les marques référentielles et/ou leurs configurations d'occurrence, susceptibles comme le suggère [Reichler-]Béguelin (art. cit.), d'être surmarqués dans les cas de pronoms « sans antécédent » (cf. 1), de renforcer polyphonie et points de vue ou encore de provoquer des effets de réalisme, comme dans (4) où le pronom semble fidèle au *verbatim* de la victime :

4) […] Soudain elle se sentit poussée à terre.

Je suis tombée sur le coude. J'ai vu les étoiles. **Il** m'a pris mon sac. **Il** a dû l'arracher parce que je le tenais bien (Tribune de Genève, in ([Reichler-] Béguelin, 1997, 47, son ex. 31)

Cela étant, la question des CR et du discours rapporté comprend toujours de nombreux angles morts. Les études citées déjà anciennes reposent sur des extraits, exception faite de l'article de Veniard : un travail mené sur corpus devrait déterminer et quantifier *in extenso* la proportion de CR initiées et/ou développées dans des DR, celle de leur « reprise » dans le discours enchâssant, voire le nombre d'interlocuteurs susceptibles de mentionner tel ou tel référent et de l'élaborer. A quoi s'ajoute la question du genre de discours qui surdétermine cette question (cf. Veniard, art. cit., Schnedecker & Landragin, 2014, Schnedecker, 2021) : le discours rapporté de presse doit rapporter des propos effectivement tenus, il est soumis à des exigences de vérité, ce qui fait que le locuteur n'a pas toujours le choix de la (re-)dénomination des référents. Celui des romans doit seulement faire vrai(semblable). Enfin, le type de discours rapporté (direct/direct libre ou indirect, indirect libre) joue également un rôle crucial, le discours rapporté direct étant, comme on l'a dit, marqué par une forme de discontinuité textuelle là où, au contraire, le discours indirect libre gomme formellement l'hétérogénéité énonciative. C'est le cas de (5) où, malgré les apparences, la suite de pronoms n'émane pas du seul narrateur : le dernier doit être imputé à Coupeau « *personnellement* comptable de l'absurde déni » (Gollut & Zufferey, 2018) :

5) Coupeau ne connaissait qu'un remède, se coller sa chopine de cric, un coup de bâton dans l'estomac, qui **le** mettait debout. Tous les matins, **il** guérissait ainsi sa pituite. La mémoire avait filé depuis longtemps, son crâne était vide ; et **il** ne se trouvait pas plus tôt sur les pieds, qu'**il** blaguait la maladie. ***Il*** *n'avait jamais été malade*. (Zola, *L'Assommoir*, 1877, in Gollut & Zufferey, 2018)

La question des CR et du discours rapporté comporte ainsi plusieurs facettes et demanderait à prendre en compte les différents paramètres mentionnés ci-dessus. Le corpus Democrat ne comportant pas dans sa version actuelle, comme mentionné plus haut, une annotation des plans énonciatifs (délimitation du discours direct) ni des plans secondaires (subordination), une étude à grande échelle n'est pas possible. Nous proposons ici une analyse qualitative des contextes

sur un nombre restreint de textes issus du corpus ayant des caractéristiques de même type, afin de limiter les paramètres de variation (narratifs et romans), mais de siècles différents pour mener l'analyse en diachronie : *Eneas* (12e), *Jehan de Paris* (15e), *Clèves* (17e), *Voyage de Cyrus* (18e), *Ventre de Paris* (19e) et *Némoville* (20e). *Eneas* est en vers (forme dominante dans les écrits littéraires médiévaux), les autres en prose. Ces textes ont fait l'objet d'une étude sur le lien entre les CR et la structuration textuelle (Capin *et al.* 2021), ce qui nous permettra de compléter l'analyse en croisant les paramètres déjà observés. Les situations pouvant être très diverses, nous avons choisi de nous concentrer sur l'analyse des maillons de reprise de personnages après discours rapporté au style direct (désormais DD), en particulier deux cas de figure : a) la reprise du locuteur qui produit le discours et présent dans le discours (Loc1) (Loc1 « … Loc1 P1… » Loc1) ; b) la reprise de l'interlocuteur destinataire du discours et cité dans le discours (Loc2) (Loc1 « …Loc1 P1 / Loc2 P2… » Loc2).

Pour a), le maillon de reprise après le DD apparait sous forme de maillon de forte accessibilité (sujet nul pour les textes anciens) à partir de *Jehan* seulement. Dans *Eneas*, la reprise est un maillon de faible accessibilité, nom propre (NPro), peut-être en lien avec le type de narration « plate » dont on avait déjà observé l'effet sur les CR (Capin *et al.* 2021) :

> 6) Danz **Eneas** formant s'escrie. / « Par deu, fait **il**, […] ne **sai** ou **ge** la puisse querre ; […] si com fortune **me** demoine. » / Molt se dementot **Eneas**, (*Eneas*, v. 209-230)

Deux passages se distinguent dans *Eneas* : Didon, loc1, est reprise par un pronom (ProPer) (en subordonnée) et sujet nul ou ProPer dans la principale (v. 1271-1277 et 1278-1323). A partir de *Jehan*, la reprise de Loc1 est un maillon de forte accessibilité (sujet nul ou ProPer) de manière régulière.

Pour b), la possibilité de reprise par un maillon de forte accessibilité apparait plus tardivement, avec un seul cas de reprise par maillon de forte accessibilité dans *Jehan* et dans *Cyrus*. Dans *Ventre*, au contraire, on trouve majoritairement un maillon de forte accessibilité en reprise de Loc2 après DD comme de Loc1, et la présence des ProPer est l'usage répandu dans l'incipit. Cela peut être dû à deux facteurs. D'une part, l'absence d'ambiguïté liée au genre différent des deux personnages principaux dans l'incipit, alors que les reprises de Loc2 dans la suite du texte se font avec le NPro lorsque les deux locuteurs sont de même genre (Florent et Claude). D'autre part, dans ce texte la redénomination joue un rôle de changement de focalisation de point de vue (cf. Capin *et al.* 2021), qui peut aussi entrer en compte dans les reprises après DD. Les reprises par un SNDéf sont dans cet extrait liés à l'introduction d'un nouveau personnage, dont la première mention apparait dans le DD (ex. p. 7 *« ...**vous**... » et comme **l'homme** s'en allait…*). Dans *Nemoville*, on observe une alternance de reprise de Loc1 comme de Loc2 par ProPer, SNDéf (le prêtre) ou NPro (Roger ou Paul), mais ce texte est également caractérisé par leur forte présence dans les incises (dans *Ventre* en comparaison, les maillons des incises – quand il y en a – sont plutôt des ProPer). La reprise du locuteur après le DD par un maillon de forte accessibilité ne semble ainsi se développer qu'à la fin de la période médiévale et encore plus tardivement pour l'interlocuteur. Cela pourrait confirmer le statut spécifique du DD dans les textes anciens et sa plus grande intégration dans le fil de la narration dans la suite.

Conclusions

Dans cet article, nous avons adopté une approche diachronique longue pour présenter le corpus Democrat, fournir quelques statistiques sur les annotations en références et CR de ce corpus, et proposer de nouvelles directions d'analyse, en commençant par les liens entre discours direct et référence.

Bien que le corpus soit disponible depuis mi-2019, c'est la première fois que nous calculons des statistiques et observons des tendances – notamment celle de la diminution de la densité référentielle – en diachronie longue. Ces données viennent enrichir les annotations Democrat, qui peuvent encore permettre de nombreuses analyses qualitatives et quantitatives.

Nous avons également souligné l'intérêt de croiser des annotations – en l'occurrence les mentions référentielles et le discours direct – pour les extraits Democrat qui reprennent les mêmes textes que ceux annotés et analysés par d'autres initiatives, notamment par la Base de Français Médiéval. Cet aspect, qui montre l'importance de l'interopérabilité des données, se veut comme une preuve de faisabilité de ce qu'il est possible de faire avec les corpus dont nous disposons. Comme souvent, l'analyse linguistique n'est pas complète et ouvre des perspectives, d'une part en tenant compte d'autres notions comme le plan narratif principal ou secondaire, d'autre part en la systématisant sur une plus grande portion du corpus.

Références bibliographiques

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

Capin, D., Glikman, J., Schnedecker, C. & Todirascu, A. (2021). Le rôle des chaînes de référence dans la structuration textuelle : étude diachronique de l'ancien français au français moderne. *Langages*, 224, 87-107.

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

Gollut, J.-D. & Zufferey, J. (2018). « Le statut énonciatif des unités dites 'transposées' en discours indirect libre », *SHS Web of Conferences*, 46, EDP Sciences.

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

Landragin, F. (dir.) (2021) Un corpus annoté en chaînes de référence et son exploitation : le projet Democrat. *Langages* 2021/4 n°224.

Marnette, S. (2006). La signalisation du discours rapporté en français médiéval. *Langue française* 149, 31-47.

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, 67-141.

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, 91-110.

Perret, M. (2008). « Les marques de retour à la narration en français médiéval », *L'information grammaticale*, 118/1, 22-26.

Prévost, S. (2020). Une grammaire fondée sur un corpus numérique. In Marchello-Nizia, C., Combettes, B., Scheer, T. & Prévost, S. (2020). *Grande Grammaire Historique du Français (GGHF)*, De Gruyter, p. 37-53.

[Reichler]-Béguelin, M.-J. (1997). « Anaphores pronominales en contexte d'hétérogénéité énonciative : effets d'(in) cohérence », in De Mulder et al. (éds), *Relations anaphoriques et (in)cohérence*, Amsterdam-Atlanta, Rodopi, 31-54.

Schnedecker, C. (1993). « Le discours rapporté a-t-il des incidences sur les chaînes de référence ? Quelques observations », *Verbum*, 13/3, 165-190.

Schnedecker, C. (1997). *Nom propre et chaînes de référence*, Paris, Klincksieck.
Schnedecker, C. (2022). *Les chaînes de référence en français*, Ophrys.
Schnedecker, C., Glikman, J. & Landragin, F. (dir.) (2017) *Les chaînes de référence en corpus*. *Langue Française* 2017/3 n°195.
Veniard, M. (2009). « Anaphores lexicales en contexte d'hétérogénéité énonciative et effets pragmatiques associés », *Ci-Dit | Discours rapporté, citation et pratiques sémiotiques*, Nice.

Auteurs

Frédéric Landragin, CNRS, Lattice

Julie Glikman, Université de Strasbourg, LiLPa

Catherine Schnedecker, Université de Strasbourg, LiLPa

Amalia Todirascu, Université de Strasbourg, LiLPa