Le discours direct au Moyen Âge : vers une définition et une méthodologie d'analyse

Céline Guillot, Alexei Lavrentiev, Bénédicte Pincemin, Serge Heiden

Céline Guillot
ENS de Lyon-UMR ICAR, celine.guillot@ens-lyon.fr
Alexei Lavrentiev
CNRS-UMR ICAR, alexei.lavrentev@ens-lyon.fr
Bénédicte Pincemin
CNRS-UMR ICAR, benedicte.pincemin@ens-lyon.fr
Serge Heiden
ENS de Lyon-UMR ICAR, slh@ens-lyon.fr

LE DISCOURS DIRECT AU MOYEN AGE : VERS UNE DEFINITION ET UNE METHODOLOGIE D'ANALYSE

Introduction

Bien que le français oral 'authentique' antérieur au XXe siècle et aux premiers enregistrements de la parole reste à jamais hors de notre portée, on s'est beaucoup intéressé – dans la lignée des travaux pionniers de P. Zumthor (1983, 1984 et 1987) – aux multiples rapports et influences qu'il est possible d'établir entre les premiers témoins de la langue vernaculaire et la langue et la littérature orales.

De nombreuses recherches menées en parallèle dans un cadre linguistique ou ethnographique ont par ailleurs montré que la dichotomie entre oral et écrit était trop simple : d'une part, il est certainement plus approprié d'établir un continuum entre ces deux pôles (Rankovic *&* *al.* 2010), d'autre part il convient d'établir des catégories plus fines et plus précises, en distinguant plusieurs types d'oral et d'écrit (cf. notamment les travaux de C. Blanche-Benveniste) et en dissociant le canal par lequel se fait la communication et le mode de conception du message lui-même (cf. en particulier Söll 1974, Koch & Österreicher 1990 et 2001).

La recherche dont nous présentons les premiers résultats ici repose sur l'exploration outillée d'un corpus de textes médiévaux. Fondée sur une approche contrastive des données, elle s'articule autour de trois grandes questions de recherche :

- Quel accès pouvons-nous avoir à l'oral et à quelle(s) forme(s) d'oral au Moyen Age ? C'est ce qui nous a amenés à nous intéresser plus spécifiquement à ce qui dans le texte écrit se donne comme de l'oral et qu'on nomme habituellement « oral représenté » ;
- Quelle relation peut-on établir entre le discours direct et l'oral représenté dans les documents médiévaux ?

- Le discours direct présente-t-il une grammaire spécifique ? Comment peut-on l'analyser et quelle exploitation linguistique est-il possible d'en faire ?

Notre recherche répondra de façon encore très partielle à toutes ces questions. Elle proposera surtout une méthodologie empirique qui permette d'aborder ces différents points. Notre présentation se déroulera en deux temps. Après avoir décrit sommairement la façon dont nous avons élaboré un corpus enrichi permettant d'étudier le discours direct de manière contrastive, nous présenterons notre méthodologie d'analyse et les outils que nous avons utilisés. Nous exposerons les premiers résultats tirés de l'exploitation du corpus dans une seconde section.

1. Conduite de l'analyse

1.1. Équipement du corpus : méthodologie et choix d'encodage

Notre recherche s'appuie sur un vaste programme de codage du discours direct à l'intérieur d'un ensemble de textes français composés entre le 9ème siècle et la fin du 15ème siècle. Menée dans le cadre du développement de la Base de Français Médiéval (BFM, <http://bfm.ens-lyon.fr>) et sous l'impulsion du projet *Corpus représentatif des premiers textes français* (CoRPTeF, <http://corptef.ens-lyon.fr>), cette initiative vise à enrichir les niveaux d'annotation linguistique encodés dans ces deux corpus en vue d'une exploitation linguistique outillée et raisonnée.

Le discours direct a été annoté de manière semi-automatique dans chaque texte du corpus (liste jointe en annexe) grâce aux balises <q> et </q>[1] insérées là où les éditions de référence des textes du corpus contenaient des guillemets ouvrants et fermants. Nous avons traité les cas – plus fréquents qu'on aurait pu le penser – où un personnage cite les paroles d'un autre à l'aide d'une hiérarchisation des discours directs (imbrication de balises <q>, qui se traduit par les valeurs 1, 2 et 3 de la propriété lexicale `q` dans les requêtes CQL, *cf.* section 1.2.1).

Nous avons également été amenés à intégrer au discours direct quelques textes dont le genre discursif entretient un rapport particulier avec l'oral représenté : les pièces de théâtre d'une part, quelques textes didactiques d'autre part, sous forme de dialogues pédagogiques (par exemple les *Dialogues du pape Grégoire le Grand*). Les prises de parole représentées dans ces deux catégories de textes ont été encodées à l'aide d'une balise particulière (<sp>). Il est indéniable que l'intégration au discours direct des pièces de théâtre se justifie davantage que celle de ces dialogues sans doute plus « artificiels », mais sur le plan formel il s'agit d'un même mode de structuration de texte (succession de prises de parole précédées d'une indication du locuteur).

---

[1] Toutes les balises utilisées pour cette recherche suivent les standards de la Text Encoding Initiative (voir la documentation de référence en ligne : http://www.tei-c.org/release/doc/tei-p5-doc/fr/html/ref-q.html).

Comme on vient de le voir, notre délimitation du discours direct s'est basée sur des critères purement formels. Les limites de ce balisage doivent rester présentes à l'esprit lors de l'exploitation du corpus, notamment parce que notre parti pris de suivre les guillemets insérés par l'éditeur nous rend tributaires des choix qu'il a effectués. Les cas litigieux restent cependant relativement limités et la démarche adoptée ici a permis de traiter des données en nombre important (1 882 231 mots) et variées.

Une fois balisé, ce corpus a été intégré à la plateforme TXM (version 0.5). Afin de permettre des contrastes entre les différents groupes de textes, cette intégration a pris en compte les métadonnées textuelles qui nous semblaient les plus pertinentes, à savoir la date de composition des œuvres, leur genre discursif, leur domaine[2] et leur forme (vers/prose). Ces métadonnées permettent de construire des sous-corpus et des partitions, regroupements de textes à l'intérieur d'un (sous-) corpus qui conduisent à opposer les résultats obtenus sur les différentes parties du (sous-) corpus.

Enfin, on verra dans la section suivante que notre exploitation du corpus repose en grande partie aussi sur un étiquetage morphosyntaxique (automatique avec correction par des médiévistes) associé à tous les mots des textes (jeu d'étiquette Cattex2009[3]). Certains calculs ont donc été réalisés uniquement sur la sous-partie du corpus qui présente ce type d'informations morphosyntaxiques.

1.2. Instrumentation : le logiciel de textométrie TXM

L'étude est menée dans une perspective contrastive : elle consiste à opposer les fréquences et le fonctionnement d'une série d'items dans différentes catégories de structures textuelles (oral représenté, narration). La textométrie offre un cadre tout à fait approprié pour ces opérations de recherche de motifs linguistiques en corpus et de contrastes évalués par des mesures statistiques adaptées aux données linguistiques. Notre analyse s'appuie sur les outils d'indexation et de statistique implémentés dans la plateforme d'analyse textométrique TXM[4] (Heiden, Magué & *al*. 2010) : calculs de fréquences et de spécificités portant sur des formes graphiques, des propriétés, des constructions, etc., contrôle et interprétation des résultats quantitatifs par des retours ciblés aux contextes d'emploi et aux réalisations effectives.

Nous présentons succinctement ci-après les principales fonctionnalités de TXM mobilisées pour cette étude.

---

[2] Le domaine décrit la visée ou la fonction du texte : édifier pour les textes du domaine religieux, enseigner pour les textes du domaine didactique, etc. Pour une description complète des domaines du projet CoRPTeF et de la Base de Français Médiéval, voir : <http://corptef.ens-lyon.fr/spip.php?rubrique60>.

[3] La liste des étiquettes Cattex2009 est accessible en ligne : http://corptef.ens-lyon.fr/spip.php?rubrique61.

[4] Cette plateforme initiée par le projet ANR Textométrie <http://textometrie.ens-lyon.fr> succède à la plateforme Weblex utilisée précédemment par l'équipe (Pincemin *& al.* 2008).

1.2.1. Langage d'interrogation CQL, pour l'expression de motifs linguistiques

Le moteur de recherche principal de TXM, appelé CQP, permet d'exprimer la recherche de motifs linguistiques complexes au sein d'un corpus à l'aide de requêtes écrites dans un langage d'interrogation formel appelé CQL. Ce langage permet de préciser la forme des motifs cherchés au plan paradigmatique (la morphologie des formes de mots et de toutes les annotations qui peuvent leur être associées dans un corpus donné – étiquette morphosyntaxique, lemme...) et au plan syntagmatique (le type de séquence de mots souhaité).

Par exemple :

- l'expression `[pos="VERcjg" & q="[123]"]` exprime la recherche d'un verbe conjugué (étiqueté « VERcjg » dans le jeu d'étiquettes morphosyntaxiques Cattex2009) qui soit dans le discours direct (marqué par les valeurs 1, 2 ou 3 de l'annotation appelée « q » dans le corpus) ;
- l'expression `[pos="VERcjg"][pos="ADV.*"]{0,2}[pos="VERppa"]` exprime la recherche d'un verbe conjugué, suivi éventuellement d'1 ou 2 adverbes, suivi d'un verbe au participe présent (la propriété `pos` des mots correspondant au jeu d'étiquettes morphosyntaxiques Cattex2009).

L'usage de ce moteur est au cœur des fonctionnalités de base de la plateforme. Son retour – ce qui a été trouvé dans le corpus comme correspondant à une requête donnée (ce que nous appelons aussi « les occurrences d'un motif ») – est utilisé de diverses manières dans TXM : sous forme de liste dans l'outil Index, de tableau comparatif pour les Spécificités ou de liste de contextes dans les Concordances.

1.2.2. Index

L'outil Index liste les différentes réalisations correspondant à un motif CQL au sein du corpus et leur associe une fréquence (un décompte du nombre d'apparitions). Quand la liste d'index est triée par la fréquence, cet outil sert à recenser les réalisations les plus fréquentes d'un motif donné. Cet outil sert également à ajuster une requête CQL par vérification pour faire correspondre au mieux un motif à un besoin de recherche donné.

1.2.3. Spécificités

On se place dans le cas où le corpus est divisé en plusieurs parties. On veut évaluer la sur-utilisation ou la sous-utilisation de mots ou de motifs dans chacune des parties. La fréquence de ces mots ou motifs est un premier indicateur intéressant, mais il est trop sensible à la différence de taille entre parties. Une règle de trois, qui rapporte la fréquence à la taille de la partie, semble être une bonne solution. On montre cependant, en modélisant mathématiquement la répartition des mots entre les différentes parties, qu'on a

un indicateur plus juste avec le calcul des spécificités (Lafon 1981). Il permet de mieux rendre compte des écarts de fréquence (entre mots rares et mots très courants) et des écarts de taille de parties.

Le calcul de la spécificité d'un mot dans une partie repose sur les quatre grandeurs suivantes :

- la fréquence *f* d'apparition du mot au sein de la partie (par exemple la partie formée par tous les mots du discours direct) ;
- la fréquence totale *F* d'apparition du mot dans l'ensemble du corpus ;
- la taille *t* de la partie (nombre total de mots dans la partie) ;
- la taille *T* du corpus complet (nombre total de mots)[5].

Il fournit une probabilité. Un score de spécificité de « +3 » signifie que le mot avait moins d'une chance sur mille (3 → 1 suivi de 3 zéros = 1 000) d'apparaître dans la partie avec une fréquence *f* aussi élevée (score positif → surreprésentation). Autrement dit, si la distribution du mot était aléatoire (selon une loi hypergéométrique), dans 99,9% des cas on aurait observé une fréquence *f* plus faible ; on a donc une fréquence *f* plus élevée de façon statistiquement significative avec un seuil de risque de 0,1%.

Et conventionnellement, un score de spécificité négatif pointe de la même façon une sous-représentation. Ainsi, un score de spécificité de « -6 » signifie que le mot avait moins d'une chance sur un million (6 → 1 suivi de 6 zéros = 1 000 000) d'apparaître dans la partie avec une fréquence *f* aussi faible (score négatif → sous-représentation). En termes de pourcentage, cela revient à dire que si la distribution du mot était aléatoire (selon une loi hypergéométrique), dans 99,9999% des cas on aurait eu une fréquence *f* plus forte, soit une significativité statistique avec un seuil de risque de 0,0001%.

Lorsque le score atteint une valeur telle qu'il dépasse les capacités de la machine, on le représente dans nos tableaux conventionnellement par la valeur *1000*.

1.2.4. Concordance

La concordance est une forme de retour au texte, qui permet d'observer les occurrences d'un motif en contexte. Sa disposition particulière (contexte sur une seule ligne, occurrences du motif alignées verticalement) aide à examiner des régularités du contexte proche. Si l'on a besoin d'un contexte plus large, la concordance donne accès à l'édition du corpus qui met en évidence les occurrences du motif directement au sein du texte.

---

[5] Dans certains cas, selon l'hypothèse linguistique, la taille (de la partie et du corpus) n'est pas mesurée par le nombre total de mots mais par celui des mots d'une certaine catégorie : on évalue alors la présence du motif par rapport à celle de cette catégorie. Par exemple, on évaluera la présence des différentes personnes par rapport aux mots des catégories fléchies selon la personne et non pas par rapport à tous les mots (cf. section 2.2.3).

1.3. Méthodologie de recherche

Pour initier la recherche et dégager les premières tendances, nous étudions le sous-corpus des neuf textes avec étiquetage morphosyntaxique vérifié. Un calcul statistique de contraste (calcul des spécificités) entre les passages de discours direct et le reste, appliqué à l'étiquetage morphosyntaxique, permet de dégager des caractéristiques de ces deux sous-parties du corpus.

Les caractéristiques repérées grâce à ce calcul, qui mobilise la cinquantaine d'étiquettes morphosyntaxiques du jeu Cattex2009, doivent être précisées dans une seconde phase, ce qui peut impliquer de changer le niveau de la description.

En particulier, l'interprétation peut conduire à considérer qu'un phénomène caractéristique du discours direct implique en réalité qu'on **regroupe** plusieurs étiquettes distinctes dans notre jeu : par exemple, le discours direct comporte une forte proportion de marqueurs interrogatifs directs, qu'ils appartiennent à la catégorie des pronoms, des adjectifs ou des adverbes.

A l'inverse, l'analyse peut conduire à examiner le **détail** du contenu d'une catégorie pour distinguer les comportements différents de ses éléments. On peut diffracter la catégorie selon la graphie, ou encore selon le contexte d'emploi (construction dans laquelle entre l'élément). Par exemple, on peut nuancer le score global de la catégorie infinitif en observant des comportements différents selon que l'infinitif entre dans telle ou telle construction.

Enfin regroupement et diffraction peuvent être combinés pour reconstruire les catégories linguistiquement les plus pertinentes pour l'analyse. Dans cet article, c'est le cas pour l'étude des personnes, qui implique de regrouper les différentes catégories concernées par la flexion en personne, puis, à partir du détail des graphies, de former de nouvelles catégories d'analyse correspondant aux différentes personnes.

Il est donc important de comprendre que le choix d'un point d'entrée de l'analyse, comme le niveau morphosyntaxique par exemple, permet de (et même doit) ouvrir sur d'autres niveaux d'analyse, en fonction de l'interprétation linguistique des observations.

De même que nous sommes conduits à varier le palier de description linguistique, nous avons à jouer avec l'échelle du corpus. Lorsqu'une tendance est dégagée dans le sous-corpus à haute qualité d'étiquetage linguistique, nous essayons, dans la mesure du possible, de confronter au corpus complet l'hypothèse formulée à partir de ce premier niveau d'analyse. Dans cette seconde phase, la recherche ne peut plus se baser de la même manière sur l'étiquetage morphosyntaxique. Elle exploite donc les formes graphiques et est limitée par les cas d'ambiguïté et d'homographie. Les résultats présentés ci-dessous montrent que beaucoup d'investigations peuvent cependant être menées de cette manière.

2. Résultats

La présentation de nos résultats suit la procédure de recherche que nous venons de décrire. Une analyse préliminaire des spécificités générales du discours direct (désormais DD) puis de tout ce qui est extérieur au discours direct (désormais non DD)[6] est exposée dans la section 2.1, et la section 2.2 détaille les résultats obtenus grâce à des recherches complémentaires portant sur des catégories et / ou sur des formes plus spécifiques.

2.1. Aperçu général

L'analyse contrastive des caractéristiques du DD et du non DD a tout d'abord été menée grâce au calcul des spécificités. Ce calcul a été réalisé à partir des étiquettes morphosyntaxiques et a été limité, pour cette raison, aux textes dont l'étiquetage avait été vérifié par des spécialistes de français médiéval.

2.1.1. Etiquettes spécifiques du DD, première interprétation et pistes d'analyse

Pos[7]	F totale	f dans le non DD	Spécif. pour le non DD	f dans le DD	Spécif. pour le DD
PONpxx	2432	4	*-1000*	2428	*1000*
PONpga	1091	83	-302,32	1008	*1000*
PONpdr	1085	84	-298,81	1001	*1000*
PROper	30414	16029	-195,37	14385	*1000*
PONfrt	13086	6945	-73,67	6141	*1000*
PROint	270	24	-71,1	246	*1000*
ADVneg	7141	3617	-68,07	3524	*1000*
INJ	232	30	-50,89	202	*1000*
NA	82	0	-33,34	82	*1000*
ADVint	106	11	-26,52	95	*1000*
ADJpos	225	70	-18,66	155	*1000*
PROper.PROper	55	3	-17,35	52	*1000*
PONfbl	28023	16446	-13,26	11577	13,26

[6] Il est certain que le regroupement dans une catégorie unique de tout ce qui est extérieur au discours direct fait de cet ensemble une catégorie très hétérogène. Le terme de non DD a été choisi par commodité, parce qu'il permet de nommer de manière univoque cet ensemble. Il ne préjuge en rien de la nature et de l'unité de son contenu. Il permet avant tout d'opérer un contraste avec le discours direct, dont la grammaire spécifique est l'objet plus particulier de notre recherche.

[7] pos : abréviation courante de « part-of-speech », et ici nom choisi au moment de l'intégration du corpus dans TXM pour l'étiquette morphosyntaxique vérifiée issue du jeu Cattex2009.

DETint	59	8	-13,07	51	13,07
PROimp	1056	526	-12,5	530	12,5
DETpos	6916	3927	-11,35	2989	11,35
PROpos	132	41	-11,32	91	11,32
DETdem	2120	1134	-11,31	986	11,31
VERinf	8321	4825	-7,1	3496	7,1
ADVneg.PROper	315	148	-6,3	167	6,3
CONsub	11211	6576	-5,77	4635	5,77
PROdem	4117	2394	-3,54	1723	3,54

Tableau 1

Résultat du calcul de spécificités sur le corpus des neuf textes à étiquetage morphosyntaxique vérifié, pour la partition opposant le DD (q=1, 2 ou 3) au reste (q=0), sur la propriété pos correspondant à l'étiquetage morphosyntaxique vérifié, trié par score de spécificité décroissant pour le DD et seuillé à la valeur 3.

Le tableau 1, qui ne rend compte que des étiquettes les plus spécifiques au DD, permet de dégager dès à présent les domaines dans lesquels le DD se distingue le plus du non DD à l'intérieur de notre corpus restreint. Un fort score de spécificité positive marque un emploi anormalement élevé pour le DD (au sens statistique), mais ce n'est pas nécessairement un emploi exclusif : à ce titre, il est toujours utile de consulter aussi les décomptes de fréquence.

Parmi les points les plus remarquables mis en évidence par le calcul des spécificités, on peut noter le score très élevé des **pronoms**, notamment des pronoms personnels (PROper, PROper.PROper), mais aussi des pronoms interrogatifs (PROint), impersonnels (PROimp), possessifs (PROpos) et démonstratifs (PROdem). Ce résultat nous semble devoir être mis en lien avec la surreprésentation du nom et des éléments constitutifs du syntagme nominal pour le non DD (cf. la section suivante, 2.1.2) : le pronom et le syntagme nominal sont en relation paradigmatique, et il semble que le DD opte de préférence pour le pronom alors que le non DD recourt davantage au syntagme nominal.

Les pronoms personnels peuvent être aussi rapprochés d'autres catégories sensibles aux marques de personne, à savoir les **possessifs** (ADJpos, DETpos, PROpos). En considérant les marques de personne dans les pronoms personnels et les possessifs réunis, nous pouvons étudier plus en détail la répartition des différentes personnes entre DD et non DD (cf. section 2.2.2).

Un autre élément très remarquable de la liste est la spécificité des **adverbes de négation** : adverbes de négation proprement dits (*ne*, *non*, *pas*, *mie*, *point*) et contractions d'un adverbe de négation et d'un pronom personnel (ADVneg.PROper, par exemple *nel* pour *ne le*). Un calcul de spécificité centré sur les négations et effectué sur le détail des graphies montre que les diverses formes (particules négatives, forclusifs) se comportent de façon globalement équivalente vis-à-vis de l'opposition DD / non DD. Les

contrastes semblent jouer sur d'autres dimensions : ainsi, avec des spécificités sur le dialecte, on peut confirmer l'analyse de la grammaire de Buridant (2007, p.713 §609) : *mie* est surreprésenté dans l'Est (dialectes lorrain, picard, wallon : spécificité maximale (*1000*)) et sous-représenté en anglo-normand (-77).

La surreprésentation des **verbes à l'infinitif** (VERinf) n'a pas non plus d'explication évidente et mérite une investigation plus poussée, en essayant de préciser quels usages et quelles constructions amènent ce suremploi (cf. section 2.2.3).

Pour d'autres catégories, le calcul confirme les attentes, si bien que les résultats sont moins remarquables, mais non sans valeur (ils rassurent sur la pertinence de la méthodologie, ils permettent de dresser un descriptif relativement complet). Les scores de spécificité les plus élevés correspondent à des catégories spécifiques par construction : les **ponctuations** PONpxx (les guillemets non orientés), PONpga (ponctuation gauche ouvrante, majoritairement des guillemets à 90%), PONpdr (ponctuation droite fermante, majoritairement des guillemets à 90%). Les autres catégories de ponctuation, plus hétérogènes, doivent être analysées dans le détail des graphies. On trouve sans surprise que les ponctuations « émotives » (point d'exclamation, point d'interrogation) sont spécifiques du DD. Plus difficile à interpréter, les points et les virgules sont également fortement surreprésentés dans le DD. Mais la ponctuation observée ici se fonde sur les pratiques des éditeurs de textes et ne reflète que très partiellement les marques et divisions du document primaire. Ce paramètre sera naturellement à prendre en compte dans nos recherches ultérieures.

La surreprésentation des **interrogatifs directs** n'est pas surprenante dans le DD. Ils apparaissent de façon dispersée dans les catégories PROint, ADVint, DETint. Il est possible de mesurer la spécificité d'une catégorie « interrogatifs » qui les regroupe : son score dépasse effectivement le plafond.

Peu surprenant aussi, les **interjections** (INJ) sont globalement spécifiques du DD. Nous observons cependant qu'elles sont présentes aussi dans le non DD. L'analyse du détail des formes montre que 27 interjections différentes sont représentées dans l'ensemble du corpus annoté. Une recherche ciblée sur un sous-ensemble de 19 formes non ambigües et appliquée au corpus intégral montre que deux d'entre elles, *He* et *ha*, sont spécifiques du non DD : un retour au texte permet de voir qu'elles ont notamment un usage rhétorique d'apostrophe au lecteur dans un texte argumentatif comme les *Miracles* de Gautier de Coinci.

Le score des **démonstratifs** (DETdem, PROdem) invite à considérer de façon plus générale les déictiques, pour lesquels un petit sondage a pu être fait en 2.2.1.

On peut noter en dernier lieu l'excédent de **conjonctions de subordination** (CONsub), que nous n'analysons pas plus avant ici.

NA correspond à un artefact dans le codage du corpus (mots ou ponctuations qui n'avaient pas reçu d'étiquette) et n'a pas à être interprété linguistiquement.

2.1.2. Etiquettes spécifiques du non DD, première interprétation et pistes d□analyse

pos	F totale	f dans le non DD	Spécif. pour le non DD	f dans le DD	Spécif. pour le DD
DETdef	16314	11861	*1000*	4453	-233,60
NOMcom	49225	31920	*1000*	17305	-88,44
CONcoo	18888	12715	*1000*	6173	-80,98
PRE	28687	18630	*1000*	10057	-51,39
DETndf	2531	1888	*1000*	643	-48,42
PRE.DETdef	4233	2996	*1000*	1237	-41,97
DETcom	166	163	*1000*	3	-30,59
VERppa	1017	773	*1000*	244	-24,29
DETcar	1187	870	*1000*	317	-19,12
PROind	3439	2326	*1000*	1113	-16,50
VERppe	10065	6505	15,48	3560	-15,45
NOMpro	8541	5544	15,05	2997	-15,06
ADJcar	348	273	11,81	75	-11,81
OUT	304	239	10,62	65	-10,62
PRE.DETcom	62	60	10,49	2	-10,49
PROcar	473	331	4,72	142	-4,72
ADJqua	11104	6957	4,71	4147	-4,71
ADJord	104	83	4,57	21	-4,57
ADVsub	370	257	3,5	113	-3,50

Tableau 2

Résultat du calcul de spécificités sur le corpus des neuf textes à étiquetage morphosyntaxique vérifié, pour la partition opposant le DD (q=1, 2 ou 3) au reste (q=0), sur la propriété pos correspondant à l'étiquetage morphosyntaxique vérifié, trié par score de spécificité décroissant pour le non DD et seuillé à la valeur 3.

A nouveau, rappelons qu'un fort score de spécificité n'implique pas que l'usage de la catégorie soit réservé au non DD, mais simplement que son emploi est plus « dense » dans le non DD relativement au DD. Complémentairement, les fréquences permettent de savoir s'il s'agit d'un trait caractéristique

(quasi)exclusif au non DD, ou simplement d'un déséquilibre entre un suremploi pour le non DD et un sous-emploi pour le DD.

La caractéristique du non DD semble être d'abord l'usage du nom, et plus généralement du syntagme nominal, avec la surreprésentation des étiquettes des **noms propre** et **commun** (NOMcom, NOMpro), des **déterminants** (DETdef, DETndf, PRE.DETdef, DETcom, DETcar, ADJcar, PRE.DETcom, ADJord), et de l'**adjectif qualificatif** (ADJqua). Ainsi que nous l'avons noté dans l'analyse des caractéristiques du DD, syntagme nominal et pronom sont en alternative paradigmatique, et le non DD opte pour le premier alors que le DD privilégie le second. Peut-être aussi l'expression du sujet est-elle moins fréquente dans le non DD : c'est une hypothèse qu'il faudrait examiner sur un corpus annoté en syntaxe. Le calcul montre cependant que deux types de pronoms font exception : pronoms indéfinis (PROind) et pronoms cardinaux (PROcar).

Les **conjonctions de coordination** (CONcoo) sont mises en avant comme caractéristiques du non DD, mais leur très fort score suscite l'étonnement et mérite un examen un peu plus approfondi, dans le détail des différentes conjonctions. En fait c'est le ET[8] qui est spécifique au non DD, et le DD semble avoir de son côté des affinités avec CAR, MES, NE (mais leur distribution semble aussi dépendre de l'auteur).

Les verbes se signalent par un excédent de **participes**, présents et passés (VERppa, VERppe). Pour mieux comprendre le phénomène linguistique sous-jacent, nous pouvons examiner plus précisément certaines constructions dans lesquelles ces participes entrent en composition (temps composés, périphrases verbales) (cf. sections 2.2.4 et 2.2.5).

Là encore, le temps et l'espace nous manquent pour examiner et interpréter la spécificité de deux dernières catégories sélectionnées par le calcul : les **prépositions** (PRE, PRE.DETdef, PRE.DETcom), qu'il faudrait analyser en détail dans leurs contextes, et les **adverbes de subordination** (ADVsub).

OUT correspond à un artefact dans le codage du corpus (mots ou ponctuations qui ne doivent pas être pris en compte par l'étiqueteur) et n'a pas à être interprété linguistiquement.

2.2. Investigations à partir de pistes pointées par le calcul de spécificités initial

Cette section présente catégorie par catégorie les investigations permettant de préciser quelques-unes des grandes tendances dégagées ci-dessus. Ces recherches complémentaires ont été menées tantôt sur le corpus restreint (chaque fois que les étiquettes morphosyntaxiques ont été mobilisées), tantôt sur le corpus total.

---

[8] La graphie *et* a un score maximal, et la graphie *Et* a un score de spécificité de +3.

2.2.1. Les déictiques

Pour étudier le comportement d'un certain nombre d'expressions déictiques (sans prétendre à l'exhaustivité mais en essayant de se focaliser sur les formes non ambigües), nous avons recherché la spécificité générale (parmi l'ensemble des mots) des graphies correspondant à l'équation CQL[9] suivante :

```
[word="[i]?c[iy]|ça|ill?u[eo][cq]u?e?s?|h?u[iy]"%c]
```
[10]

Les formes ainsi sélectionnées se répartissent sans reste en deux groupes : soit elles sont spécifiques des passages balisés <q> (et quelquefois aussi <sp>, mais plus rarement et souvent moins fortement) ; soit leur distribution est neutre, équilibrée entre les passages au DD et les passages extérieurs au DD. Voici le détail des résultats :

Word	F totale	f dans les passages <q>	Spécif. pour les passages <q>	f dans les passages <sp>	Spécif. pour les passages <sp>	f dans le non DD	Spécif pour le non DD
ça	271	163	10	16	-0	92	-32
Ça	19	10	2	0	-0	9	-1
ci + cy	1072	627	10	84	2	361	-128
Ci + CI + Cy	189	54	0	3	-2	132	-0
hui + huy + ui + uy	223	133	10	42	10	48	-48
Hui + Ui	22	10	1	6	2	6	-4
ici + icy	280	192	10	19	0	69	-53
Ici + Icy + ICY	93	12	-2	5	0	76	2
iluec + ilueques + illuec + illueques	207	34	-2	18	1	155	1
Iluec + Illuec + Illueques + Iluoc	45	7	-0	1	-0	37	1

Tableau 3

Résultat du calcul de spécificités sur le corpus total des 57 textes, pour la partition contrastant (i) les passages avec du DD balisé <q> (q=1, 2 ou 3), (ii) les passages avec du DD balisé <sp> (sp=t), et (iii) le reste, représentant le non DD (q=0 et sp=f). Le résultat est focalisé sur les lignes sélectionnées par l'équation CQL [word="[i]?c[iy]|ça|ill?u[eo][cq]u?e?s?|h?u[iy]"%c]. Le calcul est fait sur les graphies (propriété word), après regroupement des variantes d'écriture.

On constate que les formes neutres concentrent les graphies avec majuscule initiale. L'une des explications possibles à ce phénomène pourrait être que ces éléments jouent en rôle important dans la structuration des unités narratives, d'où leur affinité avec le début d'une unité syntaxique et discursive

---

9 Dans la suite de l'article, nous utiliserons de façon synonyme les termes d'*équation* et d'*expression CQL*.

10 Word : nom donné au moment de l'intégration du corpus dans TXM à la forme graphique des mots.

(cf. en particulier les analyses de Perret 1988 sur *ci*, Guillot 2004 sur les démonstratifs et 2009 sur *or(e)*). Mais aussi il s'agit majoritairement de formes relativement peu fréquentes, pour lesquelles mathématiquement le score statistique ne peut pas monter très haut.On remarque également la bonne représentation du paradigme de ILLUEC dans le non DD. Il apparaît donc très clairement que le DD ne détient pas le monopole de la deixis.

2.2.2. Les personnes

En considérant toutes les formes de pronoms personnels (PROper, PROper.PROper, ADVneg.PROper) et de possessifs (ADJpos, DETpos, PROpos), nous groupons les formes par personne et calculons une spécificité limitée à ces formes. Autrement dit, nous étudions comment se répartissent les personnes entre DD et non DD, à cela près que nous ne prenons pas en compte les personnes des formes verbales sans nom ou pronom sujet, car nous n'avons pas l'information sur la personne dans l'étiquetage des verbes.

Sur le corpus à étiquetage morphosyntaxique vérifié, nous lançons la requête : [pos=".*PROper.*|.*pos"], qui sélectionne 252 graphies différentes (pour 38198 occurrences).

Dans la table lexicale, nous fusionnons les lignes selon les personnes, en nous aidant de filtres : sélection des formes de la première personne du singulier par les équations : [jJgGmM].*, puis .*m ; puis sélection des formes de la deuxième personne du singulier par l'équation : [tT].*. La première personne du pluriel est obtenue par : [nN][ou].*[^l], la deuxième personne du pluriel par : [vV].*, la troisième personne (sans distinguer singulier et pluriel, d'autant qu'un certain nombre de formes sont homographes) correspond au reste (à l'exception de deux erreurs d'étiquetage, qu'on retire).

Le tableau suivant synthétise les résultats obtenus une fois ces regroupements pris en compte :

Personne	F totale	f dans le non DD	Spécif. pour le non DD / pronoms personnels et possessifs	f dans le DD	Spécif. pour le DD / pronoms personnels et possessifs
1s	6015	502	*-1000*	5513	*1000*
2s	1265	17	*-1000*	1248	*1000*
1p	1259	243	-138,91	1016	*1000*
2p	3804	156	*-1000*	3648	*1000*
3	25853	19390	*1000*	6463	*-1000*

Total		20308		17888	

Tableau 4

Résultat du calcul de spécificités sur le corpus des pronoms personnels et des possessifs des neuf textes à étiquetage morphosyntaxique vérifié, pour la partition opposant le DD (q=1, 2 ou 3) au reste (q=0). Un index a permis de relever toutes les graphies des pronoms personnels et des possessifs, et une table lexicale a permis de fusionner les occurrences des graphies correspondant à la même personne.

Comme attendu, les premières et deuxièmes personnes apparaissent comme très fortement caractéristiques du DD. La deuxième personne est presque absente du non DD, mais la première personne n'en est pas aussi massivement exclue[11]. Par un effet de balancement, le calcul fait apparaître la troisième personne comme étant caractéristique du non DD (bien que le détail des fréquences montre qu'elle est très utilisée aussi dans le DD).

2.2.3. Les infinitifs

Nous avons voulu préciser les constructions dans lesquelles entrent les infinitifs dans le DD et le non DD pour affiner les premières tendances observées au niveau de l'ensemble de la catégorie (l'infinitif est globalement surreprésenté dans le DD). Nous avons pour cela observé, parmi toutes les occurrences des infinitifs présents dans le corpus annoté, le comportement des constructions à verbes modaux, limités à *pooir*, *voloir* et *devoir*, et des périphrases verbales en « aller + infinitif ».

Nous avons donc calculé les spécificités dans le DD et le non DD des infinitifs entrant dans les constructions « verbe modal + infinitif », en prenant soin de prendre en compte les occurrences dans lesquelles un ou plusieurs éléments s'intercalent entre le verbe conjugué et l'infinitif[12]. Nous avons procédé de la même manière pour les infinitifs impliqués dans une périphrase « aller + infinitif »[13], et de façon générale pour la construction « verbe + infinitif »[14]. L'expression de requête permettant de repérer ces constructions du type « X (qui correspond aux formes fléchies du verbe captées grâce aux expressions données en notes) + insertion éventuelle + infinitif » est la suivante :

```
[X]([pos="DET.*"]?[pos="NOM.*"]|[pos="ADV.*|PRO.*"]){0,2}
[pos="VERinf" & q="0"]
```

---

[11] On peut se demander dans quelle mesure le fait qu'on ne se base que sur l'expression des pronoms personnels, et non sur les personnes verbales, introduit un biais dans les résultats qu'on obtient. Il faudrait naturellement compléter cette recherche en tenant compte de ces informations.

[12] Les expressions permettant de capter ces verbes modaux dans le corpus sont détaillées ici :

- pour *pooir* : [pos="VERcjg" & word="p[eouëü].*"]
- pour *voloir* : [pos="VERcjg" & word="vu(e|il|l).*|vieu?l?t|veil.*|veu.*|vel[stz]?|vo(l|r|il|e).*"] (il manque certaines formes en vau[rl].* comme *vaura*)
- pour *devoir* : [pos="VERcjg" & word="d.*" & word!=".*gn.*|di.*"] (l'équation capte quelques formes intruses, en nombre négligeable).

[13] Les occurrences de *aller* ont été repérées grâce à l'équation suivante : [pos="VERcjg" & word="ai?ll?.*|[iy]r[aeiou].*|va[^ul]*|vet|v[ou]nt"%c]. On constate qu'on manque quelques occurrences des formes ambigües : *vois, voit, veit, veis*.

[14] La formulation CQL est alors simplement [pos="VERcjg"].

(ou idem avec à la fin `q="[123]"` pour extraire les occurrences présentes dans le DD). Les résultats du calcul des spécificités[15] sont présentés dans le tableau 7 :

	F totale	f dans le non DD	Spécif pour le non DD	f dans le DD	Spécif pour le DD
Pooir + infinitif	1240	628	-13	612	13
Devoir + infinitif	521	179	-34	342	34
Voloir + infinitif	417	200	-8	217	8
Aller + infinitif	223	163	5	60	-8
Vb + infinitif	4267	2288	-22	1979	22
Taille	T=364967	t=221814		t=143153	

Tableau 5

Résultat du calcul de spécificités sur le corpus des infinitifs des neuf textes à étiquetage morphosyntaxique vérifié, pour la partition opposant le DD (q=1, 2 ou 3) au reste (q=0). Un index a permis de compter les occurrences de chaque construction et de l'infinitif seul, dans le corpus entier et dans chaque type de passage (DD ou non DD). On obtient ainsi les quatre paramètres (T,t,F,f) permettant le calcul de chaque spécificité.

Dans les limites de notre corpus de neuf textes, l'infinitif est clairement employé de façon spécifique dans le DD dans les constructions complexes « verbe modal + infinitif », tout particulièrement avec le verbe *devoir*. A l'inverse, la construction « aller + infinitif » se révèle être spécifique au non DD. Ces premiers résultats demanderaient à être précisés, en tenant compte surtout de l'évolution diachronique : il est certain que la périphrase « aller + infinitif », par exemple, connaît une évolution sémantique très importante entre la période de début (9ème siècle) et la période de fin (fin 15ème) de notre corpus.

2.2.4. Les participes passés

Le participe peut être utilisé « seul » ou faire partie d'une forme verbale complexe. C'est ce dernier emploi qui a retenu notre attention. Nous avons donc voulu examiner la fréquence de la succession « verbe conjugué + éventuellement une insertion + participe passé » dans le DD et hors du DD[16]. On constate que dans la quasi-totalité des cas, les verbes conjugués qui ont été sélectionnés sont les auxiliaires *être* et *avoir*. Les fréquences de cette construction dans nos deux sous-parties a permis la

---

[15] La version courante de TXM (0.5) ne permet pas de lancer directement ces calculs de spécificité (portant sur des constructions sur plusieurs mots). Nous avons donc utilisé TXM pour obtenir les valeurs pour chacun des paramètres *f*, *F*, *t*, *T* (cf. tableau 7), puis nous avons fait appel à Weblex (http://weblex.ens-lsh.fr/wlx/) pour effectuer le calcul du score de spécificité à partir des quatre paramètres.

[16] La requête utilisée pour cette extraction est la suivante :

[pos="VERcjg"](([pos="DET.*"]?[pos="NOM.*"])|[pos="ADV.*"]|[pos="PRO.*" & pos!="PROrel"]){0,2}[pos="VERppe"]

réalisation du calcul de spécificité[17]. Notons que pour ce calcul nous avons pris comme base pour la mesure de la taille des parties et du corpus le nombre de verbes conjugués : en effet, linguistiquement, c'est sur les verbes conjugués seuls que se marque le choix de cette construction particulière. Les résultats de ce calcul montrent que la succession « verbe conjugué + éventuellement une insertion + participe passé » est caractéristique du non DD (spécificité de +8). Il importe de garder à l'esprit que cette requête capte en réalité deux types de formes verbales : les temps composés d'une part, les formes passives d'autre part. Une recherche plus poussée, que permettrait une annotation syntaxique, serait utile pour préciser la part de chacune de ces formes verbales à l'intérieur et à l'extérieur du discours direct[18].

2.2.5. Les participes présents

Le participe présent est globalement suremployé dans le non DD. Il est intéressant néanmoins de pouvoir nuancer ce résultat en regardant dans le détail certaines constructions importantes, pour voir si leur comportement est homogène.

On a d'abord examiné le comportement de la construction « verbe conjugué + adverbe éventuel + participe présent » (dans les expressions du type *il va/est querant*)[19], en calculant sa spécificité[20] dans les

---

[17] La version courante de TXM (0.5) ne permet pas de lancer directement ce calcul de spécificité (portant sur une construction sur plusieurs mots et effectué relativement aux seuls verbes conjugués). Nous avons donc utilisé TXM pour obtenir les valeurs pour chacun des paramètres *f*, *F*, *t*, *T*, puis nous avons fait appel à Weblex (http://weblex.ens-lsh.fr/wlx/) pour effectuer le calcul du score de spécificité à partir des quatre paramètres.

Détail des paramètres du calcul de spécificité :

T = [pos="VERcjg"] = 48494

t = [pos="VERcjg" & q="[0]"] = 29337

F = [pos="VERcjg"](([pos="DET.*"]?[pos="NOM.*"])|[pos="ADV.*"]|[pos="PRO.*" & pos!="PROrel"]){0,2}[pos="VERppe"] = 7105

f = [pos="VERcjg" & q="[0]"](([pos="DET.*"]?[pos="NOM.*"])|[pos="ADV.*"]|[pos="PRO.*" & pos!="PROrel"]){0,2}[pos="VERppe" & q="[0]"] = 4529

Spécificité de +10 pour les formes verbales composées à participe passé dans le non DD.

[18] Glikman & Mazziotta trouvent ainsi que les temps composés sont spécifiques du DD. Mais ils ne travaillent que sur un seul texte, la *Queste del saint Graal*. Nous avons donc examiné si, selon notre méthodologie d'analyse, ce texte fait exception ou bien se comporte de façon comparable aux autres textes de notre petit corpus :

T = [pos="VERcjg"] = 16819

t = [pos="VERcjg" & q="[123]"] = 7972

F = [pos="VERcjg"](([pos="DET.*"]?[pos="NOM.*"])|[pos="ADV.*"]|[pos="PRO.*" & pos!="PROrel"]){0,2}[pos="VERppe"] = 2453

f = [pos="VERcjg" & q="[123]"](([pos="DET.*"]?[pos="NOM.*"])|[pos="ADV.*"]|[pos="PRO.*" & pos!="PROrel"]){0,2}[pos="VERppe" & q="[123]"] = 1179

La spécificité est de l'ordre de 1, donc non significative : dans la *Queste del saint Graal*, la construction « verbe + insertion éventuelle + participe passé » se distribue de façon équivalente entre le DD et le non DD. Aurement dit, la spécificité de cette construction n'est plus marquée : la *Queste del saint Graal* montre donc un usage particulier de cette construction par rapport aux huit autres textes du petit corpus.

[19] L'équation ayant servi à la requête est la suivante :

[pos="VERcjg"][pos="ADV.*"]{0,2}[pos="VERppa"]

[20] La version courante de TXM (0.5) ne permet pas de lancer directement les calculs de spécificité de cette section (portant sur des constructions sur plusieurs mots). Nous avons donc utilisé TXM pour obtenir les valeurs pour chacun des paramètres *f*, *F*, *t*, *T*, puis nous avons fait appel à Weblex (http://weblex.ens-lsh.fr/wlx/) pour effectuer le calcul du score de spécificité à partir des quatre paramètres.

deux sous-parties relativement à l'ensemble des mots employés. Les résultats du calcul révèlent une légère spécificité (+3) de cette construction dans le non DD[21].

Une analyse comparable de la fréquence de la séquence « en + participe présent » montre qu'elle se rencontre elle aussi surtout hors du DD (forte spécificité de +9)[22].

La comparaison des fréquences des participes présents n'entrant pas dans ces deux constructions fait apparaître une spécificité encore plus marquée (+18) de ces participes dans le non DD[23].

Lors de cette première exploration nous n'avons donc pas trouvé de construction particulière pour laquelle le participe présent serait suremployé dans le DD.

Conclusion

A l'issue de ces quelques observations et investigations statistiques, il semble que la méthode que nous avons adoptée pour la constitution et l'exploitation d'un corpus pour l'étude du DD en français médiéval est capable d'apporter des résultats intéressants. Nous avons pu rassembler et annoter en l'espace de quelques mois un corpus relativement étendu et varié de textes médiévaux et, grâce à l'usage de modèles statistiques appropriés, nous avons pu amorcer une analyse linguistique innovante, à la fois informée des connaissances traditionnelles en français médiéval et directement basée sur l'observation des textes en corpus.

Il est possible de dégager plusieurs tendances propres au DD et au non DD dans notre corpus. Certaines de ces tendances confirment nos attentes : les ponctuations émotives, les interrogatifs directs, les personnes 1 et 2 sont caractéristiques du discours direct, la personne 3 est surreprésentée dans le non DD. D'autres suscitent davantage l'étonnement : les formes verbales composées d'un auxiliaire et d'un participe sont plutôt utilisées en dehors du DD, de même que « aller + infinitif ». Dans quelques cas enfin, les tendances observées, sans être surprenantes, sont dignes de retenir notre attention : le DD

---

21 Paramètres du calcul pour la spécificité de la construction « verbe conjugué + adverbe éventuel + participe présent » :
T = [] = 364967
t = [q="0"] = 221814
F = [pos="VERcjg"][pos="ADV.*"]{0,2}[pos="VERppa"] = 231
f = [pos="VERcjg"][pos="ADV.*"]{0,2}[pos="VERppa" & q="0"] = 159

22 Paramètres du calcul pour la spécificité du gérondif :
T = [] = 364967
t = [q="0"] = 221814
F = [pos="PRE" & word="en"%c] [pos="VERppa"] = 146
f = [pos="PRE" & word="en"%c] [pos="VERppa" & q="0"] = 122

23 Paramètres du calcul pour la spécificité des autres participes présents :
[pos="VERppa"] = 1017
[pos="VERppa" & q="0"] = 773
T = [] = 364967
t = [q="0"] = 221814
F = 1017 - (231 + 146) = 640
f = 773 - (159 + 122) = 492

semble privilégier les pronoms, le non DD le syntagme nominal, les infinitifs régimes des verbes modaux sont surreprésentés dans le DD, et à l'inverse, les participes présents sont caractéristiques du non DD, y compris lorsqu'ils entrent dans des périphrases verbales ou sont utilisés après *en*. Notre démarche empirique et outillée a permis de mettre en évidence des phénomènes qu'on ne peut pas expliquer facilement, et que l'on n'aurait sans doute pas remarqués si l'on avait commencé par tester des hypothèses linguistiques conçues *a priori*.

L'un des principaux apports méthodologiques de ces quelques sondages est aussi d'avoir montré la nécessité de fouiller l'analyse au-delà des grandes catégories utilisées au départ. La mise au jour de tendances générales masque parfois des situations complexes et contrastées : les déictiques sont globalement très présents dans le DD mais certains d'entre eux sont plus spécifiques au DD alors que d'autres ne sont pas sensibles à l'opposition DD vs non DD ; la situation des conjonctions de coordination est comparable (mais inverse : formes dominantes spécifiques au non DD, et existence de formes spécifiques au DD). Dans bien des cas, il est très probable que d'autres facteurs que l'opposition DD / non DD interfèrent sur le comportement des unités linguistiques : la négation semble caractéristique du DD mais ce sont parfois des choix dialectaux qui rendent compte de la fréquence de certaines formes dans les textes.

L'analyse pourra se poursuivre d'une part en étudiant plus finement certains phénomènes pointés dans cette première étude et qu'il n'était pas possible de développer ici, d'autre part en prenant en compte des dimensions de variation textuelle comme la variation générique : l'opposition DD vs non DD pourrait alors être nuancée selon les rapports du genre à l'oralité. Par ailleurs, les avancées sur le corpus (extension de l'annotation morphosyntaxique vérifiée, annotation syntaxique, ajout de nouveaux textes) permettront également de renouveler et de consolider les premières observations.

L'annotation du DD utilisée dans cet article sera prochainement disponible aux utilisateurs de la BFM pour réaliser des opérations et calculs similaires sur les textes actuellement diffusés dans la Base.

Bibliographie sommaire

Buridant, C. 2007. *Grammaire nouvelle de l'ancien français.* Paris : SEDES.

Cerquiglini, B. 1981. *La parole médiévale : discours, syntaxe, texte.* Paris : Editions de minuit. Coll. Propositions.

Glikman, J. & Mazziotta, N. (infra) « Représentation de l'oral et syntaxe dans la prose du Queste del saint Graal (1225-1230) », *RSL V*.

Guillot, C. 2004. « Ceste parole et ceste aventure dans la Queste del saint Graal, marques de structuration discursive et transitions narratives ». *L'Information grammaticale*, 103 : 29-36.

Guillot, C. 2009. « Ecrit médiéval et traces d'oralité : l'exemple de l'adverbe or(e) ». In : E. Havu & al. (éd.), La langue en contexte. *Actes du colloque Représentation du sens linguistique IV* (Helsinki, 28-30 mai 2008), Helsinki : Société Néophilologique : 267-281.

Guillot, C. & Lavrentiev, A. 2009. *Manuel de description des textes pour la Base de Français Médiéval*, version 2.3. Lyon : Projet BFM, http://ccfm.ens-lyon.fr/IMG/pdf/Manuel_Descripteurs_BFM.pdf.

Heiden, S., Guillot, C., Lavrentiev A. & Bertrand, L. 2010. *Manuel d'encodage XML-TEI des textes de la Base de Français Médiéval*, version 4.0. Lyon : Projet BFM, http://bfm.ens-lyon.fr/IMG/pdf/Manuel_Encodage_TEI.pdf.

Heiden, S., Magué, J-P. & Pincemin, B. 2010. « TXM : Une plateforme logicielle open-source pour la textométrie – conception et développement ». In : S. Bolasco & al. (éd.), *Statistical Analysis of Textual Data - Proceedings of 10th International Conference Journées d'Analyse statistique des Données Textuelles - JADT 2010*, Rome : Edizioni Universitarie di Lettere Economia Diritto : 1021-1032.

Jucker, A., Fritz G. & Lebsanft, F. 1999. *Historical Dialogue Analysis*, Amsterdam / New York : John Benjamins Publishing Company.

Koch, P. 1993. « Pour une typologie conceptionnelle et médiale des plus anciens documents/monuments des langues romanes ». In : M. Selig & al. (éd.), *Le passage à l'écrit des langues romanes*. Tübingen : Narr : 39-81.

Koch, P. & Österreicher, W. 1990. *Gesprochene Sprache in der Romania : Französisch, Italienisch, Spanisch*. Tübingen : Niemeyer.

Koch, P. & Österreicher, W. 2001. « Gesprochene Sprache und geschriebene Sprache. Langage parlé et langage écrit ». In : G. Holtus & al. (éd.), *Lexikon der romanistischen Linguistik*. Tübingen : Niemeyer : 584-627.

Lafon, P. 1980. « Sur la variabilité de la fréquence des formes dans un corpus », *M.O.T.S* 1 : 127-165.

Llamas Pombo, E. 1996. « Écriture et oralité : ponctuation, interprétation et lecture des manuscrits français de textes en vers (XIII^{e}-XV^{e} s.) ». In : E. Aloinsi & al. (éd.), *La linguistique française : grammaire, histoire et épistémologie*. Seville : Grupo Analuz de pragmática 133-144.

Llamas Pombo, E. 2010. « Marques graphiques du discours rapporté (Manuscrits du Roman de la Rose, XV^{e} siècle) ». In : B. Combettes & al. (éd.), *Le changement en français : Études de linguistique diachronique. Actes du colloque international DIACHRO-IV* (22-24 octobre 2008, Madrid). Bern : P. Lang. 249-270.

Marchello-Nizia, Ch. (à par.). « L'oral représenté : un accès construit à une face cachée des langues 'mortes' », à par. dans les actes du colloque international DIACHRO-V (Lyon, oct. 2010).

Marnette, S. 1998. *Narrateur et points de vue dans la littérature française médiévale : une approche linguistique*. Bern : Peter Lang.

Marnette, S. 2006a. « La signalisation du discours rapporté en français médiéval ». *Langue française* 149 : 31-47.

Marnette, S. 2006b. « La ponctuation du discours rapporté dans quelques manuscrits de romans en prose médiévaux ». *Verbum* 1 : 47-66.

Perret, M. 1988. *Le signe et la mention*. Genève : Droz.

Pincemin, B., Guillot, C., Heiden, S., Lavrentiev, A., & Marchello-Nizia, Ch. 2008. « Usages linguistiques de la textométrie : analyse qualitative de la consultation de la Base de Français Médiéval via le logiciel Weblex ». *Syntaxe & Sémantique* 9 : 87-110.

Prévost, S., Guillot, C., Lavrentiev, A., Heiden, S. 2010. Jeu d'étiquettes CATTEX2009, version 1.3. Lyon : Projet BFM, <http://bfm.ens-lyon.fr/IMG/pdf/Cattex2009_1.3.pdf>.

Rankovic, S., Melve, L. & Mundal, E. (éd.) 2010. *Along the Oral-Written Continuum. Types of Texts, Relations and their Implications*. Turnhout : Brepols.

Söll, L. 1974. *Gesprochenes und geschriebenes Französisch*. Berlin : Erich Schmidt.

Zumthor, P. 1983. *Introduction à la poésie orale*. Paris : Seuil.

Zumthor, P. 1984. *La poésie et la voix dans la civilisation médiévale*. Paris : Presses Universitaires de France.

Zumthor, P. 1987. *La lettre et la voix : de la littérature médiévale*. Paris : Seuil.

Description du corpus

Identifiant[24]	Date	Domaine	Genre	Forme	DD	Non DD	Total
SermentsW2*	842	juridique	serment	prose	0	115	115
EulalieW2*	883	religieux	hagiographie	vers	0	188	188
AlexisS2	1050	religieux	hagiographie	vers	1484	3387	4871
RolMoign	**1100**	**litteraire**	**epique**	**vers**	**11579**	**17730**	**29309**
PhThCompS	1116	didactique	comput	vers	60	14618	14503
EpreuveJudicG*	1117	juridique	ceremonial	prose	0	414	414
LapidalS	1117	didactique	lapidaire	vers	11	9288	9299
PsOxfM*	1125	religieux	psautier	mixte	0	42500	42500
PhThBestWa	1128	didactique	bestiaire	vers	162	13591	13753
GormB	1130	litteraire	epique	vers	1188	2620	3808

[24] Dans la mesure du possible, nous utilisons les sigles bibliographiques du *Dictionnaire étymologique de l'ancien français,* http://www.deaf-page.de. Une note est ajoutée en cas d'absence de correspondance précise. Les numéros de volume éventuels sont donnés entre parenthèses. Les textes bénéficiant d'un étiquetage morphosyntaxique vérifiés sont présentés en gras, les dialogues et les textes dramatiques sont en italiques, les identifiants des textes ne contenant pas de marques de discours direct sont suivis d'un astérisque.

JuiseR	1137	religieux	sermon	vers	1399	2863	4262
DescrEnglB*	1140	historique	histoire	vers	0	1303	1303
LapidfpS*	**1150**	**didactique**	**lapidaire**	**prose**	**0**	**4781**	**4781**
PsOrneS*	1150	religieux	psautier	prose	0	705	705
BrutA	1155	historique	chronique	vers	1280	14356	15636
EneasS2(1)	**1155**	**litteraire**	**roman**	**vers**	**12897**	**22039**	**34936**
EneasS2(2)	1155	litteraire	roman	vers	13107	11690	24797
PsCambrM*	1157	religieux	psautier	prose	0	4311	4311
CommPsia1/2G2*	1164	religieux	commentaire	prose	0	48381	48381
SthomGuernW2	1173	religieux	hagiographie	vers	15062	38909	53971
BenDucF	1174	historique	chronique	vers	9910	15243	25153
FantosmeJ	1174	historique	histoire	vers	6406	11761	18167
TristThomL	1174	litteraire	roman	vers	8588	9989	18577
MirNDOrlM	1175	religieux	miracle	vers	192	766	958
RecMédJuteH*	1175	didactique	recettes	prose	0	2692	2692
CligesM	1176	litteraire	roman	vers	12894	27476	40370
TristBérM4	1180	litteraire	roman	vers	14092	13123	27215
AdgarK	1183	religieux	miracle	vers	9875	39006	48881
DialGregF	*1183*	*didactique*	*dialogue*	*prose*	*31342*	*1337*	*32679*
EpMontDeuH	1183	religieux	epistolaire	prose	0	30392	30392
SermMadnP	1183	religieux	sermon	prose	259	1969	2228
RoisC	1190	religieux	histoire	prose	32499	58080	90579
BodelFablBailNo[25]	1192	litteraire	fabliau	vers	248	476	724
BodelFablGombNo[26]	1192	litteraire	fabliau	vers	283	932	1215
BodelNicH4	*1195*	*litteraire*	*dramatique*	*vers*	*10101*	*342*	*10443*
AmAmD	**1200**	**litteraire**	**epique**	**vers**	**11931**	**13353**	**25284**
ElucidaireiiiD	*1200*	*didactique*	*dialogue*	*prose*	*25299*	*196*	*25495*
SAgnèsDobT	1200	religieux	sermon	prose	0	4691	4691
RobClariL	1210	historique	chronique	prose	3453	30564	34017
GuillDoleL	**1210**	**litteraire**	**roman**	**vers**	**11000**	**23545**	**34545**
AucR3	1212	litteraire	rbrefs	mixte	4691	5292	9983
CoincyI1K(1)	1222	religieux	lyrique	vers	5998	11456	17454
CoincyI1K(2)	1222	religieux	lyrique	vers	9091	33855	42946
CoincyI1K(3)	1222	religieux	lyrique	vers	16660	55965	72625
CoincyI1K(4)	1222	religieux	lyrique	vers	10971	78749	89720
SGraalIVQueste	**1227**	**litteraire**	**roman**	**prose**	**51638**	**56072**	**107710**

[25] Jean Bodel, *Le Vilain de Bailluel,* édition NoomenFabl.

[26] Jean Bodel, *Gombert et les deus clers*, édition NoomenFabl.

M[27]							
RosemLec	1274	didactique	roman	vers	41362	9572	50934
AdHaleFeuillL	*1276*	*litteraire*	*dramatique*	*vers*	*7316*	*9*	*7325*
BeaumCoutS	1283	juridique	traite	prose	2415	141084	143499
AdHaleRobL	*1285*	*litteraire*	*dramatique*	*vers*	*5051*	*17*	*5068*
FroissChron3D	1385	historique	chronique	prose	25067	192036	217103
GriseldisEstR	*1395*	*litteraire*	*dramatique*	*vers*	*15645*	*598*	*16243*
ManLangK	1396	didactique	manuel	mixte	9333	6163	15496
ManLangK	1399	didactique	manuel	mixte	3130	1678	4808
QJoyesR	**1400**	**litteraire**	**nouvelle**	**prose**	**11395**	**23319**	**34714**
ManLangK	1415	didactique	manuel	mixte	3156	0	3156
ComteArtS	1460	litteraire	roman	prose	10447	35359	45806
CentNouvS	1462	litteraire	nouvelle	prose	47796	104929	152725
PathelinD[28]	*1462*	*litteraire*	*dramatique*	*vers*	*10676*	*1*	*10677*
JParisW	**1494**	**litteraire**	**roman**	**prose**	**9987**	**15136**	**25123**
CommC	**1497**	**historique**	**memoires**	**prose**	**550**	**22243**	**22968**
TOTAL					548976	1333255	1882231

[27] Édition Ch. Marchello-Nizia (2011), Lyon : Équipe de la BFM, http://txm.risc.cnrs.fr/txm.

[28] *Farce de Maître Pierre Pathelin,* éd. J. Dufournet, Paris, Garnier Flammarion, 1986.