L'oral représenté dans un corpus de français médiéval (9e-15e) : approche contrastive et outillée de la variation diasystémique

Céline Guillot, Serge Heiden, Alexei Lavrentiev, Bénédicte Pincemin

L'oral représenté dans un corpus de français médiéval (9e-15e) : approche contrastive et outillée de la variation diasystémique

Céline Guillot, Serge Heiden, Alexei Lavrentiev, Bénédicte Pincemin
(UMR ICAR – ENS de Lyon / CNRS)

0 Introduction

La recherche que nous présentons ici fait suite à plusieurs travaux menés récemment sur l'oral représenté et le discours direct en français médiéval (notamment Marnette 2006a et b, Marchello-Nizia 2012, Glikman & Mazziotta à par., Guillot *et al.* à par.). Ces diverses études se sont focalisées sur les limites du discours direct et son bornage dans les manuscrits médiévaux (Marnette 2006a et b, Marchello-Nizia 2012), mais aussi sur les caractéristiques linguistiques internes à ce qu'on peut appeler *l'oral représenté* au Moyen Âge (Marchello-Nizia 2012, Glikman & Mazziotta à par. Guillot *et al.* à par.).
Il semble en effet légitime d'étudier de manière spécifique cette forme d'écrit particulier, qui se donne de façon claire comme de l'oral (même s'il est nécessairement figuré) et qui se délimite du reste du texte par un ensemble de marques linguistiques et graphiques explicites. L'un des enjeux de ces études est bien entendu de déterminer si l'oral représenté est doté d'une grammaire spécifique et de quelle manière on peut la relier, sinon à des usages vraiment oraux la langue, du moins à une forme de proximité communicative (Koch & Österreicher 1990 et 2001).
Outre l'intérêt de telles recherches pour l'étude du changement linguistique et l'impact que des usages « plus ou moins oraux » de la langue peuvent avoir sur les évolutions en cours ou à venir, les travaux sur l'oral représenté permettent également d'enrichir les réflexions théoriques et méthodologiques sur les spécificités des données exploitées par la recherche linguistique diachronique. L'un des objectifs de la présente étude sera ainsi de contribuer, de façon limitée et modeste, à ce vaste débat.
L'originalité de l'étude que nous présentons ici tient d'une part à la méthodologie de recherche, qui repose sur une approche expérimentale confrontant des hypothèses linguistiques à un corpus enrichi et outillé, d'autre part au cadre descriptif, qui se base sur une analyse contrastive des données linguistiques. C'est donc en comparant et en opposant la variation liée au fait qu'on se trouve à l'intérieur ou à l'extérieur de l'oral représenté avec d'autres paramètres de variation déjà bien identifiés dans la recherche diachronique, et spécialement pour la période médiévale, que nous mènerons notre étude de l'oral représenté. De ce point de vue, notre travail intègre les apports de la linguistique variationnelle pour l'étude du français médiéval en mobilisant plusieurs des dimensions de la variation linguistique (variations diachronique et diaphasique notamment).
Après avoir exposé la méthodologie employée pour réaliser notre analyse linguistique, nous présenterons les résultats en deux temps : d'abord notre analyse basée sur un corpus restreint aux étiquettes morphosyntaxiques vérifiées, puis notre analyse sur un corpus élargi, avant de conclure.

1 Méthodologie d'analyse

Les différentes hypothèses linguistiques de cette étude sont confrontées à l'observation directe d'un corpus numérique issu de la Base de Français Médiéval (BFM : http://bfm.ens-lyon.fr)[1] et étudié à l'aide du logiciel d'analyse de corpus TXM (http://textometrie.ens-lyon.fr).
Les opérations de vérification de l'étiquetage morphosyntaxique des textes de la BFM étant toujours en cours[2], nous avons adopté une stratégie opportuniste équivalente à celle utilisée dans (Guillot *et al.* à par.). Dans un premier temps nous avons travaillé sur un sous-corpus de textes où l'on n'a utilisé que l'observation de la répartition des étiquettes morphosyntaxiques (portées par chaque mot) dont on est sûrs de la justesse (sous-corpus vérifié : 20 textes, 606 705 mots), puis l'analyse a été étendue à un corpus comprenant plus de textes et sélectionné pour être plus représentatif vis-à-vis des dimensions de variation étudiées (corpus complet : 78 textes, 2 482 323 mots). Outre la dimension diachronique, nous nous sommes centrés dans cette recherche sur la variation en domaines textuels[3]. La répartition des domaines selon les périodes est représentée graphiquement dans la Figure 1 :

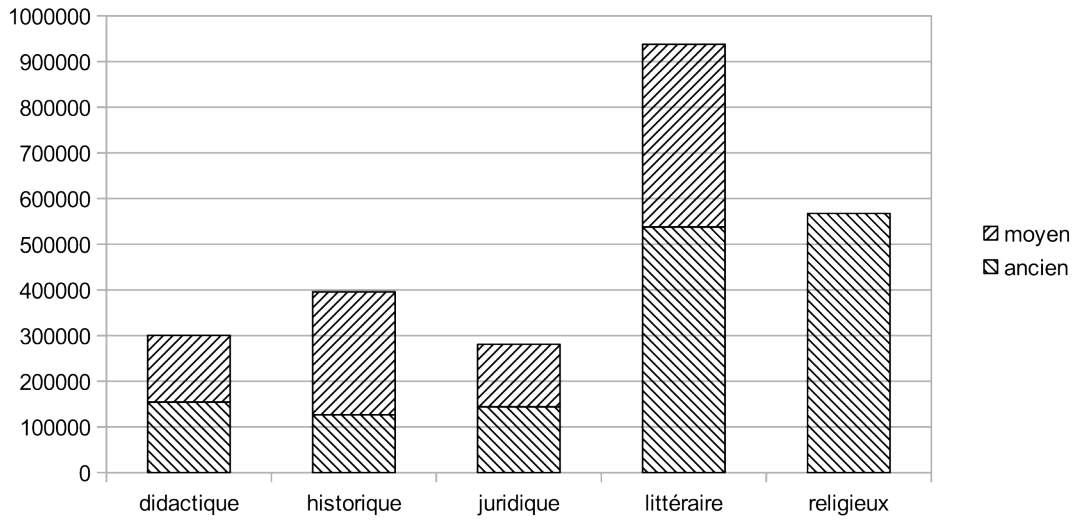

Figure 1 : Répartition des domaines textuels du corpus complet selon les périodes ancien/moyen (en nombres de mots).

Les contrastes ont été calculés à partir de la combinaison d'informations disponibles dans trois niveaux d'encodage différents :

1) chaque texte est caractérisé dans son ensemble de façon très précise en suivant les descripteurs établis par le projet CORPTEF (Guillot *et al*, 2010)[4] : la date de

[1] Dans cette étude, chaque texte sera représenté par son identifiant dans la BFM. Pour voir à quelle édition cela correspond, vous pouvez accéder à la fiche bibliographique du texte d'après son identifiant sur le portail de la BFM : http://txm.bfm-corpus.org (voir « afficher les fiches bibliographiques des textes de la BFM » dans la page d'accueil).

[2] Le jeu d'étiquettes utilisé est le système CATTEX2009 développé pour le français médiéval (Prévost et al, 2013).

[3] Le domaine textuel correspond à la destination principale du texte et au domaine d'activité auquel il se rattache : divertir → « littéraire », enseigner, instruire → « didactique », édifier → « religieux », consigner/relater les événements du passé → « historique », réguler la vie sociale → « juridique ».

[4] Voir la « Présentation des descripteurs du projet CORPTEF » : http://corptef.ens-lyon.fr/IMG/pdf/descripteurs-corptef.pdf.

composition, le domaine et le genre du texte ainsi que sa forme en « vers », « prose » ou « mixte ». Ces informations importées depuis la BFM dans TXM ont permis de contraster les textes ainsi que les séquences de mots selon les périodes ancien/moyen et les domaines ;

2) au sein de chaque texte, le discours direct a été délimité au mot près au moyen d'une balise XML appelée « <q> »[5] en s'appuyant sur les marques formelles des éditions (diverses sortes de guillemets). Ces informations importées comme structures intermédiaires des textes dans TXM ont permis de contraster ce qui est de l'ordre du discours direct (DD) de ce qui n'en relève pas (non DD) ; dans les textes entièrement dialogués de type « théâtre » ou « dialogue didactique », les prises de parole ont été délimitées au mot près au moyen d'une balise XML appelée « <sp> »[6] en s'appuyant sur les marques formelles des éditions (mise en page permettant de repérer les noms de personnages et leurs prises parole). Ces informations ont été importées de la même façon dans TXM comme structures intermédiaires des textes.

Les dénombrements des étiquettes morphosyntaxiques dans les parties s'opposant selon les différents axes de variation ont constitué les tableaux de données pour les analyses statistiques utilisées par cette étude : calcul de spécificités et analyses factorielles des correspondances (ou AFC)[7]. L'AFC est un outil de réduction de la dimensionnalité d'une matrice de décomptes. Pour chaque analyse de contraste, nous avons produit une matrice correspondante croisant :

- en colonnes les modalités du contraste : par exemple les plans textuels du discours direct texte par texte ainsi que les plans textuels ne correspondant pas au discours direct texte par texte (*cf.* Figure 2) ;
- en lignes les décomptes pour chaque étiquette morphosyntaxique possible.

Dans cette matrice, chaque colonne est comprise comme une représentation - sous forme de vecteur - d'une modalité de contraste donnée (par exemple la modalité « plan textuel du discours direct dans le texte 'roland' »). Les coordonnées du vecteur sont alors les décomptes de chaque étiquette pour cette modalité (par exemple 25 pour la ligne « PROper » : pronom personnel) ou 3 pour la ligne « DETind » : déterminant indéfini). De même, symétriquement, chaque ligne est un vecteur dans lequel une étiquette morphosyntaxique est caractérisée par sa présence dans les différents textes et plans textuels.

Le calcul d'AFC consiste à déterminer un nouveau jeu de coordonnées pour ces vecteurs, permettant de réduire le nombre de dimensions tout en maintenant le plus possible les rapports de distance d'origine entre eux. Les nouvelles coordonnées, appelées facteurs, sont ordonnées de sorte à ce que le premier facteur renseigne le mieux l'ensemble des colonnes quant à leur distance en termes de coordonnées initiales, puis le second, etc[8]. Les lignes peuvent être situées dans le même espace, qui optimise de la même façon leur représentation (*cf.* Figure 3). Dans notre analyse, nous n'avons à chaque fois utilisé que les deux premiers facteurs des résultats d'AFC, car l'information représentée était suffisante. L'interprétation de ces représentations repose alors sur une analyse de la position des modalités colonnes entre elles et par rapport au centre, de même pour les lignes entre elles et par rapport au centre, ainsi que sur diverses informations complémentaires calculées par l'AFC (la contribution d'une ligne ou colonne d'origine à un facteur, la qualité de représentation d'une ligne ou colonne sur le plan, etc.).

---

[5] Il s'agit de la balise « q » du consortium TEI : http://www.tei-c.org/release/doc/tei-p5-doc/fr/html/ref-q.html

[6] Il s'agit de la balise « sp » du consortium TEI : http://www.tei-c.org/release/doc/tei-p5-doc/fr/html/ref-sp.html

[7] TXM s'appuie sur le package FactomineR pour réaliser ce calcul : http://factominer.free.fr.

[8] Le poids d'un facteur dans cette liste est représenté par son pourcentage de contribution à l'inertie totale des colonnes.

Les AFC ont été utilisées en complément de l'analyse des spécificités dont le principe est présenté dans (Guillot *et al.* à par.).

2 Résultats

2.1 Analyse du sous-corpus vérifié

Comme on l'a indiqué plus haut, les premières investigations ont été menées sur le sous-corpus dont l'étiquetage est le plus fiable (sous-corpus vérifié). Bien qu'il soit plus limité en taille, nous utilisons ce corpus pour repérer de grandes tendances. Ces premiers résultats seront confrontés, dans un second temps, aux données fournies par le corpus complet.

2.1.1 Axes diachronie et discours direct : le calcul des spécificités

Notre objectif a d'abord été de croiser les dimensions diachronique et discours direct / non discours direct (désormais DD/non DD) afin de dégager les oppositions les plus marquées. Le calcul du score de spécificités nous a fourni un outil de mesure pertinent de la sur- ou sous-représentation des étiquettes dans les parties contrastées. Nous avons d'abord séparé les textes de chaque période (ancien et moyen français) pour opposer, dans ces périodes, le DD au non DD. Nous avons ensuite réparti le DD et le non DD dans deux ensembles pour pouvoir opposer, au sein du DD, l'ancien et le moyen français. L'opposition des périodes dans le non DD n'a pas été étudiée, puisqu'elle ne concerne pas directement notre sujet. La figure suivante illustre la façon dont les deux partitions permettant les contrastes ont été construites :

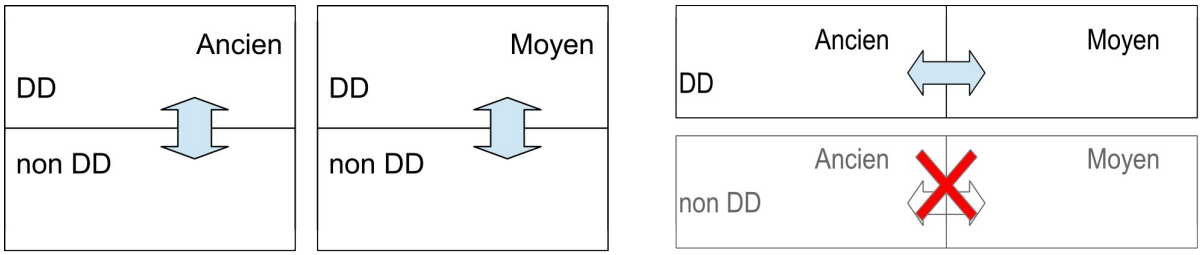

Figure 2 : Séparations opérées et oppositions étudiées au sein du corpus

Les résultats montrent peu d'évolutions dans les caractéristiques du DD et révèlent surtout des scores de spécificité toujours très élevés pour l'opposition DD/non DD. Lorsque l'on considère séparément chaque période, on observe qu'un petit nombre de catégories sont spécifiques au DD quelle que soit la période. Ces catégories sont très attendues pour certaines (interrogatifs directs, interjections), beaucoup moins pour d'autres (négation, pronoms personnels et impersonnel, pronoms adverbiaux *en* et *y*, infinitifs, conjonctions de subordination, possessifs). Le DD de l'ancien français se distingue en outre du DD du moyen français par la fréquence accrue des adjectifs possessifs et des déterminants démonstratifs. Le DD du moyen français se caractérise plus particulièrement par l'abondance des verbes conjugués et des adverbes.
Les étiquettes spécifiques au non DD à toutes les périodes sont les noms et les déterminants. Le non DD se caractérise plus spécifiquement en ancien français par les adjectifs qualificatifs et les participes passés, en moyen français par les participes présents et les prépositions.
Une analyse plus précise montre que le seul texte entièrement dialogué du sous-corpus vérifié (codé en <sp>) a des propriétés qui le rapprochent davantage du non DD que du DD. Il s'agit d'un texte didactique, une sorte de traité religieux qui prend la forme d'un dialogue fictif entre

un maître (le pape Grégoire) et son élève. Il n'est pas étonnant que ce dialogue didactique n'ait que peu de rapports avec l'oral représenté auquel le DD nous donne accès.
L'observation séparée des données selon l'axe DD/non DD conforte les résultats observés précédemment. Il apparaît que le DD se caractérise en diachronie par une certaine stabilité et qu'il s'oppose toujours très nettement au non DD. Nous verrons plus loin que, quels que soient les paramètres de variation mis en jeu et les outils statistiques employés, cette stabilité des catégories morphosyntaxiques propres au DD et au non DD est manifeste.
On note par ailleurs qu'une grande partie des catégories qui sont spécifiques au moyen français dans le DD sont également celles qui s'avéraient être spécifiques au DD aux deux périodes dans la partition précédente (pronom impersonnel, pronoms personnels, déterminant possessif, verbe à l'infinitif notamment), alors même que le moyen français est minoritaire (dans ce corpus et pour le DD). Il semble donc que les traits spécifiques au DD deviennent de plus en plus marqués au fur et à mesure que l'on progresse dans le temps.
De manière très schématique, on conclut de ce qui précède que le discours direct se caractérise par les pronoms (personnels et autres), la négation, les conjonctions de subordination, les possessifs, les infinitifs et, bien entendu, les mots interrogatifs et les interjections. Le discours non direct se distingue quant à lui particulièrement par son usage des noms et des déterminants et, en ancien français, par les adjectifs qualificatifs, autrement dit, par les éléments internes au groupe nominal.
Les résultats présentés ici doivent néanmoins être pris avec prudence. La période du moyen français n'étant représentée que par trois textes dans le sous-corpus vérifié, on ne peut exclure que les tendances observées en diachronie soient liées au style d'un auteur ou à un genre particulier. Mais nous verrons plus loin que l'analyse du corpus exhaustif confirme ces tendances. En outre, pour affiner l'analyse à l'échelle des textes, ces premiers résultats ont été complétés par l'analyse factorielle des correspondances entre les différents textes du sous-corpus.

2.1.2 Axes des textes et du discours direct : l'analyse factorielle des correspondances au niveau des textes

Le corpus vérifié étant trop peu étendu pour permettre des partitions représentatives, il a surtout l'intérêt de permettre l'étude de la variation au niveau des textes. Grâce au balisage du corpus, nous avons dissocié dans chaque texte le discours direct du reste du texte. Le calcul de l'analyse factorielle des correspondances nous a ensuite permis de construire, grâce à la comparaison des étiquettes morphosyntaxiques, la position relative de chaque plan de texte (discours direct / non discours direct) à l'intérieur d'un graphique à deux dimensions. Les regroupements des points correspondant à chaque plan de texte peuvent être mis en relation, dans la phase d'interprétation du plan factoriel, avec les différentes dimensions de variation possible. Or il s'avère que c'est bien l'opposition discours direct / non discours direct qui organise l'espace :

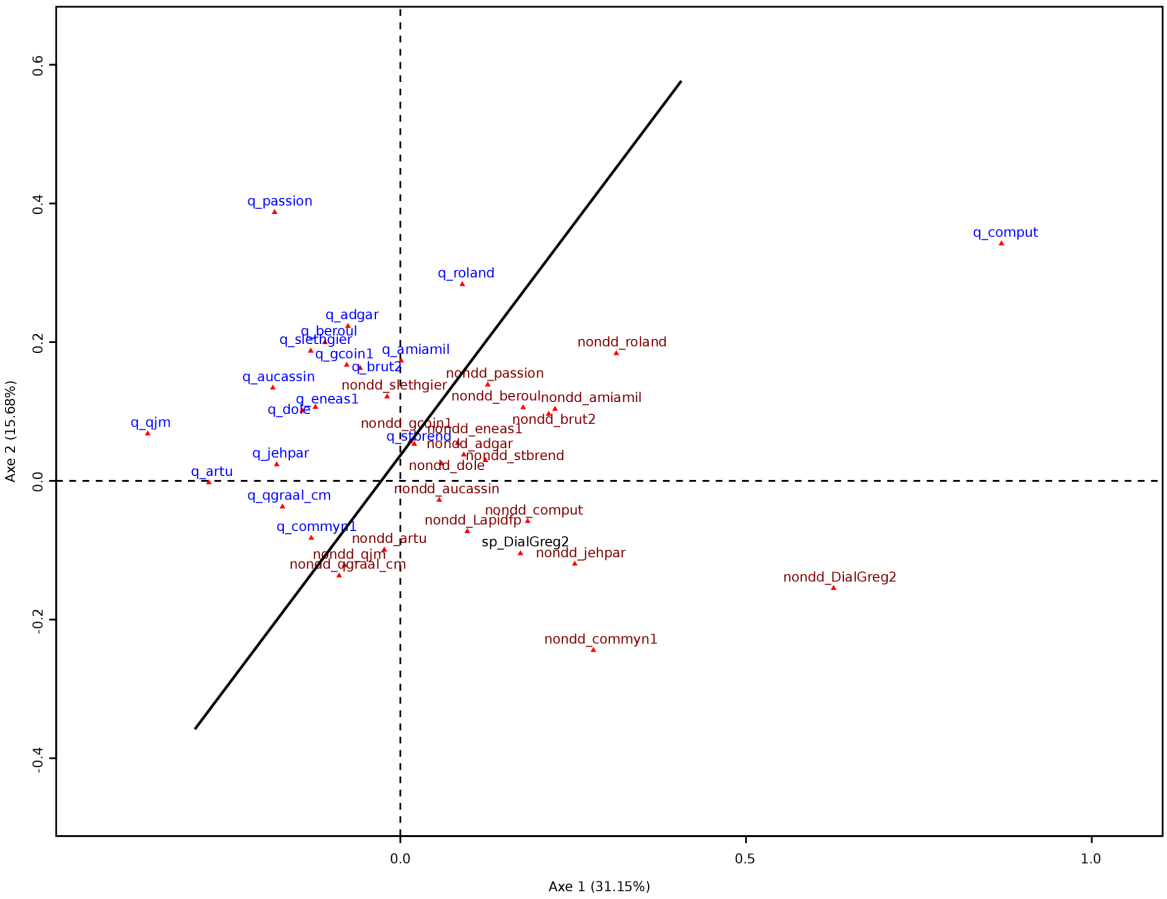

Figure 3 : Analyse factorielle des correspondances au niveau des textes dans le sous-corpus vérifié. Exemple de lecture d'un point : le point étiqueté « nondd_commyn1 » représente la position dans ce plan factoriel du plan textuel « non DD » des *Mémoires* de Philippe de Commynes représenté par le vecteur de fréquence de ses étiquettes morpho-syntaxiques.

Le calcul montre clairement que l'opposition DD/non DD constitue un contraste dominant à l'intérieur du corpus, puisque les points étiquetés « q_... » et « nondd_... » se positionnent d'eux-mêmes de part et d'autre de l'espace correspondant à cette opposition. Nous avons tracé une diagonale séparatrice pour mettre en évidence cette distribution. La seconde opposition structurante semble être le clivage entre la prose et le vers.
La position originale du Comput de Philippe de Thaon (« q_comput » dans le graphique) est liée à l'usage des guillemets propre à ce texte : ils n'y marquent pas l'oral représenté mais servent à citer des mots isolés. Par ailleurs, les quelques autres points « mal » situés sont ceux qui sont mal représentés dans le plan (cos^2<0,15). Autrement dit, ces textes se démarquent des autres selon d'autres dimensions que celles qui sont représentées ici.
L'analyse des étiquettes qui fondent la position relative de chaque point selon l'axe 1[9] montre qu'il s'agit presque exclusivement des éléments que le calcul des spécificités faisait déjà apparaître comme spécifiques au discours direct ou au non discours direct : d'un côté, les pronoms personnels, conjonctions de subordination, négation, pronom impersonnel, interjections, adverbes ; d'un autre côté, les noms propres, article défini, contradiction de l'article et des prépositions (*du*, *au*, etc.), noms communs, déterminants cardinaux et participes présents.

[9] On observe pour cela les contributions des étiquettes à l'axe 1, dans le tableau des aides à l'interprétation de l'AFC. Nous nous sommes focalisés sur les étiquettes avec une contribution d'au moins 2 %.

De même, les étiquettes qui contribuent le plus fortement à l'axe 2 sont d'une part le possessif (déterminant et pronom), les adjectifs qualificatifs, les noms propres, les verbes conjugués, les pronoms interrogatifs directs, les pronoms adverbiaux *en* et *y*, la négation et les interjections, d'autre part les conjonctions de coordination, le déterminant ledit et ses contractions (*audit*, etc.), les conjonctions de subordination, les prépositions, le pronom impersonnel et le pronom relatif.
Comme nous l'avons indiqué plus haut, ces résultats initiaux sur la nature et la force de l'opposition DD / Non DD ont été confortés lors de la seconde phase d'analyse, lorsque nous avons utilisé le même calcul statistique, l'analyse factorielle des correspondances, en l'appliquant au corpus complet, quelle que soit la dimension de variation (périodes, domaines) combinée à l'opposition DD / non DD.

2.2 Analyse du corpus complet

Le corpus complet offre un volume de données suffisamment important, pondéré et diversifié pour produire des résultats assez généraux. Mais sa faiblesse tient à son étiquetage non vérifié par des spécialistes. Il a donc fallu évaluer la qualité et le niveau de fiabilité des étiquettes du corpus pour intégrer ce nouveau paramètre à nos analyses.
L'étude du taux de réussite de l'étiquetage automatique de quelques textes vérifiés ne faisant pas partie du corpus d'apprentissage nous a permis d'identifier les étiquettes « à risque ». L'examen de l'impact des différentes étiquettes sur les résultats produits par les outils statistiques a ensuite permis de préciser de quelle façon les étiquettes sont prises en compte dans l'analyse. Les étiquettes « peu fiables » (avec un taux de réussite inférieur à 75%) se sont finalement révélées correspondre quasiment aux étiquettes peu fréquentes. Elles pèsent peu dans l'analyse factorielle et nous avons vérifié que nous pouvons les écarter sans perturber l'analyse. Les quelques étiquettes relativement peu fiables et fréquentes du corpus (nom propre, participe présent, adjectif qualificatif) ont, quant à elles, un impact limité sur la configuration des plans factoriels : nous avons contrôlé que si on les retirerait, la configuration resterait stable. Les analyses factorielles effectuées à partir des textes étiquetés du corpus complet ne sont donc pas affectées par les limites de la qualité de l'étiquetage automatique.
Sur la base de cet examen méthodique de la fiabilité des étiquettes, les calculs statistiques ont été optimisés en appliquant un seuil quantitatif à toutes nos analyses, allégeant le calcul et facilitant la mise en évidence des principaux contrastes. Seules les étiquettes ayant plus de 4000 occurrences ont été prises en compte, de sorte à écarter l'essentiel des étiquettes peu fiables et peu discriminantes et de garder l'essentiel des étiquettes fiables et pertinentes pour l'analyse contrastive. Les analyses présentées dans les sections suivantes reposent donc sur un ensemble de 24 étiquettes morphosyntaxiques[10].

2.2.1 Axes périodes et discours direct : l'analyse factorielle des correspondances au niveau des périodes

L'*analyse factorielle des correspondances* va nous permettre à présent de représenter à l'intérieur d'un même espace la configuration des étiquettes grammaticales et des plans textuels en fonction de leurs associations réciproques. Le texte est réparti selon trois plans textuels à l'intérieur de chaque période identifiée dans le corpus : discours direct (en ancien et en moyen français), prises de parole codées en <sp> (dans les dialogues didactiques et le théâtre en ancien français, dans le théâtre seulement en moyen français), non discours direct

[10] Jeu complet : 59 étiquettes. Si on enlève les ponctuations, mots étrangers et hors jeu d'étiquettes : restent 52 étiquettes. Avec le seuil à une fréquence de 4000, cela élimine 28 étiquettes : 22 non fiables, 1 moyennement fiable et 5 fiables mais peu discriminantes. Il reste 24 étiquettes, soit 18 étiquettes fiables et pertinentes pour l'analyse contrastive, et 6 moyennement fiables mais peu perturbatrices.

(en ancien et en moyen français). La Figure 4 représente la position des points correspondant à ces six parties prédéfinies et aux 24 étiquettes morphosyntaxiques :

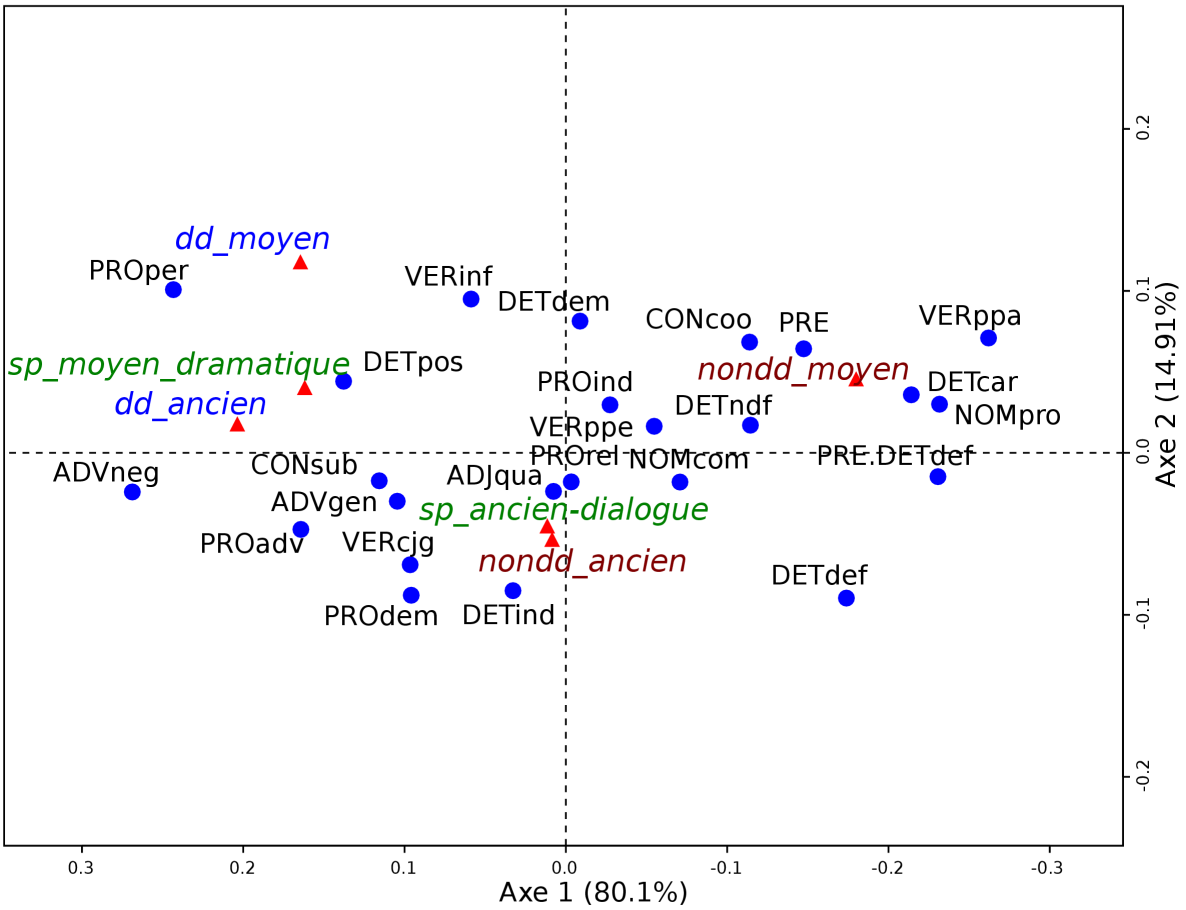

Figure 4 : Analyse factorielle des correspondances au niveau des périodes dans le corpus complet

Le premier axe est celui qui rend compte, selon le calcul mathématique, des contrastes les plus significatifs dans le corpus. Sur cet axe horizontal, ce qui organise la répartition des points c'est l'opposition DD (à gauche) / non DD (à droite).
On note également que tous les points sont correctement représentés dans le plan, à l'exception des textes dialogués de l'ancien français[11], qui semblent avoir des difficultés à se positionner par rapport aux autres selon les mêmes caractéristiques. Or il s'agit bien d'un sous-ensemble hétérogène, qui rassemble majoritairement des dialogues didactiques (les dialogues du pape Grégoire et d'autres du même type) et quelques textes de théâtre profane. Les textes dialogués du moyen français sont tous des textes dramatiques et n'ont, quant à eux, aucune difficulté à se ranger du côté du discours direct.
Le graphique permet également d'observer la position des étiquettes morphosyntaxiques sur les deux axes. Les étiquettes les moins bien représentées (adjectif qualificatif, déterminant démonstratif, pronom indéfini et pronom relatif) sont proches du centre. Elles n'adoptent pas une position tranchée sur l'axe 1 et influent peu sur l'interprétation.
Ce nouveau plan factoriel confirme de manière évidente la stabilité de la configuration des catégories morphosyntaxiques. D'une analyse factorielle à l'autre, la répartition des étiquettes varie très peu et peut s'interpréter en fonction des mêmes paramètres syntaxiques. Il semble

[11] L'indice Q_{12} de qualité de représentation de ce point sur le plan des deux premiers axes factoriels ne vaut que 0,13.

en effet qu'une opposition globale s'établisse entre les catégories qui sont liées au nom et au groupe nominal (noms propres et communs, déterminants hormis le possessif, prépositions), et les catégories qui gravitent autour du verbe (pronom personnel, verbe conjugué et à l'infinitif, adverbe - dont la négation, conjonction de subordination). Il s'agit là d'un type de configuration bien connu des approches textométriques et déjà observé dans plusieurs autres corpus (Brunet 2002, 2009[12]). Nos analyses retrouvent cette configuration et montrent l'affinité du non DD avec le groupe nominal, par opposition au DD qui accorde davantage de place au verbe.

2.2.2 Axes domaines et discours direct : l'analyse factorielle des correspondances au niveau des domaines

Notre dernière analyse repose sur le croisement de l'axe des domaines discursifs et de la partition DD / non DD. L'analyse factorielle des correspondances (Figure 5) est à nouveau d'abord structurée par l'opposition DD / non DD :

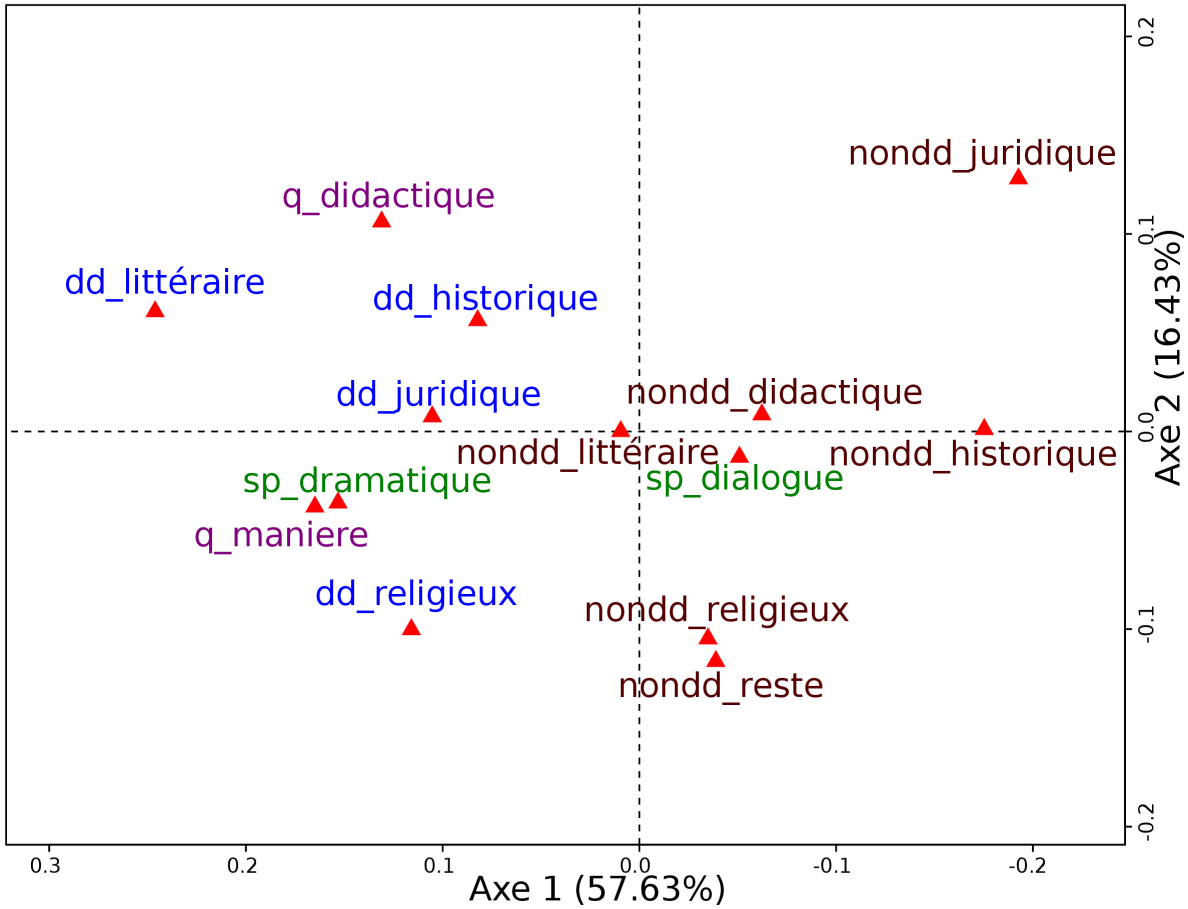

Figure 5 : Analyse factorielle des correspondances au niveau des domaines dans le corpus complet

[12] On trouvera dans le recueil (Brunet, 2009) de nombreuses AFC sur les catégories grammaticales pour différents corpus, et le commentaire de Brunet souligne régulièrement la stabilité de l'opposition entre « le clan du nom » et celui du verbe. Voir par exemple p.52 (Rabelais), p. 74 (Balzac), p.116 (Hugo), p.138 et 147 (Flaubert), p.164 (Zola), p.186 (Rimbaud). La plupart de ces articles datent de 2002-2003, au moment où ont pu être introduits dans le logiciel Hyperbase des corpus étiquetés avec l'analyseur morphosyntaxique Cordial. Mais de premières observations avaient déjà été menées sur la base d'un étiquetage plus rudimentaire du lexique (étiquetage hors contexte), l'étude sur Zola reprise dans ce recueil date par exemple de 1985, celle sur Hugo de 1988.

Pour avoir des regroupements de textes homogènes, ont été dissociés des domaines discursifs les deux genres (dialogues didactiques et théâtre) qui avaient posé problème chaque fois qu'ils étaient regroupés dans un même sous-ensemble dans les calculs précédents.
L'analyse de la Figure 5 confirme les tendances observées précédemment. Le premier axe du graphique est structuré d'un côté par le DD littéraire, didactique, religieux et le théâtre profane, d'un autre côté par le non DD historique et juridique. Ces différents types de DD et de non DD sont bien représentés dans le corpus et contrastent vigoureusement. Les points plus proches du centre du graphique correspondent aux points moins bien représentés dans le plan factoriel. Ils devraient être appréhendés grâce à d'autres facteurs de description pour être pleinement pris en compte.

3 Bilan et perspectives de recherche

Les quelques expériences présentées ici ont mobilisé un ensemble de méthodologies et d'outils statistiques dont l'usage s'étend dans les sciences sociales mais est encore relativement rare dans le champ de la linguistique. Outre que ces outils permettent de détecter, d'évaluer et de visualiser des phénomènes précis, ils offrent ici l'avantage de permettre la combinaison d'un faisceau de traits ou de paramètres variationnels prédéfinis. C'est ainsi que nous avons pu étudier de manière conjointe et contrastive l'impact sur l'usage des catégories morphosyntaxiques de l'opposition discours direct / non discours direct et des dimensions périodes, textes et domaines discursifs. Or il est apparu de manière très évidente que c'est toujours l'axe discours direct / non discours direct qui l'emporte sur tous les autres. Que l'analyse ait porté sur le corpus restreint, dont les étiquettes ont toutes été vérifiées, ou sur le corpus complet, plus vaste et moins bien étiqueté, les résultats ont montré une convergence rarement atteinte dans ce type d'analyse.
Par ailleurs, il est apparu très nettement que ce sont à peu près toujours les mêmes étiquettes qui construisent les oppositions les plus significatives. Nous avons montré que la répartition des catégories morphosyntaxiques s'ordonne selon un principe relativement clair, avec du côté du discours direct les catégories qui relèvent plutôt du domaine verbal et du côté du non discours direct les catégories liées au groupe nominal. Ces premiers résultats nous paraissent ouvrir des pistes de recherche prometteuses, dans la mesure où ils étayent l'hypothèse d'une grammaire propre au discours direct, dont les grands linéaments semblent se dessiner déjà. Il serait nécessaire, bien entendu, d'entrer dans une analyse plus fine des catégories, par exemple en intégrant le niveau des lemmes et des formes elles-mêmes.
Les grandes tendances que nous avons dégagées, et surtout, la netteté avec laquelle le discours direct semble se distinguer du reste, nous invitent à réfléchir aussi à l'importance de ce paramètre dans l'exploitation linguistique que nous faisons des ressources médiévales. Il très rare qu'on prenne en compte le facteur discours direct / non discours direct dans les recherches diachroniques ou synchroniques sur le français médiéval. Il serait pourtant très utile de l'intégrer de manière plus systématique à l'analyse linguistique, au même titre que la variation vers/prose ou la variation dialectale par exemple.
Pour conclure, nous voudrions insister sur l'apport, pour la recherche diachronique, d'une approche expérimentale fondée sur corpus et mobilisant aussi bien les outils d'enrichissement linguistique (l'étiquetage morphosyntaxique en particulier) que les outils d'analyse statistique (analyse factorielle des correspondances, etc.). Nous n'aurions pas obtenu des résultats aussi clairs si nous n'avions pas utilisé un corpus étiqueté en morphosyntaxe. Les variations phonétiques et graphiques auraient été telles, entre textes provenant de différentes régions et

de différentes époques, qu'elles auraient alourdi et encombré l'analyse, brouillant ainsi les grandes tendances rendues visibles grâce à l'analyse en parties du discours. Par ailleurs, il est évident que l'équipement et la préparation du corpus (le balisage du discours direct) étaient un préalable nécessaire à la recherche présentée ici. Tous ces éléments nous semblent plaider, de manière évidente, pour le développement d'une linguistique de corpus, qui se base toujours davantage sur des corpus enrichis et outillés.

Références

Brunet, E. (2002). « Le lemme comme on l'aime », Actes des 6èmes Journées internationales d'Analyse statistique des Données Textuelles (JADT 2002), p. 221-232. [En ligne : http://lexicometrica.univ-paris3.fr/jadt/jadt2002/PDF-2002/brunet.pdf]

Brunet, E. (2009). Comptes d'auteurs. Etudes statistiques, de Rabelais à Gracq. Ecrits choisis, tome I, Damon Mayaffre (éd.), coll. Lettres numériques n°10, Paris : Honoré Champion.

Dupuis, F. & Lebart, L. (2008).« Visualisation, validation et sériation. Application à un corpus de textes médiévaux », Actes des 9èmes Journées internationales d'Analyse statistique des Données Textuelles (JADT 2008), p. 433-444. [En ligne : http://lexicometrica.univ-paris3.fr/jadt/jadt2008/pdf/dupuis-lebart.pdf]

Glikman, J. & Mazziotta, N. (à par.). « Représentation de l'oral et syntaxe dans la prose de la Queste del saint Graal (1225-1230) », Actes du colloque international Représentations du sens linguistique V (25- 27 mai 2011, Chambéry).

Guillot C., Lavrentiev A., Pincemin B. & Heiden S. (à par.). « Le discours direct au Moyen Age : vers une définition et une méthodologie d'analyse », Actes du colloque international Représentations du sens linguistique V (25- 27 mai 2011, Chambéry). [Version auteur disponible sur HAL-SHS : http://halshs.archives-ouvertes.fr/halshs-00820262].

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, p. 1021-1032. [En ligne : http://www.ledonline.it/ledonline/JADT-2010/allegati/JADT-2010-1021-1032_025-Heiden.pdf]

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, p. 584-627.

Lebart, Ludovic et Salem, André (1994). Statistique textuelle, Paris : Dunod. [En ligne : http://ses.telecom-paristech.fr/lebart/ST.html]

Marchello-Nizia, Ch. (2012). « L'oral représenté : un accès construit à une face cachée des langues 'mortes' ». In : Guillot C., Combettes B., Lavrentiev A., Oppermann-Marsaux E. & Prévost S. (éd.) Le changement en français. Etudes de linguistique diachronique. Bern/Berlin/Bruxelles : Peter Lang, p. 247-264.

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 XXVIII, 47-66.

Prévost, S., Guillot, C., Lavrentiev, A., Heiden, S. (2013). Jeu d'étiquettes CATTEX2009, version 2.0. Lyon : Projet BFM. [En ligne : http://bfm.ens-lyon.fr/article.php3?id_article=176].