Représentation de l'oral en français médiéval et genres textuels

Céline Guillot-Barbance, Bénédicte Pincemin, Alexei Lavrentiev

Représentation de l'oral en français médiéval et genres textuels

Céline GUILLOT-BARBANCE[1], Bénédicte PINCEMIN[2], Alexei LAVRENTIEV[2]
[1] ENS de Lyon, laboratoire IHRIM
[2] CNRS, laboratoire IHRIM

1. Introduction : objectifs, méthodologie et corpus de recherche

1.1. Les enjeux du rapport entre langue parlée et langue écrite au Moyen Âge

Le rapport entre langue parlée et langue écrite est essentiel pour la linguistique historique et diachronique, particulièrement aux tout débuts du français, langue d'abord exclusivement orale qui s'éloigne progressivement du latin parlé et écrit des lettrés (Banniard, 1992). Or le fait qu'on ne possède que des témoignages écrits des états anciens du français amène naturellement à mettre en question le rapport entre la langue qu'on étudie au travers de ces sources et ses usages oraux contemporains. Et l'hypothèse relativement ancienne et courante selon laquelle le changement linguistique commence d'abord à l'oral souligne l'importance de cette question pour l'approche diachronique en général.

La relation qu'on peut établir entre oralité et scripturalité semble évoluer au fil de l'histoire du français (voir l'introduction de ce numéro). Le Moyen Âge se distingue par le fait que la langue écrite s'y diffuse sous deux formes concurrentes, latine et française, qui suivent une évolution parallèle et entretiennent des rapports constants. La place de l'écrit demeure en même temps marginale dans une société encore en grande partie semi-orale. Une distance très grande sépare les lettrés des non lettrés et l'on peut supposer que cette distance éloigne d'autant les variétés écrites des variétés orales du français.

De cette situation particulière découlent deux conséquences pour ainsi dire opposées. D'une part, le français écrit constitue dès le départ une variété haute, relativement homogène et coupée de la langue de communication usuelle. Sans avoir encore le statut d'usage normé, il se présente comme une langue de prestige (même si ce prestige est inférieur à celui du latin, cf. Lusignan, 2012) et distincte des usages oraux, de diffusion beaucoup plus locale. D'autre part, la période la plus ancienne est marquée par la « scripturalité à destin vocal » (Koch, 1993), et les premiers textes en langue vernaculaire restent étroitement liés aux conditions de l'oralité puisqu'ils sont conçus pour donner lieu à une performance vocale. Ces textes relèvent de plusieurs genres discursifs (formule de serment, déposition de témoin, bénédiction, sermon, hagiographie, théâtre religieux, chanson de geste et poésie profane des trouvères) et leur fonction (rituelle, édificatrice et/ou poétique) est indissociable de leur réalisation sous forme orale. Par la suite, d'autres genres discursifs, comme les fabliaux, les miracles, et, bien sûr, le théâtre non religieux, semblent également avoir été destinés, à un moment ou à un autre de leur histoire, à des performances du même type. On peut dès lors se demander quelles sont les conséquences de cette situation sur le mode de conception des textes (Koch & Österreicher, 2001), et si les genres oralisés se distinguent des autres et entre eux du point de vue linguistique.

1.2. Travaux antérieurs et hypothèses de recherche

Malgré l'importance du rapport entre langue parlée et langue écrite pour la linguistique historique et diachronique, la nature de ce rapport n'a pour l'instant été que très peu théorisée et étudiée (voir Combettes ici-même). Les recherches menées dans les années 80 par P. Zumthor (notamment 1986 et 1987) font exception, mais ces travaux étaient réalisés dans une perspective exclusivement littéraire et anthropologique *via* l'étude de la lyrique, du vers et de la voix. L'essor des corpus numériques et des méthodologies d'analyse instrumentées permet à présent d'envisager cette question sous un angle renouvelé.

Les recherches actuelles qui portent sur l'« oral représenté » (Marchello-Nizia, 2012) se centrent sur les séquences écrites qui se donnent de manière explicite comme la restitution de paroles orales et recouvrent, de manière plus ou moins exacte, les passages au discours direct des textes. Ces séquences sont attestées dès les premiers textes français. Elles sont explicitement identifiées et balisées grâce à des marques graphiques et/ou linguistiques qui ont varié au cours du temps mais qui existent depuis les origines (Marchello-Nizia, à par. ; Guillot *et al*., 2014 et à par.) et sur lesquelles les éditeurs s'appuient généralement en insérant les marques modernes (guillemets et tirets). Bien qu'on ne puisse prétendre que ces séquences reflètent l'oral réel, elles donnent accès à une oralité construite à l'intérieur de l'écrit. Leur délimitation montre bien que les auteurs et scribes médiévaux avaient clairement conscience d'un type de discours distinct du reste des textes. Il nous paraît de ce fait justifié d'étudier ces séquences par contraste avec ce qui les entoure, afin d'établir si des traits linguistiques leur sont spécifiques.

Deux études précédentes (Guillot *et al*., 2013 et 2015) ont été menées dans cette perspective, à partir d'un corpus numérique dans lequel le discours direct a été intégralement balisé. Le balisage a permis de séparer les unités discursives (UD) d'oral représenté et de les opposer aux UD du reste des textes. La catégorisation morphosyntaxique des unités lexicales et des outils statistiques ont permis d'étudier les fréquences des catégories et de faire ressortir par contraste les spécificités des UD d'oral représenté. Les résultats sont très nets : la distribution des catégories varie très clairement selon qu'elles se trouvent ou pas dans des UD d'oral représenté, les catégories sur-représentées et sous-représentées sont pour l'essentiel toujours les mêmes et cette opposition prime sur toutes les autres (sur la dimension diachronique et la répartition des textes en domaines notamment). Ils amènent à conclure que l'oral représenté constitue un paramètre de variation important, dont les ouvrages descriptifs pourraient rendre compte tout autant que des variations diachroniques ou régionales. La recherche que nous présentons ici vise à préciser ces premières conclusions grâce au croisement de cette opposition avec celle des genres discursifs. Notre analyse s'appuiera sur deux hypothèses principales : (i) les UD d'oral représenté doivent se grouper vers le pôle [+oralité] de l'axe, les autres UD vers le pôle [-oralité] ; (ii) les UD des textes oralisés, liés de près ou de loin à des performances orales, doivent plutôt se situer du côté [+oralité], les UD des textes non oralisés du côté [-oralité].

Les travaux de P. Koch et W. Österreicher (2001) offrent un cadre d'analyse des genres discursifs qui complète cette étude. Les 10 facteurs pragmatiques mis en évidence (communication privée ou publique, émotionnalité forte ou faible, (non) coprésence situationnelle, (non) ancrage situationnel, (non) ancrage référentiel etc.) doivent permettre de prédire la position des genres sur un continuum allant du mode conceptionnel de la proximité à celui de la distance communicative. Ce continuum ne recouvre pas parfaitement la dichotomie oralité / scripturalité, mais il prévoit une relation privilégiée entre, le pôle de la distance communicative et l'écrit et le pôle de la proximité communicative et l'oralité. De nouvelles prédictions peuvent être faites sur les genres du corpus qui devraient se situer aux deux extrémités de l'axe : le théâtre (genre dramatique) du côté de la proximité communicative et du pôle [+oralité], les genres traité, commentaire, bestiaire, lapidaire, coutumier, plaid, journal juridique à l'autre extrémité. Et, à la lumière également de nos premières observations (Guillot *et al*., à par.), on peut supposer une tendance des domaines discursifs[1] à s'ordonner du proximal au distal selon le schéma suivant : domaine littéraire → didactique, religieux, historique, juridique → actes de la pratique[2] (avec une partie centrale moins clairement ordonnée que les deux extrêmes).

1 Le domaine discursif décrit la fonction principale du texte : divertir pour le littéraire, édifier pour le religieux, enseigner pour le didactique, etc.

2 Les actes de la pratique visent à régler une question concrète grâce à l'application de la norme juridique. Le genre des chartes en fait partie.

1.3. Méthodologie de recherche

L'étude repose sur la combinaison de ressources textuelles et logicielles. Le corpus se compose de textes équipés en fonction des objectifs de la recherche : (i) chaque texte est doté de nombreuses métadonnées, dont son rattachement à un domaine discursif et à un genre, qui peuvent être reportés sur les UD[3] ; (ii) les séquences au discours direct sont balisées, ce qui permet de typer les UD d'oral représenté et les autres ; (iii) chaque unité lexicale est étiquetée grâce à un jeu d'étiquettes morphosyntaxiques (Cattex2009)[4]. Les ressources logicielles sont intégrées à la plateforme d'analyse TXM (Heiden *et al*., 2010) et exploitent ces informations grâce à quatre outils principaux : (i) la création de *partitions*[5] permet de délimiter les UD à contraster, en séparant l'oral représenté du reste des textes et en distinguant les genres textuels ; (ii) les *Spécificités* permettent de repérer statistiquement les éléments (ici, des catégories grammaticales) sur- ou sous-représentés dans une partie de la partition relativement à l'ensemble du corpus ; (iii) l'*Analyse factorielle des correspondances* (AFC), en calculant une représentation géométrique mathématiquement optimisée pour concentrer le maximum d'informations sur le minimum de dimensions, permet de visualiser la configuration d'ensemble du corpus dans une représentation plane synthétique, mais aussi de dégager la dimension de plus forte variation structurant les données ; (iv) la *Concordance* et l'*Édition* permettent d'observer les mots et catégories en contexte selon les besoins de l'interprétation.

Le balisage XML-TEI du discours direct est réalisé en mode semi-automatique à partir des marques typographiques (guillemets, balise <q>) et de la mise en page du texte (textes du théâtre et dialogues avec l'indication des locuteurs, balise <sp>). Cette méthode permet d'annoter rapidement un corpus de taille considérable, mais amène un certain nombre d'erreurs. Par exemple, dans le *Récit du voyage en Terre Sainte* d'Ogier d'Anglure, la majorité des passages marqués par des guillemets dans l'édition de référence correspond aux citations de noms géographiques ou d'autres termes utilisés par la population de l'endroit visité par le narrateur, comme dans l'exemple suivant :

> Celle fonteine appellent mesmes les Sarrasins la « fonteine Saincte Marie » (*Anglure*, p. 57).

La procédure automatique n'a pas permis de repérer cette anomalie et ces passages sont initialement traités comme les autres. En revanche, leur nature différente est mise en évidence par l'analyse statistique, si bien qu'une rectification *a posteriori* a été possible (on a recomposé les UD concernées et relancé l'analyse statistique sur les données corrigées).

L'étiquetage morphosyntaxique repose également sur une procédure automatique. Seule une partie des textes a fait l'objet d'une correction manuelle par des médiévistes[6]. Les études précédentes ont toutefois permis d'évaluer la qualité et le niveau de fiabilité des étiquettes apposées automatiquement, grâce à l'examen et à la quantification des erreurs visibles sur les textes vérifiés. On a fait en sorte que les AFC effectuées sur le corpus ne soient pas affectées par les limites de la qualité de l'étiquetage automatique (Guillot *et al*., 2015), en écartant les étiquettes les moins fréquentes et les moins fiables qui, mathématiquement, pèsent peu dans l'analyse factorielle de même que les catégories qui influent sur la construction de la première dimension mais qui ne sont pas suffisamment fiables (nom propre et déterminant *ledit*). Il reste donc au final un jeu de 34 étiquettes mobilisées pour la présente étude.

---

3 La grille des genres et des domaines discursifs est accessible en ligne : http://bfm.ens-lyon.fr/article.php3?id_article=301. Pour cette étude elle est cependant légèrement redéfinie (note 9) pour gagner en consistance et en représentativité.

4 Le jeu d'étiquettes mophosyntaxiques Cattex2009 est spécialisé pour le français médiéval. Il est accessible en ligne : http://bfm.ens-lyon.fr/spip.php?article176.

5 La partition utilisée ici est complexe (découpages fins au sein des textes), et pour cela nous avons préféré recourir à un script (qui garde trace des opérations effectuées) plutôt qu'à la fonctionnalité Partition de l'interface.

6 La vérification de l'étiquetage a été financée grâce à plusieurs programmes de recherche, sans lesquels notre étude n'aurait pas été possible : projet ANR *Corpus représentatif des premiers textes français* (http://corptef.ens-lyon.fr), projet ANR-DFG *Syntactic Reference Corpus of Medieval French* (http://srcmf.org) et projet ANR-DFG *Passage du latin au français* (https://www-app.uni-regensburg.de/Fakultaeten/SLK/Medieninformatik/PaLaFra/?lang=en).

1.4. Corpus de recherche

Le corpus BFM016DD utilisé pour cette recherche est issu de la Base de français médiéval (http://txm.bfm-corpus.org). Sa composition est très proche de celle du corpus public BFM2016, à quelques exceptions près (9 suppressions et 25 ajouts postérieurs). Il compte 137 textes composés entre le IX^{e} et le XV^{e} siècle, soit environ 4 225 000 tokens[7] (mots et signes de ponctuation). Si les textes datés du IX^{e} au XI^{e} siècle sont très rares (environ 12 000 tokens au total), à partir du XII^{e} siècle, la répartition chronologique des textes est relativement équilibrée. Un tiers du corpus (1 402 000 tokens) appartient au domaine littéraire. Les autres domaines ont des dimensions qui varient du simple au double et ne sont pas très également répartis sur tous les siècles[8]. Ce déséquilibre s'explique en partie par l'histoire de la production des textes (diminution progressive de la part de la littérature religieuse, qui passe dans le corpus de 40 à 2,3% entre le XII^{e} et le XV^{e} siècle, accroissement du domaine juridique à la fin du Moyen Âge). En ce qui concerne le genre discursif, tout espoir d'obtenir une répartition diachroniquement équilibrée nous semble illusoire, les genres évoluant, apparaissant et disparaissant au fil des siècles. Très peu de genres sont représentés à tous les siècles dans le corpus : l'hagiographie apparaît dès les textes les plus anciens, le roman et la chronique se développent à partir du XII^{e} siècle.

La majorité des genres est limitée à un domaine particulier, mais certains se manifestent dans plusieurs (comme le roman littéraire ou didactique et le traité didactique ou juridique). Suivant les préconisations de F. Rastier (2011 : 77), nous n'avons pas utilisé la métadonnée *genre* indépendamment de l'information de *domaine*, et nous avons combiné systématiquement les deux informations[9].

Par la suite, la notation adoptée pour la désignation des UD, par exemple q_rbrefsLn, combine quatre informations : (i) le type d'oral représenté ou non (« q » pour les passages entre guillemets, « sp » pour les dialogues avec indication de locuteur, « z » pour le reste des textes) ; (ii) le genre issu de la métadonnée BFM (dans l'exemple, « rbrefs ») ; (iii) une lettre majuscule correspondant à l'initiale du domaine (« L » dans notre exemple) ; (iv) un indicateur mnémonique du nombre de textes pour ce type d'UD dans le corpus (1 pour un seul texte, 2 pour deux textes, n pour trois textes ou plus).

2. Résultats et analyses

2.1. Part de l'oral représenté dans les genres discursifs

Sur l'ensemble des genres disponibles dans le corpus, la part de l'oral représenté est assez faible (elle ne dépasse pas les 10 % dans 18 des 32 genres), en particulier pour les genres qui relèvent des domaines juridique et didactique. À l'opposé, le genre dramatique (religieux ou littéraire) se limite pour l'essentiel à des dialogues, les *Serments de Strasbourg* peuvent être considérés comme reproduisant dans leur intégralité des paroles orales et les *Manieres de langage* (manuels didactiques de conversation) présentent un taux particulièrement élevé de passages au discours direct (70%). Ce taux est de 47 % dans les chansons de geste (épique). Les romans et récits brefs littéraires s'en approchent, de même que l'histoire religieuse (autour de 40 %).

7 Nous présentons les chiffres arrondis, car d'éventuelles erreurs de tokenisation et la variation des pratiques de segmentation dans les éditions de référence (Lavrentiev *et al* à par.) ne permettent pas de garantir une précision absolue.

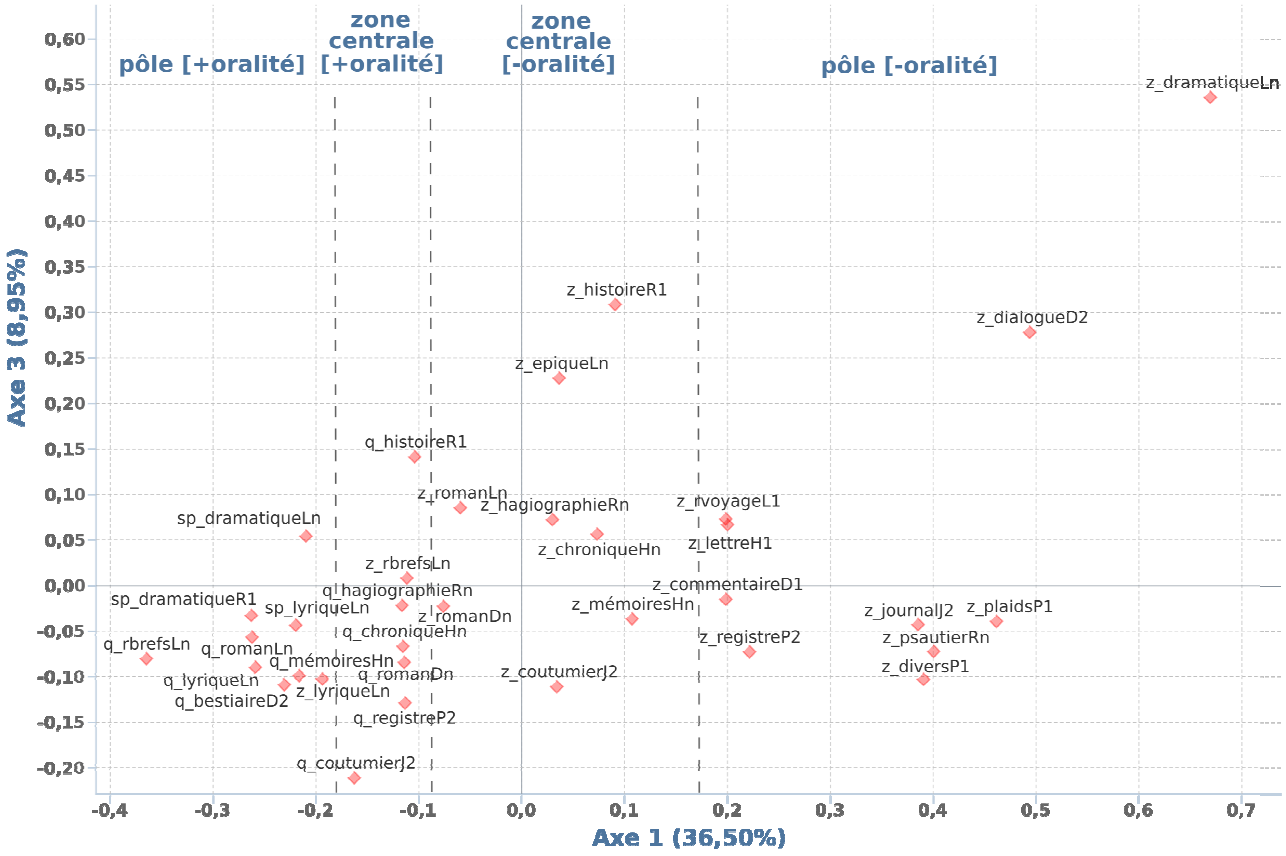
8 Un tableau de présentation du corpus détaillant la répartition des genres par siècle et la part du discours direct dans chaque genre peut être consulté dans l'archive ouverte HALSHS (https://halshs.archives-ouvertes.fr/halshs-01495132, Annexe 1).

9 L'ensemble des genres que l'on distingue selon les domaines est : le commentaire (soit didactique, soit religieux), le théâtre (soit littéraire, soit religieux), l'histoire (soit historique, soit religieuse), la lettre (soit littéraire, soit religieuse), le roman (soit didactique, soit littéraire). Nous avons en revanche regroupé les traités didactiques et religieux sous le domaine didactique, car le traité religieux se trouve de fait à l'intersection de ces deux domaines.

2.2. Émergence d'un axe oralité

La figure 1 permet de situer les UD du corpus sur un espace à deux dimensions, défini par l'axe 1 et l'axe 3[10] obtenus par l'analyse factorielle.

Figure 1. Analyse factorielle des correspondances sur le tableau UD x catégories morphosyntaxiques : plan 1 x 3 avec uniquement les UD mathématiquement bien représentées[11]



Pour la représentation graphique, les UD mal représentées (Q13 < 0,3) et contribuant peu aux dimensions représentées (contribution < 2 %) ont été effacées, de façon à optimiser la lisibilité et la fiabilité de lecture (éviter les illusions d'optique créées par la projection, typiquement de fausses proximités entre points). Le tableau 2 fournit les résultats complets et permet de situer toutes les UD sur le gradient d'oralité. Les quatre zones délimitées verticalement par les pointillés sont une annotation manuelle du graphique pour rendre compte de l'interprétation faite au § 2.3.2.

Deux résultats généraux se dégagent de la figure. D'une part, une concentration des UD d'oral représenté (préfixe « q_ » ou « sp_ ») est très nettement visible d'un côté de l'axe 1, tandis que les UD correspondant aux autres parties de textes (préfixe « z_ ») se regroupent de l'autre côté. Pour évaluer d'éventuels effets diachroniques, nous avons lancé la même analyse séparément sur la partie ancien français et la partie moyen français de notre corpus, et constaté dans les deux cas la même configuration globale des points, avec une première dimension opposant les points q_ et sp_ aux points z_. Il semble donc que la première dimension, qui, mathématiquement, dégage la dimension de plus fort contraste dans le corpus, peut s'interpréter comme un continuum entre un pôle [+oralité] et un pôle [-oralité].

10 En effet, la dimension 2 de cette analyse factorielle sert essentiellement à distinguer un seul genre (le psautier), montré comme fortement associé à l'usage des adjectifs possessifs. Une fois cette information connue, il est plus intéressant de visualiser les variations globales du corpus par le plan croisant les dimensions 1 et 3.

11 Nous remercions Serge Heiden pour son aide pour la réalisation de cette figure, notamment l'automatisation par macro TXM (CAFilter) de l'opération de filtrage des points mal représentés.

L'analyse factorielle des correspondances nous permet de dégager simultanément les catégories morphosyntaxiques contribuant le plus à chaque pôle.

Tableau 1. Catégories morphosyntaxiques caractéristiques des pôles [+oralité] et [-oralité][12]

POS les plus contributives de [+oralité]		***POS les plus contributives de [-oralité]***	
PROper :	pronom personnel	PRE :	préposition
ADVgen :	adverbe général	NOMcom :	nom commun
ADVneg :	adverbe de négation	PRE.DETdef :	article défini contracté
VERcjg :	verbe conjugué	DETdef :	article défini
PROadv :	pronom adverbial	VERppe :	participe passé
DETpos :	déterminant possessif	ADJpos :	adjectif possessif
CONsub :	conjonction de subordination	DETcar :	déterminant cardinal
VERinf :	verbe à l'infinitif	VERppa :	participe présent

Par ailleurs, la disposition des points sur le plan (qui s'élargit davantage sur l'axe 3 du côté du pôle [-oralité]) et, plus précisément, les valeurs de l'indicateur de distance à l'axe 1 (le cos^2) dans le tableau 2 ci-après, montrent une dissymétrie entre les deux pôles : le pôle [-oralité] apparaît plus hétérogène, plus dispersé, que le pôle [+oralité].

2.3. Axe oralité et genres discursifs : analyse globale

Le plan factoriel révèle que la distribution des UD repose principalement sur leur rapport à l'oral représenté. Au sein de cette première opposition, la distribution des genres ne semble toutefois pas être aléatoire. Pour examiner la situation sur l'ensemble des UD, on s'appuie sur le tableau complet listant toutes les UD dans l'ordre de leur position sur l'axe 1 (tableau 2).

Tableau 2. Liste complète des UD triées du [+oralité] au [-oralité], avec indicateurs pour l'interprétation de l'analyse factorielle

Colonnes	***c1***	***Cont1***	***Cos²1***	***Q13***	***Mass***
q_rbrefsLn	**-0,36**	**2,28**	**0,76**	0,79	0,45
q_dramatiqueR2	**-0,27**	0,03	0,20	0,20	0,01
q_lyriqueLn	**-0,26**	0,20	**0,62**	0,70	0,08
q_romanLn	**-0,26**	**27,33**	**0,87**	0,91	**10,44**
sp_dramatiqueR1	**-0,26**	0,84	**0,74**	0,75	0,32
q_lapidaireD2	**-0,24**	0,00	0,02	0,10	0,00
q_bestiaireD2	**-0,23**	0,09	**0,58**	0,71	0,05
q_mémoiresHn	**-0,22**	0,11	**0,42**	0,51	0,06
sp_lyriqueLn	**-0,22**	0,05	**0,48**	0,50	0,02
sp_dramatiqueLn	**-0,21**	**1,00**	**0,58**	0,62	0,60
z_lyriqueLn	**-0,19**	**3,56**	**0,55**	0,70	**2,49**
q_manuelDn	**-0,17**	0,51	0,27	0,27	0,44
q_coutumierJ2	**-0,16**	0,07	0,15	0,41	0,07
q_rbrefsRn	**-0,15**	0,46	0,24	0,25	0,57

12 Il s'agit des catégories morphosyntaxiques à plus forte contribution sur l'axe 1 de l'analyse factorielle, en gardant les 8 premières à coordonnée positive et les 8 premières à coordonnée négative. Elles sont présentées dans chaque colonne par valeur de contribution décroissante.

Colonnes	*c1*	*Cont1*	*Cos²1*	*Q13*	*Mass*
z_proverbesD2	**-0,15**	0,17	0,17	0,21	0,20
q_epiqueLn	**-0,13**	0,66	0,26	0,29	**1,00**
q_chroniqueHn	**-0,12**	0,88	**0,48**	0,64	**1,72**
q_hagiographieRn	**-0,12**	0,49	**0,45**	0,46	0,96
q_sermonRn	**-0,12**	0,19	0,18	0,18	0,35
q_registreP2	**-0,11**	0,05	0,21	0,48	0,09
q_romanDn	**-0,11**	**1,27**	**0,47**	0,73	**2,56**
q_sermentJ1	**-0,11**	0,00	0,06	0,14	0,00
z_rbrefsLn	**-0,11**	0,39	**0,36**	0,37	0,83
q_histoireR1	**-0,10**	0,36	0,08	0,22	0,87
z_romanDn	**-0,08**	0,93	**0,40**	0,44	**4,20**
z_bestiaireD2	**-0,07**	0,19	0,28	0,29	0,95
z_commentaireR1	**-0,07**	0,24	0,09	0,09	**1,26**
z_dramatiqueR2	**-0,06**	0,01	0,04	0,12	0,06
z_romanLn	**-0,06**	**2,07**	0,22	0,67	**15,25**
q_dialogueD2	**-0,05**	0,00	0,01	0,04	0,00
q_traitéDRn	**-0,05**	0,05	0,05	0,23	0,50
q_commentaireD1	**-0,03**	0,00	0,00	0,06	0,00
z_rbrefsRn	**-0,02**	0,03	0,02	0,13	**2,29**
z_lapidaireD2	**-0,01**	0,00	0,00	0,03	0,39
z_sermonRn	**0,00**	0,00	0,00	0,00	**3,26**
z_coutumierJ2	**0,03**	0,19	0,03	0,31	**4,37**
z_hagiographieRn	**0,03**	0,12	0,06	0,45	**3,44**
z_traitéDRn	**0,03**	0,14	0,01	0,05	**4,91**
z_epiqueLn	**0,04**	0,06	0,02	0,65	**1,22**
q_proverbesD2	**0,05**	0,00	0,00	0,00	0,00
z_computD1	**0,05**	0,03	0,02	0,07	0,39
z_chroniqueHn	**0,07**	**1,88**	0,29	0,47	**9,24**
z_lettreR1	**0,07**	0,17	0,11	0,11	0,84
z_histoireR1	**0,09**	0,47	0,05	0,57	**1,50**
z_manuelDn	**0,10**	0,08	0,07	0,19	0,19
sp_dialogueD2	**0,11**	0,28	0,16	0,16	0,65
z_mémoiresHn	**0,11**	**2,36**	0,26	0,29	**5,40**
q_journalJ2	**0,14**	0,02	0,06	0,10	0,03
sp_hagiographieR1	**0,14**	0,32	0,13	0,13	0,44
z_commentaireD1	**0,20**	0,32	**0,63**	0,64	0,21
z_lettreH1	**0,20**	0,10	**0,38**	0,43	0,06
z_rvoyageL1	**0,20**	0,98	**0,31**	0,36	0,66
z_registreP2	**0,22**	**16,29**	**0,63**	0,69	**8,73**

Colonnes	*c1*	*Cont1*	*Cos²1*	*Q13*	*Mass*
z_diversP1	**0,39**	**1,46**	**0,68**	0,72	0,25
z_journalJ2	**0,39**	**19,82**	**0,81**	0,82	**3,51**
z_psautierRn	**0,40**	**7,85**	0,12	0,12	**1,29**
z_plaidsP1	**0,46**	**2,30**	**0,68**	0,69	0,28
z_dialogueD2	**0,49**	0,07	**0,38**	0,50	0,01
z_dramatiqueLn	**0,67**	0,16	**0,32**	0,52	0,01

2.3.1. Guide de lecture du tableau 2

L'axe [+oralité] / [-oralité] correspond à l'axe 1, le tri se fait sur la coordonnée sur cet axe (colonne **c1**). On note que l'orientation de l'axe (le fait que l'oralité soit à gauche du graphique et corresponde aux coordonnées négatives) est mathématiquement arbitraire, la seule information pertinente est l'opposition entre les deux pôles, pas leur position à gauche ou à droite.

La contribution d'une UD à l'axe 1 (**Cont1**) est la participation relative (pourcentage) de cette UD à la formation de l'axe 1. Autrement dit, plus Cont1 est élevé, plus l'UD a été importante pour définir l'axe [-oralité] / [+oralité] ici observé. On a mis en gras toutes les Cont1 supérieures à 1 %.

Cette participation peut être d'autant plus forte que l'UD représente un grand volume de texte (*cf.* colonne **Mass**, en gras les masses supérieures à 1) et qu'elle se démarque de la moyenne bien dans la direction de l'axe.

La colonne **Cos²1** mesure le cosinus carré de l'angle formé par le point représentatif de l'UD : il varie entre 0 et 1. Plus le $Cos^2 1$ est proche de 1, plus l'axe 1 suffit à représenter l'UD. On a mis en gras les cosinus carrés les plus forts (supérieurs à 0,30), qui signalent donc les UD les mieux décrites par l'opposition [-oralité] / [+oralité] telle que définie par l'axe 1.

L'indicateur **Q13**, somme des cosinus carrés sur les axes 1 et 3, traduit la qualité de la représentation sur ce plan (celui de la figure 1) : une valeur proche de 1 indique que l'UD est très bien représentée (elle est peu décrite par d'autres dimensions de variation). À l'inverse, une valeur faible signale une UD qui mêle des traits des deux pôles [+oralité] et [-oralité], ou met en avant d'autres caractéristiques linguistiques.

Des délimitations horizontales séparent les quatre groupes distingués par l'interprétation (ci-après, § 2.3.2).

2.3.2. Tendances générales du positionnement des genres et des domaines sur l'axe oralité

L'examen de la répartition des genres sur la figure 1 (plus synthétique) et sur le tableau 2 associé (plus complet) permet de dégager trois résultats généraux :

(i) Deux zones (ou « pôles »), représentées aux deux extrémités du graphique, regroupent les UD dont le positionnement est très net (fort cos^2). Le pôle [+oralité] va du q_rbrefsLn au z_lyriqueLn, le pôle [-oralité] va de z_commentaireD1 à z_dramatiqueLn.

(ii) La partie intermédiaire, plus difficile à caractériser, peut également se séparer en deux parties, allant de q_manuelDn à q_histoireR1 d'une part (concentrant les UD d'oral représenté), et de z_romanDn à sp_hagiographieR1 d'autre part (pour comprendre une majorité d'UD de type z_, non oral représenté). Dans cette partie centrale, et plus encore du côté [-oralité], les cos^2 sont souvent faibles, traduisant qu'une portion importante de ces UD se laissent mal décrire par l'opposition [+oralité] / [-oralité] telle que modélisée par l'axe 1 : ces UD ont une facture morphosyntaxique différente, équilibrant les catégories grammaticales d'autres façons, plus originales par rapport au reste du corpus.

(iii) Les UD des genres qui relèvent du domaine littéraire (suffixe L) se situent majoritairement dans la zone [+oralité], les UD des genres des actes de la pratique (suffixe P) se portent dans la zone opposée. Les UD des domaines religieux et didactique, très nombreuses, tendent à se concentrer dans la large zone centrale [-oralité]. Les quelques UD du domaine historique se trouvent aux frontières entre zones centrales et pôles. Les UD du domaine juridique sont celles qui présentent la

répartition la plus étale sur l'axe 1, bien que peu nombreuses et absentes du pôle [+oralité][13]. Nos prédictions de départ sont donc partiellement vérifiées : les positions extrêmes et opposées du littéraire et des actes de la pratique sont confirmées, et pour les autres domaines aucun ordre ne se dégage clairement.

2.4. Axe oralité et genres discursifs : analyse détaillée

2.4.1. Pôle [+oralité]

Le pôle [+oralité] rassemble les UD d'oral représenté de certains genres oralisés (**dramatique**, **fabliaux**) et toutes les UD du genre **lyrique**. Les fabliaux apparaissent à travers le genre des **récits brefs**. Les **nouvelles**, dont le rapport à l'oralité a été maintes fois souligné (voir notamment Azuela, 1997), ont un comportement analogue à celui des fabliaux, c'est pourquoi tous ces textes ont été regroupés sous le même genre. Les UD d'oral représenté du **roman littéraire** se positionnent également dans la même zone.

Deux faits surprenants peuvent être soulignés. Tous les textes oralisés ne sont pas représentés dans cet espace, en particulier ceux qui relèvent du domaine religieux (hagiographie, récit bref religieux, sermon) et le genre épique, qui sont décalés vers la zone centrale. À l'inverse, la position excentrée des UD d'oral représenté de deux genres didactiques, le **lapidaire** et le **bestiaire**, et d'un genre historique, les **mémoires**, n'était pas prévisible. L'oral représenté du lapidaire est en fait extrêmement bref (une dizaine de mots) et se laisse mal décrire par l'axe 1 ($cos^2 = 0,02$), il y a peu de pertinence à le considérer pour l'interprétation du pôle. Quant à la présence des UD d'oral représenté des genres bestiaire et histoire, elle est due principalement à un trait d'oralité particulier qu'elles accentuent fortement, leur emploi massif des pronoms personnels.

2.4.2. Zone centrale [+oralité]

La zone centrale [+oralité] comporte encore presque exclusivement des UD d'oral représenté. Les textes oralisés non présents dans la zone précédente (**hagiographie**, **miracle**, **sermon, épique**) se situent dans cet espace. En font également partie les passages des **récits brefs littéraires** qui ne correspondent pas à de l'oral représenté.

À genre constant, on note ainsi des décalages entre les domaines. Les miracles religieux (q_rbrefsRn) se démarquent des récits brefs littéraires, excentrés dans le pôle [+oralité]. Ils développent en effet une autre dimension morphosyntaxique, qui met en avant les adjectifs qualificatifs (*grant, sainte, douce, las, bon, cher, chaitif, bele*...) et les déterminants indéfinis (*tel, nul, tut*...). De même, l'oral représenté des **romans didactiques** correspond moins complètement aux traits d'oralité dégagés par l'axe 1 que celui des romans littéraires, puisqu'il fait un usage abondant des conjonctions de coordination, et recourt moins que le roman littéraire aux adverbes, verbes conjugués et interrogatifs.

La présence des **proverbes**, seul point à la fois hors oral représenté et hors domaine littéraire, détonne. C'est leur fort usage des verbes à l'infinitif, des adverbes de négation, et, dans une moindre mesure, des verbes conjugués et des adjectifs qualificatifs, qui les déporte du côté [+oralité]. Leur originalité par rapport à l'axe 1 (cos^2 de 0,17) est de coupler cela avec une fréquence importante des pronoms relatifs, indéfinis et impersonnels (au lieu des pronoms personnels). Certaines constructions typiques expliquent en partie les sur-emplois observés : *Fol(z)/Mauvais est qui ne + verbe conjugué, N'est pas*.

13 L'évaluation de la répartition des domaines par rapport aux quatre zones découpées sur l'axe 1 a été contrôlée statistiquement par un calcul de spécificité portant sur la fréquence des différentes UD. Les indices les plus marqués sont la sur-représentation du littéraire au pôle [+oralité] (+2,3) et celle des actes de la pratique au pôle [-oralité] (+1,9). Les indices suivants concernent le didactique (+1,1) et le religieux (+0,89) dans la zone centrale [-oralité]. Les indices sont globalement peu élevés du fait des petits effectifs considérés (une soixantaine d'UD réparties sur les quatre zones).

2.4.3. Zone centrale [-oralité]

La zone centrale [-oralité] comporte une très forte majorité d'UD hors discours direct, correspondant à des UD au discours direct se trouvant du côté [+oralité]. Dans le cas du roman didactique, la distance entre les UD au discours direct et les UD des autres parties de texte semble assez faible. Ce sont les quelques UD d'oral représenté qui s'insèrent encore dans cet ensemble, issues des domaines didactique et religieux[14], qui retiennent notre attention.

Le **dialogue didactique** regroupe deux œuvres, les *Dialogues du pape Grégoire* et le *Dialogue de l'âme*, dans lesquelles le dialogue fictif du maître avec son élève (en réalité, seul le maître parle) permet de faire passer le contenu de l'enseignement. Le caractère très artificiel de ce dialogue ressort ici et sa position s'explique par son sur-emploi de catégories liées au pôle [-oralité] (noms, déterminants, participes) et son sous-emploi d'autres catégories typiques du pôle [+oralité] (conjonction de subordination, infinitif, pronom personnel). Ce genre n'est cependant pas très bien représenté par l'axe 1 (cos^2 de 0,16), car ses traits les plus marqués, le déterminant relatif (*li queiz, la queile, lo quei*l, etc.) et l'adjectif indéfini (essentiellement dans la construction *cel(e)/cez meisme(s)*) sont peu liés à cet axe. Ces remarques valent également pour les UD étiquetées **sp_hagiographieR1** correspondant à une autre partie du même texte (livre 2) racontant la vie de saint Benoît.

Le **commentaire religieux** est représenté par un seul texte (*Commentaire en prose sur les psaumes*, anonyme, XII^{e} s.) et n'est pas très bien décrit par l'axe 1 (cos^2 de 0,09). Certains traits le poussent plutôt du côté [+oralité] (le sur-emploi des verbes conjugués, le sous-emploi des prépositions, des participes et des déterminants), mais il présente également des traits du pôle [-oralité] (fort usage des conjonctions de coordination, sous-emploi des infinitifs, adverbes et pronoms adverbiaux). Plus ou moins indépendamment de l'axe 1, une dimension forte de sa description est son sur-emploi des pronoms démonstratifs et relatifs, ainsi que des adjectifs indéfinis (*meïsme(s), altre(s), autre(s)*), et possessifs.

Le **traité** correspond à un cas de figure analogue : tant pour ses UD d'oral représenté que pour ses parties narratives, ses traits les plus saillants (les démonstratifs, l'adjectif qualificatif, les déterminants relatifs, etc.) ne sont pas ceux de l'opposition [+oralité] / [-oralité], ou mêlent des caractéristiques des deux pôles.

L'oral représenté du **commentaire didactique** et celui des **proverbes** se caractérisent par leur brièveté (moins de cent mots). Ce faible volume fait que chaque occurrence – et chaque absence – prend une importance considérable dans le profil morphosyntaxique, qui est déséquilibré et mêle les traits relevant des deux pôles. Leurs contributions et cos^2 sont quasi nuls. Ces unités discursives trop faiblement représentées ne sont donc pas en mesure de contribuer à l'analyse.

2.4.4. Pôle [-oralité]

Le pôle [-oralité] ne présente aucune UD d'oral représenté. De manière relativement attendue, il regroupe les genres du domaine des **actes de la pratique** (seules les parties au discours direct du registre se trouvent dans la zone centrale [+oralité]). Les autres domaines sont également représentés, mais de manière plus faible : la **lettre** uniquement pour le domaine **historique**, le **journal** pour le domaine juridique, le **commentaire** et le **dialogue** pour le domaine **didactique**. On note surtout la présence d'un texte littéraire, le **récit de voyage**, dont la position peut surprendre : c'est que les repérages de l'itinéraire dans le temps et l'espace font appel à de nombreuses prépositions, et à des tournures comme *le lundi* (*mardi*, etc.) *ensuivant* qui démultiplient les participes présents, caractéristiques du [-oralité] Enfin, le **psautier**, qu'on aurait volontiers rapproché du genre lyrique, se trouve placé aux antipodes, du fait d'un fort usage des noms communs, déterminants définis et possessifs, prépositions, conjonctions de coordination. Mais, comme on l'a vu plus haut (note 10), la position de ce texte se définit aussi surtout par rapport à une

14 L'oral représenté du journal juridique se trouve être très majoritairement des passages en latin (1227 tokens) plutôt qu'en français (336 tokens). Aussi le décompte des catégories est-il fantaisiste et la position de cette UD ne correspond pas à la réalité, qui nous échappe ici.

tout autre dimension, liée aux possessifs. Quant au **z_dramatiqueLn**, il correspond aux didascalies des pièces, qui présentent les traits les plus marqués du [-oralité].

3. Conclusion

La méthodologie élaborée pour cette recherche permet d'étudier un type de discours défini par son rapport à l'oralité. Elle s'appuie sur un vaste corpus (137 textes, plus de 4 millions de mots) et des outils de mesure dont la mise en œuvre innove au fil des études, avec ici une exploitation plus approfondie de l'analyse factorielle. Elle confirme la prégnance de l'oralité représentée comme dimension contrastive dominante au sein des textes et met en évidence un continuum (axe d'oralité) sur lequel les UD se positionnent. Se confirme une certaine homogénéité des domaines discursifs, avec principalement l'affinité du littéraire avec les traits d'oralité, et le positionnement opposé des actes de la pratique. À une échelle plus détaillée et précise, les genres ressortent comme une unité pertinente pour l'étude de la variation qui nous intéresse, dans la mesure où leur distribution s'organise de façon cohérente et éclairante sur l'axe oralité, dans la limite des données disponibles. Les résultats les plus nets concernent les positions extrêmes : certains genres sont fortement polarisés [+oralité] (récits brefs, théâtre, lyrique), et d'autres marquent une affinité avec le pôle [-oralité] (traité, commentaire, voire le dialogue dont la forme, très artificielle, relève plus de l'écrit rédigé). La zone intermédiaire, assez vaste, rend notamment compte du fait que la diversité des genres ne saurait se réduire à une modélisation bi-polaire : la palette linguistique des genres est riche, et beaucoup d'entre eux s'écartent de l'axe oralité quantitativement dominant pour développer des formes d'expressions originales.

L'analyse statistique confirme par ailleurs que certains genres qui se trouvent dans des domaines différents doivent bien être distingués : romans littéraires et didactiques, récits brefs littéraires et religieux. Les faits linguistiques restent nuancés : si les genres oralisés sont tous du côté de l'oralité, ils n'ont pas tous une position très marquée, et se comportent finalement de manière très différente les uns des autres. De même, pour ce qui concerne nos hypothèses du côté [-oralité], le lapidaire et le bestiaire ne se détachent pas de la zone centrale, et leurs UD d'oral représenté font même partie du pôle [+oralité]. Contre toute attente, le psautier s'oppose au lyrique et adopte une position atypique, qui ouvre une dimension de description importante et complémentaire à l'échelle du corpus (deuxième axe de l'analyse factorielle). Ainsi, l'observation du corpus fait évoluer et enrichit nos conceptions des genres et de leur rapport linguistique à l'oralité construite.

La modélisation s'appuie ici sur un décompte de catégories morphosyntaxiques, sans cependant s'y réduire, grâce aux possibilités de retour au texte. Dans certains cas on a ainsi pu mettre en évidence que telle saillance morphosyntaxique traduisait telle lexicalisation précise, telle phraséologie récurrente, tel trait stylistique d'un auteur. Un des prolongements les plus prometteurs serait une exploration plus systématique des régularités lexicales en utilisant une description en lemmes, qui neutralise les variations graphiques omniprésentes en français médiéval. Cette voie pourrait être bientôt envisageable pour des corpus issus de la Base de français médiéval.

Bibliographie

AZUELA C. (1997). « L'activité orale dans la nouvelle médiévale. Les *Cent nouvelles nouvelles*, le *Décaméron* et les *Contes de Canterbury* », *Romania* 115, 519-535.

GUILLOT C., LAVRENTIEV A., PINCEMIN B. & HEIDEN S. (2013). « Le discours direct au Moyen Âge : vers une définition et une méthodologie d'analyse », *in* D. Lagorgette & P. Larrivée (éds), *Représentations du sens linguistique* 5, Chambéry : Presses universitaires de Savoie, 17-41, <halshs-00820262>.

GUILLOT C., PRÉVOST S. & LAVRENTIEV A. (2014). « Oral représenté et diachronie : étude des incises en français médiéval », *in* F. Neveu, P. Blumenthal, L. Hriba, A. Gerstenberg, J.

Meinschaefer & Sophie Prévost (éds), *Actes du 4e Congrès Mondial de Linguistique Française*, Paris: EDP Sciences, 259-276, <http://dx.doi.org/10.1051/shsconf/20140801284>.
GUILLOT C., HEIDEN S., LAVRENTIEV A. & PINCEMIN B. (2015). « L'oral représenté dans un corpus de français médiéval (9e-15e) : approche contrastive et outillée de la variation diasystémique », *in* K. Jeppesen Kragh & J. Lindschouw (éds), *Les variations diasystémiques et leurs interdépendances dans les langues romanes. Actes du Colloque DIA II à Copenhague (19-21 nov. 2012)*, Strasbourg : Éditions de linguistique et de philologie, 15-27, <halshs-00760647v2>.
GUILLOT C., LAVRENTIEV A., PINCEMIN B. & HEIDEN S. (à par.). « Diachronie de l'oral représenté : délimitation et segmentation interne du dialogue (12e-15e siècles) », Actes du colloque 2014 de la Société internationale de diachronie du français (Cambridge), <halshs-01313822>.
HEIDEN S., MAGUÉ J.-P., PINCEMIN B. (2010). « TXM : Une plateforme logicielle open-source pour la textométrie – conception et développement », *in* I. C. Sergio Bolasco (éd.), *Proceedings of 10th International Conference on the Statistical Analysis of Textual Data - JADT 2010,* vol. 2, Rome : Edizioni Universitarie di Lettere Economia Diritto, 1021-1032.
KOCH P. (1993). « Pour une typologie conceptionnelle et médiale des plus anciens documents/monuments des langues romanes », *in* M. Selig, B. Frank & J. Hartmann (éds), *Le passage à l'écrit des langues romanes*, Tübingen : Narr, 39-81.
KOCH P. & ÖSTERREICHER W. (2001). « Gesprochene Sprache und geschriebene Sprache. Langage parlé et langage écrit », *in* G. Holtus, M. Metzeltin & C. Schmitt (éds), *Lexikon der romanistischen Linguistik* 1-2, Tübingen : Niemeyer, 584-627.
LAVRENTIEV A., GUILLOT C. & HEIDEN S. (à par.). « Enjeux philologiques, linguistiques et informatiques de la philologie numérique : l'exemple de la segmentation en mots », à par. dans *Diachroniques*.
LEBART L. & SALEM A. (1994). *Statistique textuelle*, Paris : Dunod.
LUSIGNAN S. (2012). *Essai d'histoire sociolinguistique : le français picard au Moyen Âge*, Paris : Éditions Classiques Garnier numérique.
MARCHELLO-NIZIA C. (2012). « L'oral représenté : un accès construit à une face cachée des langues 'mortes' », *in* C. Guillot *et al*. (éds), *Le changement en français. Études de linguistique diachronique*. Bern/Berlin/Bruxelles : Peter Lang, 247-264.
MARCHELLO-NIZIA C. (à par.), « Les débuts de l'<oral représenté> en français : marquage du discours direct dans les plus anciens textes », *Mélanges Soutet*, à paraître.
RASTIER F. (2011). *La mesure et le grain. Sémantique de corpus*. Paris : Honoré Champion.
ZUMTHOR P. (1984). *La poésie et la voix dans la civilisation médiévale*. Paris : PUF.
ZUMTHOR P. (1987). *La lettre et la voix. De la littérature médiévale*. Paris : Seuil.

Résumé français

La relation entre oralité et scripturalité au Moyen Âge est approchée par une analyse en corpus (137 textes, 4 millions de mots). Les passages d'oral représenté (discours direct, théâtre, etc.) sont contrastés au reste des textes, en prenant en compte les 32 genres textuels qui les contextualisent. L'analyse factorielle des correspondances basée sur l'étiquetage morphosyntaxique (34 catégories) met en évidence comme première dimension de variation au sein du corpus un axe s'interprétant comme un gradient d'oralité, par rapport auquel chaque genre est automatiquement positionné, en distinguant ses parties d'oral représenté et ses parties complémentaires. Il ressort notamment que le caractère littéraire ou oralisé d'un genre (récité, chanté ou joué) accentue ses traits caractéristiques d'oralité ; que les façons de marquer la non-oralité sont plus dispersées et hétérogènes ; que la statistique distingue très clairement de l'oralité un genre comme le dialogue didactique, où la prise de parole est un artifice conventionnel de mise en forme du genre ; ou encore que le psautier, contrairement au lyrique dont on aurait pu le penser très proche, présente les traits opposés à l'oralité.

Mots clés
genres textuels, linguistique de corpus, français médiéval, discours direct, genres oralisés

Title
Relationships between represented oral speech in medieval French texts and textual genres

Abstract
Relationships between speech and writing in medieval French are analysed through a corpus composed of 137 texts (4 millions tokens). Text chunks representing speech (quotes, speech turns, etc.) are contrasted with remaining text parts, taking into account the genre of the text to contextualize every chunk (from a 32-genre typology). A correspondence analysis is performed on part-of-speech tags (34 tags). It reveals an orality axis as the first dimension of variation ; every genre, divided into reported speech and the rest, automatically gets a coordinate on this axis. Among the results, we observe that if a text is from the literary domain or is intended to oral performance (such as a song, a play or a recital), then its orality features are emphasized ; that the ways of expressing non-orality are more diverse and heterogenous than those of orality ; that statistics clearly sets apart from orality a genre like the didactic dialog, in which speech turns are used as a conventional and artificial layout ; or also that psalms, which could be supposed to be very close to poems and therefore to orality, are on the opposite side of the orality axis and present main features for non orality.

Keywords
text genres, corpus linguistics, Old French, direct speech, spoken genres