Diachronie du latin au français médiéval : pour une approche empirique et contextuelle (corpus, outils et philologie numérique)

Céline Guillot-Barbance

Diachronie du latin au français médiéval : pour une approche empirique et contextuelle (*corpus, outils et philologie numérique*)

Céline GUILLOT-BARBANCE

Maître de conférences à l'ENS de Lyon, UMR ICAR

Mémoire de synthèse

en vue de l'Habilitation à diriger les recherches

Jury

Marie-José Béguelin

Claude Buridant

Bernard Combettes

Walter De Mulder

Annie Kuyumcuyan

Christiane Marchello-Nizia

Catherine Schnedecker (directrice)

Décembre 2013

Remerciements

Je tiens à remercier particulièrement Catherine Schnedecker, qui a bien voulu encadrer ce travail et qui m'a soutenue intellectuellement et personnellement depuis de nombreuses années. Son énergie positive et ses conseils m'ont aidée bien des fois.

Je remercie tout spécialement aussi Christiane Marchello-Nizia pour sa générosité et son attention affectueuse. Elle m'a accompagnée depuis mes débuts en Sciences du langage et m'a souvent poussée dans des directions nouvelles pour moi. Je lui dois beaucoup.

Merci à Bernard Combettes. Son soutien chaleureux, ses remarques et ses conseils m'ont toujours été très précieux.

Merci aussi à Walter de Mulder pour son écoute et sa gentillesse et, surtout, pour sa grande disponibilité pour parler du démonstratif français !

Enfin, je tiens à remercier tout particulièrement la petite équipe lyonnaise (Alexei Lavrentiev, Serge Heiden, Bénédicte Pincemin et Matthieu Decorde) ainsi que Sophie Prévost (qui n'est pas de Lyon mais en fait un peu partie aussi). L'essentiel de ce mémoire provient de nos réflexions et de nos activités communes. Leur bonne humeur et leur soutien, leur amitié, ces derniers mois en particulier, m'ont beaucoup apporté.

Sommaire

Introduction 7

Chapitre 1 : Linguistique diachronique et linguistique des usages 13

1. Le rôle du locuteur et du contexte énonciatif du point de vue de la linguistique diachronique 13
   1.1. Le rôle du locuteur dans le changement linguistique 14
   1.2. La place des paramètres de l'énonciation dans la description linguistique 18
2. L'opposition oral *vs* écrit revue à la lumière de la linguistique diachronique 19
   2.1. L'oral représenté et la linguistique diachronique 20
      2.1.1. Spécificités de l'oral représenté 21
      2.1.2. Oral représenté et changement linguistique 23
   2.2. L'écrit et la linguistique diachronique 26
      2.2.1. Ecrit et changement linguistique 27
      2.2.2. Ecrit et contacts linguistiques avec le latin 32

Chapitre 2 : Linguistique diachronique et méthodologie de corpus 39

1. Le corpus d'analyse 39
   1.1. Définition du corpus 39
   1.2. La représentativité du corpus 41
   1.3. Les critères de variation externe / interne 43
   1.4. L'équilibrage du corpus 47
2. L'approche contrastive 48
   2.1. Présentation de l'approche 48
   2.2. Le cadre typologique 49
   2.3. Les unités de description 55
3. L'annotation linguistique des corpus 58
   3.1. Les buts et les fonctions de l'annotation linguistique 59
   3.2. L'apport heuristique de l'annotation 65
4. Les outils d'analyse 72
   4.1. Outils permettant l'étude d'un phénomène linguistique ciblé 72
   4.2. Outils permettant l'étude d'un type de discours 77
   4.3. La navigation entre outils et niveaux d'analyse différents 79

Chapitre 3 : Linguistique diachronique et développement des infrastructures de recherche 83

1. Le maniement des ressources linguistiques 84
   1.1. La préparation et le formatage des données primaires 84
      1.1.1. La segmentation en unités 86
      1.1.2. Les graphies 89
      1.1.3. La délimitation du texte 90
   1.2. La normalisation des formats 92
2. La politique de diffusion des ressources 96
   2.1. Le principe de l'ouverture 97
   2.2. Le principe de la paternité 98
3. L'organisation communautaire et les échanges scientifiques 100
   3.1. Les normes d'édition des textes médiévaux 100
   3.2. La création d'un corpus bilingue latin / français 103

Conclusion 107

Bibliographie 109

Curriculum vitae 115

Articles cités 123

Introduction

J'aimerais revenir en quelques mots, en ce début de synthèse, sur mon parcours personnel, parce qu'il me semble assez bien refléter des évolutions qui ont marqué, à des époques différentes, le champ disciplinaire des Sciences du langage.

Ma formation première s'est déroulée à l'Ecole nationale des chartes, où l'on m'a fait découvrir les « sciences auxiliaires de l'histoire », parmi lesquelles la paléographie, la codicologie, l'histoire du droit et des institutions, la diplomatique, l'histoire du livre, le latin médiéval… et la philologie romane. Les études que j'ai entamées dans le même temps en Sciences du langage à l'Université Paris V puis à l'EHESS pour le DEA m'ont tout naturellement poussée à orienter ma thèse d'école des chartes vers l'édition d'un texte médiéval, la traduction en 1409 par Laurent de Premierfait d'un des textes latins de Boccace les plus célèbres au Moyen Âge, le *De casibus virorum illustrium* (*Des cas des nobles hommes et femmes*). L'édition d'une partie de ce texte m'a permis de m'initier aux pratiques de l'édition « traditionnelle » et de commencer à travailler sur la langue du 15ème siècle (mon DEA portait sur l'analyse des temps verbaux dans ce texte). C'est également à la faveur de cette thèse que j'ai eu la chance de rencontrer C. Marchello-Nizia pour la première fois.

Si ma curiosité et mon goût personnel m'attiraient nettement du côté de l'Université et de la linguistique encore relativement florissante à cette période, ma scolarité à l'Ecole des chartes a tout de même laissé en moi quelques traces dont je mesure mieux aujourd'hui tous les apports possibles. Elle explique en tout cas de manière très directe le fait que je me sois dirigée du côté de la linguistique diachronique et de l'histoire du français, lorsqu'après une interruption de plusieurs années et un séjour plus ou moins heureux dans les archives j'ai finalement choisi d'entamer une thèse de doctorat.

L'intérêt de ce parcours est sans doute qu'il reproduit, à quelques années de distance, ce qu'un grand nombre de collègues ont certainement eu l'impression de vivre aussi (et que racontent très bien J.-C. Chevalier et P. Encrevé dans leur ouvrage de 2006) : formation classique et philologique, découverte puis saut dans les Sciences du langage en pleine effervescence. Mais ce qui était peut-être assez courant dans les années 1960 l'était certainement moins une trentaine d'années plus tard, et je m'aperçois aujourd'hui que cette trajectoire non rectiligne m'a donné l'occasion de côtoyer des approches et des milieux très différents. Le sentiment de n'appartenir pleinement à aucun d'eux m'ayant progressivement quittée, j'aimerais à présent

concilier les deux, et, si possible, favoriser le dialogue entre deux disciplines trop longtemps perçues comme antagonistes (en raison sans doute de la scission créée par l'émergence et l'autonomisation de la linguistique dans l'Université française).

En somme, ce qui n'était au départ qu'un simple attrait pour le Moyen Âge, ses textes, son histoire et sa spiritualité, s'est peu à peu mué, au hasard de ma formation et de mes rencontres, en intérêt personnel pour la langue médiévale, ses mutations de tous ordres, mais aussi, les questions spécifiques qu'elle pose au linguiste.

Je m'aperçois également, chemin faisant, que les limites temporelles que les habitudes universitaires assignent aux états de langue enseignés dans les cursus académiques (français du Moyen Âge, français du 16ème siècle, etc.) gagnent à être dépassées. Les recherches que j'ai réalisées sur le passage du latin tardif au français me convainquent de la nécessité d'étudier les changements linguistiques sur la longue durée. Il me semble que ces recherches montrent aussi en quoi la diachronie peut parfois – souvent ? – éclairer la description synchronique du système à un moment donné de son histoire (une bonne compréhension du passage du système déictique du latin à celui du français permet de toute évidence de mieux décrire et comprendre l'un et l'autre).

Outre qu'il est à l'origine de ma spécialisation progressive dans le domaine de la diachronie du français, mon passage à l'Ecole des chartes a sûrement déterminé aussi mon rapport à la langue et au texte médiéval, à son support graphique, à sa variabilité intrinsèque, aux multiples résistances qu'il offre au lecteur d'aujourd'hui (déchiffrement des graphies, résolution des abréviations, délimitation des unités linguistiques, etc.). La fréquentation des manuscrits médiévaux français et latins explique sans doute l'attention que j'aimerais accorder dans mes recherches personnelles aux aspects matériels de ces témoins du passé. Elle explique peut-être aussi une forme de résistance aux théories et aux descriptions linguistiques très abstraites et formelles, totalement détachées du contexte linguistique (entendu sans son sens le plus large). Enfin, elle m'aide peut-être à mieux percevoir l'importance du bilinguisme médiéval et la nécessité de lui faire une place, y compris dans les études exclusivement consacrées à l'histoire du français[1].

Ces préoccupations personnelles rejoignent assez bien, pour certaines d'entre elles, les thématiques qui sont au centre de la linguistique de corpus. C'est ce que j'ai peu à peu découvert à l'ENS de Lyon grâce à l'UMR ICAR à laquelle j'appartiens depuis plus de dix

[1] L'étude conjointe du français et du latin se justifie doublement pour le Moyen Âge : d'une part, parce qu'elle offre une perspective diachronique beaucoup plus longue et plus riche, d'autre part, parce qu'elle permet d'étudier les contacts entre ces deux langues. Et l'on sait que les contacts et échanges sont intenses et constants sur l'ensemble de la période.

ans et, surtout, grâce à mes collègues et amis les plus proches. Comme je l'ai dit plus haut, la démarche diachronique, appliquée notamment aux périodes les plus anciennes de l'histoire du français, présente des difficultés importantes, qui sont liées principalement à l'accès aux données qu'on étudie. Cette position particulière amène naturellement à accorder une importance toute spéciale aux aspects méthodologiques du travail de recherche. Plus qu'ailleurs sans doute et par nécessité, elle rend le linguiste médiéviste conscient que le point de vue construit l'objet.

Le hasard a également fait que j'ai « hérité » avec mes collègues en 2006 de la Base de français médiéval, au développement de laquelle j'avais participé un certain temps, comme beaucoup d'autres. Je me suis dès lors trouvée insérée dans le réseau des médiévistes travaillant sur l'histoire du français et œuvrant à la création d'outils de recherche. Ces collaborations nationales et internationales ont permis la réalisation de plusieurs programmes de recherche financés par l'Agence nationale pour la recherche. J'y ai beaucoup appris et ai apprécié le travail collaboratif à destination de la communauté des chercheurs.

J'aimerais aborder dans cette synthèse plusieurs facettes des travaux réalisés dans ce cadre. Il n'est pas si fréquent que les aspects techniques et parfois très pratiques de l'ingénierie de recherche trouvent leur place dans des publications académiques. Ce sera aussi une manière pour moi d'apprécier et de mieux reconnaître les évolutions d'une discipline dont les méthodes et les pratiques cherchent à devenir, me semble-t-il, davantage expérimentales et collectives.

Ce faisant, j'essaierai surtout de mettre en évidence les spécificités de la démarche diachronique et ses apports pour la recherche linguistique en général. Il m'apparaît finalement que c'est peut-être là sa valeur principale ou, en tout cas, ce qui lui donne le plus d'attraits à mes yeux (bien au-delà de l'amour pour la langue ancienne, de la beauté et de l'exotisme des manuscrits, etc.). Et, comme je l'ai dit plus haut, une partie de son intérêt tient sans doute aux difficultés qu'elle pose et aux stratégies qu'elle conduit à inventer pour aborder nos objets et questions de recherche.

L'une de ses caractéristiques est de confronter souvent le chercheur à l'absence de données, au manque de preuves, par suite des aléas de la conservation matérielle des documents anciens et en raison de la lente progression de l'écrit en langue vernaculaire. Les ressources fragmentaires qui nous sont parvenues laissent, par ailleurs, bien des questions ouvertes : quel écart y a-t-il entre les textes que nous conservons et les usages oraux de la langue ? De quels types d'usages ces données écrites rendent-elles compte ? Comment les gens écrivaient-ils et lisaient-ils ces textes ? Quels sont les changements qui ont affecté la mémoire et le traitement

cognitif de l'information depuis les origines ? Quel a été rôle de ces facteurs dans les formes prises par l'écrit dans l'histoire du français ? etc. Les réponses à ces questions sont pourtant essentielles pour notre compréhension du système linguistique auquel les ressources existantes nous donnent accès.

Certains de ces points seront abordés, ou effleurés, dans la suite de ce travail, mais de façon relativement indirecte puisqu'on voit difficilement comment les appréhender autrement. D'autres éléments examinés dans cette synthèse sont plus spécifiquement reliés au changement linguistique et à la façon dont on peut l'étudier. Plus qu'un véritable panorama de mes travaux de recherche, ce mémoire rassemble les questions qui me paraissent sous-tendre la recherche diachronique en général, ces questions étant d'ordre méthodologique (relatives aux données disponibles et à la façon dont on les analyse) autant que scientifique. L'introduction rédigée par notre équipe au numéro de la revue *Corpus* (2008) consacré à la constitution et l'exploitation des corpus d'ancien et de moyen français ([doc 5] Guillot *et al*. 2008) contenait déjà plusieurs des points développés dans la suite de ce travail.

Ce mémoire de synthèse est divisé en trois chapitres. Le premier chapitre porte sur les aspects communicatifs et énonciatifs des énoncés médiévaux, et surtout, sur l'importance d'une approche pragmatique pour l'analyse linguistique des ressources médiévales. Les développements récents de la linguistique diachronique amènent en effet à accorder une place toute spéciale à ces phénomènes et à cette approche, et, corollairement, au rôle joué par le locuteur dans les évolutions linguistiques (section 1). L'examen des aspects linguistiques et matériels propres aux documents médiévaux m'amènera à revenir sur la pertinence de l'opposition écrit/oral pour l'étude du changement linguistique et sur les effets sur les textes rédigés en français des échanges avec le latin (section 2).

Le second chapitre aborde différents aspects du traitement et de l'analyse des corpus linguistiques médiévaux et tente de mettre en évidence, d'une part, les problèmes spécifiques qu'ils posent au linguiste, d'autre part, leur apport possible pour le développement d'une recherche plus expérimentale. Je reviendrai notamment sur mes expériences passées en matière de constitution de corpus (section 1) et de méthodologie d'analyse (section 2), ainsi que sur l'annotation des corpus diachroniques (section 3) et les outils d'analyse et de recherche (section 4).

La dernière partie de ce travail (chapitre 3) prolonge les réflexions méthodologiques abordées dans les chapitres précédents par une présentation des développements récents des infrastructures de recherche dans le domaine des corpus linguistiques, et plus spécifiquement des corpus écrits. Le formatage des données primaires et la normalisation des formats (section

1), la diffusion des ressources linguistiques (section 2) et les modalités des échanges communautaires (section 3) seront successivement abordés, dans la double perspective du chercheur utilisateur et producteur de ressources linguistiques numériques.

Ce panorama peut sembler très vaste et certains points seront traités plus rapidement que d'autres. Ce qui m'importait le plus dans ce travail était de dégager ce qui m'est progressivement apparu, au fil des années et des étapes de mon parcours personnel, comme un cadre de recherche particulièrement riche, sans doute multiforme, mais peut-être plus apte à appréhender la complexité des phénomènes qui nous intéressent. Il se pourrait qu'il aide aussi à mieux en apprécier les limites …

Chapitre 1 : Linguistique diachronique et linguistique des usages

Ce premier chapitre vise à montrer en quoi la linguistique diachronique favorise le développement d'une approche pragmatique des phénomènes langagiers, l'étude du changement linguistique tendant à relier les phénomènes observés à l'activité du locuteur dans le contexte pragmatique de l'énonciation. La recherche diachronique – et cela est peut-être d'autant plus vrai qu'on s'intéresse aux périodes les plus reculées – conduit également à s'interroger sur les sources dont nous disposons et sur leurs rapports avec les conditions matérielles de leur production. J'essaierai de montrer dans la section 2 qu'une des manières d'intégrer cette dimension est de reprendre et de redéfinir l'opposition très débattue entre oral et écrit. Les contacts et échanges linguistiques avec le latin, la langue qui domine largement et pendant longtemps à l'écrit, seront également abordés dans ce cadre.

1. Le rôle du locuteur et du contexte énonciatif du point de vue de la linguistique diachronique

> « *The assumption of endogeny, being generally the preferred hypothesis, functions in practice as the* default *hypothesis. Thus, if some particular change in history cannot be shown to have been initiated through language or dialect contact involving speakers, then it has been traditionally presented as endogenous. Usually, we do not know all the relevant facts, and this default position is partly the consequence of having insufficient data from the past to determine whether the change concerned was endogenous or externally induced or both : endogeny is the* lectio facilior *requiring less argumentation, and what Lass has called the more* parsimonious *solution to the problem.* » (Milroy 2003 : 144)

Les chercheurs qui travaillent en diachronie butent généralement sur le problème des causes du changement linguistique. Aux tenants d'une approche endogène, qui explique les

évolutions par des raisons internes à la langue-même, s'opposent ceux qui attribuent au locuteur, et non aux insuffisances du système, un rôle de premier plan. L'opposition entre causes endogènes et externes peut paraître trop simpliste, mais il me semble qu'elle nous met en garde contre une pente assez naturelle de la linguistique diachronique. J'essaierai donc de défendre dans la section suivante une linguistique diachronique qui soit centrée sur le rôle du locuteur et du couple qu'il forme avec l'allocutaire dans le contexte de l'énonciation.

1.1. Le rôle du locuteur dans le changement linguistique

Les recherches menées depuis une trentaine d'années au sein de la linguistique diachronique, dans le cadre en particulier de la théorie de la grammaticalisation, ont permis des avancées significatives dans notre connaissance du changement linguistique, dans ses aspects théoriques et descriptifs. Elles ont également participé à des débats plus vastes, sur l'articulation possible des différents niveaux d'analyse par exemple (niveaux phonologique, morpho-syntaxique, sémantique et pragmatique).

Si les recherches réalisées dans le sillage d'E. Traugott, notamment, reposent bien, et de manière constante, sur une claire séparation entre pragmatique et sémantique, ce nouveau cadre de recherche se signale en même temps par l'importance qu'il accorde aux aspects énonciatifs (centrés sur le locuteur), et parfois même interactifs (centrés sur le couple locuteur/destinataire) du message linguistique. Pour prendre un exemple relativement concret et récent, les notions de *subjectification* et d'*intersubjectification*, qui semblent susciter un intérêt croissant de la part des chercheurs, sont définies comme des processus de « sémanticisation », de sédimentation ou de codification d'une valeur sémantique nouvelle héritée d'une inférence pragmatique contextuelle :

> « *I have hypothesized that subjectification and intersubjectification involve the reanalysis as coded meanings of pragmatic meanings arising in the context of speaker-hearer negotiation of meaning. Subjectification is the development of meanings that express speaker attitude or viewpoint, while intersubjectification is the development of the speaker's attention to addressee self-image*. » (Traugott 2010 : 60)

Dans la mesure où elle implique la transformation d'une valeur pragmatique en valeur sémantique, la subjectification se distingue de la notion de *subjectivité*, qui caractérise plus

généralement les usages linguistiques en contexte. Ce n'est qu'à partir du moment où les effets pragmatiques se généralisent et intègrent la composante sémantique d'une unité linguistique que l'on peut affirmer qu'une évolution a eu lieu. Et la théorie de la grammaticalisation identifie dans ce processus l'une des voies principales du changement linguistique.

L'étude de ces phénomènes témoigne d'une attention constante aux aspects communicatifs du langage, mais surtout, elle leur assigne une place et un rôle centraux dans les processus d'évolution. Elle renouvelle ainsi l'affirmation du lien entre changement et usage (d'où une prise en compte constante des fréquences) et fait du locuteur, d'une part, du couple locuteur-allocutaire, d'autre part, les véritables initiateurs et acteurs des modifications du système. La tendance universelle des sujets parlants à exploiter les ressources linguistiques pour mettre en avant leur point de vue personnel et, parfois aussi, pour intégrer celui du destinataire semble être au cœur du système et réguler son fonctionnement en situation[2].

Les déictiques se prêtent, par définition, particulièrement bien à ces évolutions, leur interprétation se calculant en grande partie grâce à leur contexte d'occurrence. C'est peut-être ce qui explique que ces éléments évoluent beaucoup dans toutes les langues du monde (voir Diessel (1999) pour une étude des différentes formes de grammaticalisation des démonstratifs dans un grand nombre de langues). L'émergence de l'article défini en français à partir du démonstratif ILLE du latin, par exemple, est décrite par Carlier & De Mulder (2010) comme un phénomène d'intersubjectification, qui repose sur une définition particulière, ou étendue, de cette notion.

L'évolution sémantique des démonstratifs en français semble pourtant suivre une évolution différente, et peut-être même inverse. Leur valeur personnelle, héritée du système ternaire du latin, se perdant peu à peu au profit d'un sens très affaibli en français moderne, leur trajectoire semble les mener d'un sens subjectif à un sens plus objectif (Marchello-Nizia 1997). Cette évolution sémantique se double d'une réorganisation morphosyntaxique, qui leur confère en même temps une valeur nouvelle, mais uniquement grammaticale. Si l'on peut voir dans cette évolution une forme de grammaticalisation, cette évolution semble suivre un chemin opposé à celui de la subjectification :

> « *On pourrait résumer grossièrement cette évolution ainsi : le passage au français moderne a donné le primat au modèle méta-morphologique dominant, et a abandonné*

[2] A cette tendance s'ajoutent très certainement d'autres facteurs, comme la volonté de suivre les normes et les usages de groupes sociaux par exemple.

> *une distinction sémantique spécifiant la situation d'énonciation de référence. L'ordre morphologique l'a emporté sur la sémantique, et spécialement sur la référence explicite à la situation de l'énonciation et donc à l'énonciateur* ». (Marchello-Nizia 1997 : 130).

La valeur morphologique acquise au cours de cette évolution confère bien aux démonstratifs une valeur objective, le choix des formes ne reposant plus sur l'expression de l'attitude ou du point de vue personnel du locuteur mais sur un critère purement grammatical (la catégorie morphosyntaxique dont relève la forme).

Mes recherches sur l'évolution sémantique des démonstratifs m'ont conduite à distinguer deux phases dans ce long processus d'affaiblissement, ce qui permet de préciser dans quelles conditions cette mutation a eu lieu. Dans un premier temps, les séries médiévales CIST et CIL perdent leur valeur personnelle (inclusion/exclusion de la sphère personnelle du locuteur) et développent une valeur instructionnelle plus forte (sens *token*-réflexif). Dans le stade antérieur, la sphère personnelle du locuteur correspondait à une zone abstraite, définie par le locuteur dans le contexte de l'énonciation. Le locuteur avait toujours le choix d'inclure ou d'exclure le référent de cette sphère, en fonction de la visée communicative de son discours. Lorsque les démonstratifs perdent leur trait personnel, c'est la relation du démonstratif à son propre contexte d'occurrence qui motive plutôt l'usage de CIST ou de CIL. CIST tend à s'employer lorsque le référent est saisi à l'intérieur du contexte d'occurrence, CIL lorsqu'il est saisi à l'extérieur de ce contexte[3]. Le locuteur n'a donc plus vraiment le choix. En revanche, cette nouvelle valeur des séries CIST et CIL (et surtout des formes complexes mêlant au démonstratif les particules adverbiales CI et LA) donne plus d'indications à l'allocutaire sur le mode d'identification du référent. On peut peut-être considérer que cette évolution facilite le travail cognitif du destinataire et que, de ce point de vue, elle intègre davantage son point de vue ou le coût de traitement engendré par le processus d'identification référentielle. Il s'agit-là d'une interprétation un peu hasardeuse, qui oblige à admettre une définition de l'intersubjectivité assez différente de celle que j'ai citée plus haut[4].

Dans une seconde phase (mais qui semble suivre de très près la première), et sans doute sous l'impulsion principale de la troisième série (CE) sémantiquement neutre dès le départ, l'affaiblissement sémantique de CIST et CIL parvient à son terme, et, dans le même temps, leur spécialisation morphosyntaxique s'achève. Les formes ne s'opposent plus par leur sens et

[3] Cette valeur spatiale des démonstratifs est renforcée par les particules adverbiales CI et LA en voie de grammaticalisation.

[4] Il me semble que cette interprétation va dans le même sens que celle que proposent Carlier & De Mulder (2010) pour la transformation de ILLE en article défini en français.

« l'ordre morphologique » l'a définitivement emporté. Seules les formes complexes qui contiennent les particules adverbiales CI et LA maintiennent une distinction sémantique (dont les contours restent à préciser en français moderne). Le sens du démonstratif français se limite alors pour l'essentiel à une valeur *token*-réflexive unique et à sa capacité à présenter le référent comme n'étant pas préalablement classifié (propriété qui découle très probablement de son sens *token*-réflexif ou indexical, De Mulder 1997).

S'il semble difficile de voir dans ce processus d'ensemble un phénomène de subjectification ou d'intersubjectification au sens restreint, on peut s'intéresser en revanche aux différents contextes d'usage du démonstratif et à la façon dont ils infléchissent peu à peu ses mutations sémantiques. Il semble en effet que les usages dominants des deux séries (deixis situationnelle et discursive pour CIST, anaphore et emploi mémoriel pour CIL) ont pu favoriser l'évolution sémantique décrite plus haut (primat de la valeur instructionnelle et spatiale). D'autre part, le fait qu'en situation immédiate le pôle locutorial se superpose au pôle constitué par l'occurrence-même du déictique pourrait expliquer, en partie au moins, que la valeur indexicale *token*-réflexive passe au premier plan et élimine progressivement la valeur subjective et dénotative du démonstratif. On peut faire l'hypothèse aussi que la fréquence et la prégnance de l'opposition deixis spatio-temporelle (CIST) / anaphore spatio-temporelle (CIL) et la fonction d'outil de structuration textuelle et spatiale assurée par le démonstratif dans le document écrit médiéval ont peu à peu accentué la saillance de sa valeur *token*-réflexive au détriment de sa valeur personnelle.

Tous ces facteurs paraissent susceptibles d'avoir joué un rôle dans l'évolution sémantique qui m'intéresse ici. Le fait qu'en bout de chaîne le sens subjectif du démonstratif se soit perdu en français (mais cette perte n'est probablement pas totale, même si elle n'organise plus le système en lui-même), ne remet évidemment pas en cause le rôle du locuteur et des facteurs pragmatiques dans le processus de changement. Ces quelques remarques invitent au contraire à intégrer pleinement ces facteurs et les paramètres de l'énonciation dans la description linguistique du système, en diachronie comme en synchronie.

1.2. La place des paramètres de l'énonciation dans la description linguistique

Dans plusieurs des recherches que j'ai réalisées sur le démonstratif et son histoire en français, j'ai essayé de montrer l'utilité d'une analyse linguistique associant la sémantique des expressions référentielles à la visée communicative du sujet parlant dans le contexte de l'énonciation. Cet objectif est naturellement plus difficile à atteindre pour les périodes anciennes du français, pour lesquelles, d'une part, nous n'avons plus de locuteurs, d'autre part, nous savons souvent peu de choses de la réalité des textes qui nous sont parvenus, de la façon dont ils ont été produits, dont ils ont circulé, etc.

Lorsqu'on s'intéresse à des phénomènes tels que la deixis, ce type d'approche paraît pourtant indispensable. Ce thème de recherche explique sans doute mon intérêt pour ces questions, à moins que ce ne soit l'inverse. Ce que j'aimerais souligner ici, c'est que la barrière temporelle qui nous sépare du passé nous fait peut-être prendre conscience de façon plus nette – et paradoxale – de l'importance des paramètres et facteurs énonciatifs dans le fonctionnement langagier. Ces éléments, sur lesquels nous n'avons qu'une prise très limitée et qu'on doit reconstituer comme on le peut, s'avèrent en réalité nécessaires à la compréhension du système en synchronie et en diachronie.

Il me semble aussi qu'une approche de ce type, loin d'éclater la description en micro-analyses contextuelles, permet au contraire de rassembler des usages qui peuvent sembler épars et amène à préciser ce qui relève du noyau sémantique d'une unité. C'est ce que j'ai essayé de montrer en adaptant la théorie de la sphère personnelle du locuteur à des usages du démonstratif qui n'en avait jamais été rapprochés (le démonstratif de notoriété, [doc. 11] Guillot 2012a, le pronom *cil* marquant le changement de rôle, [doc. 12] Guillot 2012b). L'analyse du contexte énonciatif, et, plus spécifiquement, de la façon dont le locuteur se positionne par rapport au destinataire de son message, a permis, en particulier, d'étendre la définition de la sphère personnelle pour proposer l'hypothèse de la « +/- prise en charge du contenu informatif du SN par le locuteur » (cf. notamment [doc. 9] Guillot 2010b). Il me semble que cette approche rend mieux compte de l'emploi en discours des démonstratifs et des effets pragmatiques du choix de l'une ou l'autre série en français médiéval, mais je crois aussi qu'elle améliore notre connaissance de ce qui constitue la composante sémantique interne de ces unités.

J'ajouterai pour clore cette section que l'attention portée aux phénomènes énonciatifs peut nous permettre de repérer des faits encore peu étudiés. Pour dire les choses de façon un peu

schématique, il me semble que les textes français les plus anciens se distinguent par un ancrage du message dans la situation de l'énonciation explicite et constant. La plupart des textes sont anonymes et la fonction d'auteur recouvre à cette période une réalité bien différente de celle à laquelle nous sommes habitués, mais cela n'empêche pas que les figures du locuteur et du destinataire soient très présentes et directement représentées dans les premières œuvres écrites en langue vulgaire. On peut certainement y voir une conséquence de la mise par écrit progressive d'une langue longtemps cantonnée au registre oral. Il s'agit peut-être d'un trait plus largement partagé par les écrits rédigés dans des sociétés semi-orales. Le fait qu'on rencontre dans les premiers textes scientifiques, par exemple, l'usage constant de l'adverbe déictique *or(e)* (« maintenant ») aux charnières de l'exposé didactique ([doc. 6] Guillot 2009), peut donner l'impression que ce discours se calque à ses débuts sur le récit littéraire, lui-même caractérisé par l'oralisation du message écrit (chansons de geste, vies de saint, etc., voir la section 2.2.1. de ce chapitre). Ce pourrait être l'un des traits qui distinguent le plus l'écrit en langue vernaculaire de l'écrit en langue savante (je reviendrai dans la section 2.2.2. sur l'importance des contacts linguistiques entre le français et le latin à l'écrit). Je ne fais que mentionner ici des hypothèses et pistes de recherche qui méritent plus ample réflexion et discussion. Mais il me semble qu'il s'agit là d'un champ et d'un cadre de recherche qui pourraient d'une part infléchir le point de vue du linguiste sur les textes et sur la langue qu'il étudie, d'autre part le conduire à s'intéresser à des phénomènes nouveaux. Les quelques éléments présentés ici amènent naturellement aussi à définir, ou à revoir, ce qu'on entend précisément par discours oral et écrit.

2. L'opposition oral *vs* écrit revue à la lumière de la linguistique diachronique

Qu'ils travaillent en diachronie ou en synchronie, sur le français moderne ou sur le français ancien, la plupart des linguistes ont désormais bien admis que l'opposition oral *vs* écrit, formulée en des termes aussi généraux, est beaucoup trop simpliste. Les nombreuses recherches menées sur le français oral d'aujourd'hui ont notamment montré la grande diversité des usages et la nécessité de bien différencier, à côté des typologies de l'écrit, différentes variétés de l'oral. Cette diversité des usages invite à remplacer l'opposition binaire oral *vs* écrit par un continuum permettant de caractériser différents types de discours oraux et écrits.

Plus récemment, le développement de l'analyse historique du dialogue (*Historical Dialogue Analysis*, voir notamment Jucker *et al.* 1999) a donné une certaine impulsion, dans les pays anglo-saxons surtout, à l'étude diachronique de l'oral. C'est ainsi que la recherche historique et diachronique nous conduit elle aussi, de manière un peu paradoxale du fait qu'elle repose, pour les périodes les plus reculées, sur des sources toujours écrites, à revoir ou à préciser l'opposition oral *vs* écrit. Une telle approche nous permet de poser sur des bases en partie renouvelées des questions tout à fait essentielles pour la linguistique diachronique, comme par exemple : quel rapport peut-on établir entre les documents écrits qui nous sont parvenus et la langue parlée par les locuteurs contemporains ? quel est le statut de l'écrit vernaculaire dans une société semi-orale ? quelle est la place de l'écrit et de l'oral dans le changement linguistique ? Nous verrons ci-dessous également que toutes ces avancées et recherches plaident pour la prise en compte, dans l'analyse linguistique, des conditions de production, de réception et de diffusion textes qui nous ont été transmis, et tout spécialement dans le cas des œuvres médiévales.

2.1. L'oral représenté et la linguistique diachronique

Il peut sembler paradoxal voire illusoire de prétendre approcher les usages oraux quand on travaille sur une langue morte telle que le français médiéval. Mais j'aimerais à nouveau montrer que la linguistique historique et/ou diachronique, peut-être en raison des difficultés ou de la résistance que les sources documentaires offrent au linguiste et peut-être aussi parce que l'étude du changement linguistique implique par nécessité qu'on tienne compte de la pragmatique de l'énonciation, peut renouveler notre l'approche de l'écrit et nous amener à nous intéresser tout particulièrement au discours direct, cette sorte d'écrit qui se donne lui-même comme une représentation de l'oral.

Je reprendrai ici le terme d'*oral représenté*, assez répandu et adopté en particulier par C. Marchello-Nizia (2012, à par. a et b), pour désigner les séquences au discours direct qui sont explicitement présentées comme les paroles d'un locuteur, même si celui-ci est bien évidemment fictif. Ces séquences sont généralement bien distinctes du reste du texte, y compris dans les manuscrits médiévaux, leur balisage s'effectuant au moyen de marqueurs qui leur sont propres et qui, d'ailleurs, évoluent dans le temps (signes de ponctuation spécialisés en français moderne, marques linguistiques spécifiques en français médiéval).

A la différence de C. Marchello-Nizia, je ne distinguerai pas nettement oral représenté et discours direct dans ce travail. Il ne fait pas de doute pourtant qu'il s'agit là d'une distinction très pertinente, les recherches de C. Marchello-Nizia ayant bien montré que certains éléments qui encadrent le discours direct dans les textes médiévaux (annonces, incises, rappels) et qui ne sont généralement pas intégrés dans les séquences délimitées par des guillemets dans les éditions modernes partagent avec le discours direct un grand nombre de caractéristiques communes (présence de verbes de parole, emploi massif du présent, etc.). Les « épisodes d'oral représenté » définis par C. Marchello-Nizia dépassent donc les limites du discours direct proprement dit puisqu'ils intègrent aussi ces éléments. Notre choix repose, quant à lui, sur des raisons pratiques. Les premières investigations menées par notre équipe ont été réalisées grâce à un pré-balisage du discours direct dans un corpus de textes médiévaux et ce balisage a coïncidé avec les guillemets insérés dans le texte médiéval par les éditeurs modernes (cf. chapitre 2, section 2.3.). Nous avons par conséquent été amenés à limiter nos analyses au discours direct *stricto sensu*.

Les deux études que nous avons menées jusqu'à présent visaient à mieux décrire les spécificités de l'oral représenté en français médiéval par contraste avec tout ce qui lui est externe et que j'appellerai ici par commodité *récit* ([doc. 13] Guillot *et al*. à par. a et [doc. 14] Guillot *et al*. à par. b). L'objectif de telles recherches est double : elles permettent d'une part de mieux caractériser la grammaire de l'oral représenté et d'en étudier la diachronie ; elles visent, d'autre part, à mieux distinguer différents types d'écrits grâce à un ensemble de traits linguistiques internes, en reliant ces traits aux conditions de production des séquences étudiées et au contexte pragmatique de l'énonciation. A terme, ces recherches doivent bien entendu nous renseigner sur les variétés linguistiques qu'il est possible d'étudier en français médiéval et dont les usages observés rendent compte (l'oral représenté représentant un usage linguistique parmi d'autres), mais elles pourront peut-être aussi nous fournir, grâce à l'étude plus fouillée de cette forme d'oral fictif, un lieu d'observation privilégié du changement linguistique.

2.1.1. Spécificités de l'oral représenté

Ces premières recherches, encore très embryonnaires, ont cependant tout de suite montré les très fortes spécificités du discours direct dans les textes médiévaux. Il apparaît tout d'abord que certains phénomènes linguistiques sont limités à ce contexte d'usage particulier : paradigmes lexicaux tels que celui des termes d'adresse par exemple, actes illocutoires

comme l'injonction, l'interrogation directe, la réponse, le refus, etc. Il arrive également que le discours direct soit le contexte d'occurrence privilégié d'une valeur ou d'un effet pragmatique particulier d'une unité linguistique. C'est le cas, pour le démonstratif CIL de l'ancien français par exemple, du rejet de l'allocutaire que provoque l'exclusion du référent hors de la sphère personnelle du locuteur et son association à celle du destinataire.

L'énoncé très célèbre de la *Chanson de Roland* dans lequel Charlemagne rabroue l'archevêque Turpin après que celui-ci s'est porté volontaire pour partir en ambassade me servira d'illustration :

> (1) Li empereres respunt par maltalant :
> « Alez sedeir desur **cel palie** blanc !
> N'en parlez mais, se jo nel vos cumant ! » (*Roland,* v. 271-273)
> *L'empereur répond avec colère : « Allez vous asseoir sur ce tapis blanc ! N'en parlez plus, si je ne vous le commande ! »*

Lorsque Charlemagne s'adresse directement à l'archevêque, l'usage de la série CIL marque clairement la colère de l'empereur qui relègue Turpin sur un vague tapis dont il n'a que faire.

De manière générale, l'oral représenté est un terrain d'observation privilégié de la deixis, même si, comme j'y ai fait allusion dans la section précédente, ce n'est apparemment pas le seul dans les textes médiévaux (voir en particulier l'usage de l'adverbe *or(e)* dans le récit ou le discours didactique).

Mais la grammaire de l'oral représenté se caractérise surtout par une sur- ou une sous-représentation de certains éléments qui sont également présents ailleurs. Il s'agit là, à vrai dire, d'une tendance qui se manifeste plus généralement dans tous les types de discours[5]. Nos recherches passées ont déjà permis d'établir une première liste de caractéristiques typiques du discours direct ([doc. 13] Guillot *et al*. à par. a) : sur-représentation des pronoms (personnels et autres), des possessifs, de la négation, des verbes à l'infinitif, des conjonctions de subordination ; sous-représentation du nom (nom propre et nom commun), des déterminants, de l'adjectif qualificatif, des participes, des conjonctions de coordinations et des adverbes subordonnants (adverbes de l'interrogation indirecte : comment, combien, etc.). Ces premiers résultats semblent indiquer qu'on peut établir une grammaire de l'oral représenté en français

[5] Il est rare en effet qu'un type de discours se distingue par la présence ou l'absence d'une marque. Il se laisse en général mieux décrire par des fréquences particulièrement (et relativement à d'autres) hautes ou basses.

médiéval, dont les caractéristiques paraissent relativement stables tout au long de la période et assez fortement marquées ([doc. 14] Guillot *et al*. à par. b).

Parmi tous les éléments énumérés plus haut, on peut s'étonner de trouver les conjonctions de subordination en si grand nombre dans le discours direct. Ce constat semble d'une certaine façon concorder avec les recherches menées par l'équipe du GARS sur la syntaxe de l'oral et sa relative complexité, y compris dans ses réalisations les moins planifiées (Blanche-Benveniste 1997 : 58-60)[6]. Il faudrait bien entendu mener une étude plus poussée pour confirmer ce point, mais ce premier résultat peut nous conduire à rechercher, parmi tous les traits que j'ai donnés et d'autres encore sans doute passés inaperçus, des caractéristiques intemporelles du discours oral ou des caractéristique qui seraient communes à certaines formes d'oral contemporain et à l'oral représenté des textes médiévaux. S'il s'avérait qu'il est possible d'en dégager une liste, cela confirmerait le statut très particulier de l'oral représenté en français médiéval et donnerait du crédit à l'idée qu'il constitue bien une variété écrite que l'on peut rapprocher de l'oral plus ou moins spontané.

2.1.2. Oral représenté et changement linguistique

On a à plusieurs reprises avancé l'hypothèse que les premières phases des changements se trouveraient dans la langue orale, l'écrit plus conservateur opposant une résistance plus forte aux innovations. De là à penser que l'oral représenté pourrait offrir un stade plus avancé du français, il n'y a qu'un pas. A l'appui d'une telle hypothèse vient le fait que les premières attestations d'un élément ou d'une construction nouvelle s'y rencontrent parfois. Cette idée, très ancienne, peut paraître simpliste ou opportuniste, mais elle est confortée par certains faits. Elle se trouve, sous une forme un peu différente puisqu'il y est question de théâtre et non de discours direct, dans la *Petite syntaxe de l'ancien français* de L. Foulet, à propos de l'article partitif et de ses premières réalisations dans le corpus étudié :

> « *Les exemples les plus nombreux et les plus probants du nouvel article partitif se trouvent dans le* Garçon et l'Aveugle, Courtois d'Arras *et surtout le* Jeu de la Feuillée, *c'est-à-dire dans des œuvres dramatiques. En revanche, il n'y en pas un seul exemple*

[6] Il faut bien entendu se garder de généralisations trop faciles. Une analyse un peu plus poussée montre que certaines conjonctions de subordination sont au contraire sur-représentées dans le récit médiéval. C'est le cas par exemple de *quant,* surtout lorsque la conjonction se trouve en position initiale. Ce résultat n'est pas très étonnant, la fréquence de la structure *quant… si* ayant été déjà amplement mise en évidence dans les récits en prose du Moyen Âge. Il semble que ce soit surtout la conjonction *que* qui apparaisse massivement dans le discours direct. Tous ces points seraient naturellement à reprendre plus en détail.

> *dans la* Chastelaine de Vergi. *Peut-être pourrait-on en tirer une conclusion qui de soi est assez vraisemblable : c'est que l'article partitif au sens moderne a pris naissance dans la langue de la conversation, s'y est développé tout d'abord et n'a pénétré qu'ensuite et graduellement dans la langue littéraire.* » (Foulet 1919/1967 : 83).

Dans ses recherches sur la naissance de l'article partitif en français, A. Carlier reprend la même thèse, même si elle ne centre pas son analyse sur le discours direct. Elle revient sur l'idée que le partitif se développerait en premier lieu dans les contextes marqués par une forme d'oralité :

> « *Nous ferons l'hypothèse que, globalement, ce style plus spontané, plus proche de l'oral permet d'expliquer la plus haute fréquence du partitif dans cette traduction. Ainsi se manifeste la tendance générale selon laquelle les innovations se répandent d'abord à l'oral et dans les textes écrits usant d'un registre proche de l'oral et tendent à être refoulées dans le registre soutenu*. » (Carlier 2004 : 125)

Il serait bien entendu très utile d'examiner en détail les séquences d'oral représenté et d'y chercher des occurrences précoces de l'article partitif en français.
D'autres exemples illustrent de façon plus claire encore le lien qu'on peut établir entre discours direct et changement linguistique. Dans ses recherches sur l'évolution aspectuo-temporelle du passé composé en français, L. Schøsler montre que cette évolution se manifeste en tout premier lieu dans le discours direct :

> « *on constate que les valeurs originelles des temps étudiés (celle du PC comme temps du présent accompli et celle du PS comme perfectum praesens et perfectum historicum) persistent dans le récit, surtout dans le récit en vers, alors que les valeurs innovatrices de ces mêmes formes (celle du PC comme perfectum praesens et celle du PS comme perfectum historicum) se rencontrent dans les parties en discours direct. Cette distribution confirme donc l'analyse selon laquelle les parties du texte qui se présentent comme proches de l'oral sont innovatrices, alors que les parties narratives sont conservatrices.* » (Schøsler 2012 : 328)

Je ne fais que mentionner ici quelques travaux et pistes de recherche bien connus. Il me semble néanmoins que ces hypothèses méritent qu'on s'y arrête et qu'elles justifient à elles seules d'entreprendre une analyse systématique de l'oral représenté et de ses traits distinctifs.
Une telle étude paraît en outre d'autant plus nécessaire que l'on admet la place centrale de la fonction communicative dans le changement linguistique. De ce point de vue, le cadre théorique proposé par Koch et Österreicher (notamment 1990 et 2001) pour préciser et complexifier l'opposition entre oral et écrit me semble fournir des arguments supplémentaires. En proposant une série de critères pragmatiques et situationnels permettant de caractériser le mode de conception du message linguistique, Koch et Österreicher nous donnent les moyens de mieux caractériser l'oral représenté, dans ses différentes formes et par contraste avec le récit. Les critères dont je reprends la liste dans le tableau qui suit permettent de remplacer l'opposition binaire entre oral et écrit par une opposition graduelle entre un pôle de la proximité communicative d'un côté (souvent réalisée grâce au médium oral, mais pas toujours) et un pôle de la distance de l'autre (préférentiellement assuré par le médium écrit) :

Tableau 1 : paramètres mesurant la distance/proximité communicative (Koch & Österreicher 2001 : 586)

1.	communication privée	communication publique	1
2.	interlocuteur intime	interlocuteur inconnu	2
3.	émotionnalité forte	émotionnalité faible	3
4.	ancrage actionnel et situationnel	détachement actionnel et sit.	4
5.	ancrage référentiel dans la situation	détachement réf. de la situation	5
6.	coprésence spatio-temporelle	séparation spatio-temporelle	6
7.	coopération communicative intense	coopération communicative minime	7
8.	dialogue	monologue	8
9.	communication spontanée	communication préparée	9
10.	liberté thématique	fixation thématique	10

La caractérisation de l'oral représenté au moyen de ces critères montre bien qu'un grand nombre de paramètres (communication privée, ancrage actionnel et situationnel, ancrage référentiel, coprésence spatio-temporelle, dialogue) le situent plutôt du côté gauche du tableau, même s'il s'agit toujours d'un ancrage, d'une coprésence, etc., fictifs.
L'un des grands apports des travaux de Koch et Österreicher est également qu'ils permettent d'articuler les paramètres de la proximité communicative que je viens d'énumérer avec les

patrons plus ou moins codifiés des traditions discursives (les domaines et les genres discursifs). Je reviendrai plus en détail sur ce point dans le chapitre 2.

Je me contenterai de souligner ici la dissociation clairement établie dans le tableau 1 entre le monologue et le dialogue. Les critères pragmatiques définis par Koch et Österreicher semblent en effet fortement rapprocher (dans le monologue) la représentation des paroles d'un locuteur unique (en principe à la première personne) de la narration proprement dite (souvent à la troisième personne). Cette distinction entre monologue et dialogue me paraît essentielle pour différentes raisons, et en tout premier lieu parce que ces deux modalités communicatives semblent se situer aux deux extrémités du spectre de la proximité/distance conceptuelle. Mais son importance tient aussi au fait que la présence de plusieurs locuteurs dans le dialogue (ou dilogue) et le passage de l'un à l'autre induisent un mode d'organisation et de structuration du discours très particuliers. L'étude de ces modes de structuration – encadrement de l'ensemble de l'épisode d'oral représenté et organisation interne de l'alternance des tours de parole – me paraît constituer un objet d'étude très important pour notre connaissance de l'oral représenté mais aussi, de manière plus large, pour l'analyse des modes d'organisation textuelle du document écrit au Moyen Âge. En cela, l'étude de la structuration de l'oral représenté gagnerait certainement à être menée dans une approche contrastive et par comparaison avec celle d'autres types d'écrit.

2.2. L'écrit et la linguistique diachronique

Le statut original de l'oral représenté et son signalement explicite dans les manuscrits médiévaux, son importance pour l'analyse du changement linguistique et ses spécificités linguistiques ont été soulignés dans la section précédente. Ces éléments semblent confirmer qu'il s'agit bien d'un type d'écrit très particulier, qui mérite une analyse détaillée et dont les rapports avec l'oral réel restent encore largement à préciser pour la période médiévale.

Je voudrais défendre à présent une approche complémentaire, centrée sur les caractéristiques du texte écrit tel qu'il nous a été transmis par la culture manuscrite médiévale, sur ses particularités et ses contraintes propres et sur ses apports pour la linguistique diachronique.

2.2.1. Ecrit et changement linguistique

> *« La conviction que la langue parlée influence la langue écrite, au risque d'y laisser infiltrer des fautes typiques de l'oral, est souvent admise sans examen. L'inverse semble pourtant tout aussi probable. »* (Blanche-Benveniste 1997 : 147)

L'influence que l'écrit peut prendre dans l'évolution linguistique dépend naturellement de la place qu'il occupe dans la culture et la société. Et cette place est elle-même susceptible d'évoluer, d'autant plus fortement que la période envisagée est longue : sept siècles séparent les premiers textes écrits en français au 9ème siècle et la fin du Moyen Âge, ce qui excède la durée qui sépare le français du 21ème siècle de celui du 16ème siècle. La période médiévale offre un terrain d'étude privilégié puisqu'elle voit l'émergence, le développement et la codification progressive d'une vaste littérature en langue vernaculaire, les variétés régionales de la langue d'oïl donnant peu à peu naissance à une langue écrite davantage standardisée qui supplante le latin d'un côté, la langue d'oc de l'autre (l'ordonnance de Villers-Cotterêt de 1539 et le développement des premières grammaires françaises au 16ème siècle sont finalement le point d'aboutissement de cette lente et longue évolution).

On a observé parallèlement que les premières œuvres écrites dans les langues romanes, et spécialement en français, semblent être fortement marquées par une forme d'oralité (voir notamment les nombreux travaux de P. Zumthor sur le sujet). Les recherches de P. Koch sur le passage à l'écrit des principales langues romanes l'amènent ainsi à proposer quatre « constellations communicatives fondamentales » (Koch 1993) pour les débuts de l'écrit en langue vernaculaire : « l'oralité mise par écrit » (catégorie peu étendue en général et non représentée pour la langue d'oïl), les listes (tarifs de péage de Sens, par exemple, mais peu de documents en français), la « scripturalité à destin vocal » (catégorie très vivante, voir plus loin) et les documents « comportant des tensions ou des contrastes linguistiques qui impliquent très souvent une prise de conscience métalinguistique ou métacommunicative » (glossaire de Tours au 12ème siècle, traduction en anglo-normand dans la deuxième moitié du 12ème siècle des *Quatre livres des rois*, etc.).

La catégorie intitulée « scripturalité à destin vocal » mérite une attention particulière, notamment parce qu'elle rassemble la grande majorité des premiers textes français. Elle se distingue aussi par le fait qu'elle témoigne d'une avancée significative du français dans le

domaine de la distance communicative. Mais ces nouvelles formes de discours en langue vernaculaire restent en même temps étroitement liées aux conditions de l'oralité, puisqu'elles sont toutes conçues pour donner lieu à une forme de performance vocale à l'adresse d'un public non lettré. Qu'il s'agisse des formules de serment, des dépositions de témoins, des bénédictions, et surtout, des sermons, de la poésie religieuse (*Cantilène de sainte Eulalie*, *Vie de saint Alexis*, etc.), du théâtre religieux (*Jeu d'Adam*), des chansons de geste (*Chanson de Roland*) ou de la poésie profane des trouvères, la fonction de tous ces textes est indissociable de leur réalisation sous forme orale (fonctions rituelle, édificatrice et/ou poétique).

L'étape suivante permet le développement de discours écrits en français qui se détachent peu à peu de la performance, même si les habitudes de lecture (à haute voix et souvent collective) ne les coupent pas complètement de la voix humaine. On peut évidemment se demander dans quelle mesure cette scission progressive d'une partie de l'écrit avec les conditions pragmatiques qui caractérisent l'oralité accentue la différenciation interne des types de discours. On peut supposer que l'écart se creuse au fur et à mesure que la culture écrite vernaculaire prend son essor. Devenu plus autonome, le français écrit peut dès lors exercer sa propre influence sur le système linguistique.

Outre ces éléments, qui sont déjà assez bien connus, on peut ajouter que les modifications techniques et matérielles du document et du support écrits ont pu avoir, elles aussi, des répercussions sur certains changements linguistiques. Les modes de linéarisation, de structuration et de progression discursive de l'écrit évoluant avec le temps, la disposition du texte et de ses différentes parties à l'intérieur du codex change peu à peu. On peut supposer que ces changements graphiques et de mise en page, qui semblent à première vue externes aux données linguistiques, s'accompagnent d'évolutions internes. Si l'oral constitue bien un lieu privilégié d'étude du changement linguistique, il est donc possible en même temps que certains faits, totalement dépendants du médium écrit, lui échappent. Cela implique qu'on tienne également compte de ces facteurs dans l'étude du changement linguistique.

Il n'est peut-être pas inutile de rappeler ici brièvement les grandes lignes de l'évolution matérielle des manuscrits médiévaux. Cette évolution dépend partiellement de la fonction des textes rassemblés dans le codex (textes littéraires à but récréatif, textes scolastiques, traités scientifiques, œuvres historiques, etc.) et du type de manuscrit considéré (manuscrit d'apparat ou d'étude, manuscrit de grande ou de petite taille, manuscrit destiné à un public de lettrés habitués à certains codes de l'écrit ou à un public de laïcs moins entraînés, etc.)[7]. Elle dépend

[7] Les caractéristiques de la mise en page semblent, en revanche, être relativement indépendantes de la région où le manuscrit a été réalisé (Hasenhor 1990).

aussi de la langue utilisée (les manuscrits latins sont plus fortement abrégés que les manuscrits français, par exemple, cf. Hasenohr 1990 : 232-233 et Hasenohr 2002), même s'il apparaît malgré tout que les textes vernaculaires adoptent globalement les habitudes graphiques des textes latins[8]. En réalité, ce sont surtout les types de textes écrits dans les deux langues qui expliquent pendant longtemps les différences apparentes (les textes scientifiques, universitaires et scolastiques sont écrits en latin, comme le sont la plupart des manuscrits de travail et d'étude)[9].

Parmi tous les changements qui modifient la mise en page des livres manuscrits, je m'intéresserai surtout ici aux éléments de repérage qui s'insèrent peu à peu à l'intérieur ou à la marge du texte. Ces éléments sont de deux sortes et servent deux fonctions principales : soit ils aident à naviguer dans le texte et en permettent une lecture fragmentée et partielle (index, concordances, répertoires, tables des matières et des incipits, titres courants, etc.), soit ils marquent la succession et la hiérarchisation des subdivisions internes au texte dont ils facilitent la lecture continue (enluminures, lettrines de tailles diverses, enluminées ou historiées, rubriques de toutes sortes, pieds de mouche, variations dans le type et le calibre des lettres, etc.). Les éléments du premier type se développent surtout dans les manuscrits à fonction scientifique et universitaire (Bibles, traités et commentaires scolastiques, recueils didactiques divers) et sont utilisés, pour cette raison, dans les textes latins essentiellement. Ceux du second type vont peu à peu être importés dans la prose vernaculaire, l'importance qu'y prennent les illustrations et enluminures révélant la fonction souvent complexe de ces éléments (fonction d'organisation textuelle mais aussi fonction esthétique)[10].

On sait que le livre reste pendant tout le Moyen Âge un objet de luxe, très coûteux et relativement rare, même si la production croît de façon très significative à partir du 13ème siècle et que les procédés de fabrication évoluent dans le même temps (Bozzolo & Ornato 1980). Parmi ces procédés, le système de la *pecia*, mis en place dans le milieu universitaire parisien, découpe le livre en cahiers réalisés simultanément et rassemblés en bout de chaîne. Il

[8] Il s'agit d'un des nombreux aspects du contact linguistique qui s'opère entre le latin et les langues vernaculaires tout au long du Moyen Âge (voir la section suivante). De ce point de vue, il est peut être utile de distinguer les procédés de mise en page, largement inspirés des modèles latins, et les « principes de transcription » (Hasenhor 1990 : 232), qui semblent plus perméables à la langue. Les remarques faites par C. Ruby sur les systèmes de notation, d'abréviation et de ponctuation en partie divergents dans les parties françaises et latines des premiers psautiers bilingues vont dans le même sens (Ruby 2010).

[9] De manière générale, les livres d'étude sont ceux dont l'appareil méta-textuel (voir Genette et la notion de « péritexte ») est le plus riche et le plus diversifié. Les gloses et systèmes de citation des autorités, particulièrement développés au Moyen Âge, en font naturellement partie.

[10] Dans les textes littéraires par exemple, les rubriques sont à la charge de l'enlumineur et non du copiste. G. Hasenhor montre d'ailleurs très bien comment elles évoluent au fil du temps : souvent liées à l'enluminure dont elles indiquent au départ la légende, elles sont peu à peu intégrées au texte et jouent le rôle de titre.

permet la réalisation rapide d'un grand nombre d'ouvrages, mais il impose en même temps un système de repérage interne des différents cahiers qui soit suffisamment développé pour permettre l'assemblage final. Cet exemple de la *pecia* montre lui aussi que l'évolution de la mise en page médiévale est tributaire des aspects du livre les plus divers (aide à la lecture, agrément esthétique, procédés de fabrication matérielle).

Ce sur quoi il est peut-être utile d'insister ici est le caractère apparemment très compact du folio médiéval. Alors même que se développent un grand nombre de procédés d'aide au repérage inter- et intra-textuel (dont les principaux ont été énumérés plus haut), les copistes semblent animés, d'un bout à l'autre du Moyen Âge, par une même horreur du vide[11]. Ces deux tendances conjuguées pourraient peut-être expliquer, de manière partielle au moins, que se mettent également en place des éléments de repérage internes au texte-même. Ces éléments ne se signalent pas tous par une mise en forme particulière, mais ils jouent un rôle dans l'organisation discursive et guident le lecteur dans sa lecture et son parcours textuels.

Certains de ces éléments ont déjà été mis en évidence et étudiés : adverbe démonstratif *(i)ci*, utilisé dans les formules marquant la fin et/ou le début d'un épisode et peu à peu intégré dans les titres à la fin du Moyen Âge (Perret 1988), éléments de rappel du type *a cest mot* marquant la fin du discours direct et la reprise du récit (Marchello-Nizia 2012), formules et syntagmes utilisés de manière répétitive dans la prose ([doc. 2] Guillot 2004), cadres de discours ouvrant une séquence discursive d'étendue plus ou moins vaste (*ce vendredi… ceste meisme nuit… le samedi au matin…*), etc. Il semble que la liste de ces éléments évolue et que certains d'entre eux changent de valeur au fil du temps (c'est ce que j'ai essayé de montrer en étudiant l'évolution sémantique des démonstratifs et le développement de leur valeur spatiale à la fin du Moyen Âge).

L'étude de ces éléments implique qu'on s'intéresse à leur disposition matérielle et à leur mode d'inscription dans la linéarité du texte, leur fonction principale étant de baliser le document et de servir de frontière discursive. Mais, comme je l'ai dit plus haut, ils ne se distinguent pas toujours par leur aspect graphique. Ils partagent néanmoins avec les procédés de mise en page évoqués plus haut une même fonction topologique. Il semble dès lors légitime de les étudier ensemble. Et l'on peut faire l'hypothèse que le développement de ces procédés de différents types entraîne (révèle ?) un changement dans les modes de lecture, le

[11] Le coût du support (parchemin surtout) n'est évidemment pas pour rien dans cette tendance à remplir le moindre espace. Mais il n'explique pas tout (la tendance, par exemple, à remplir la fin d'une ligne restée sans texte par un trait horizontal). Ce propos doit toutefois être nuancé, les espaces et alinéas étant régulièrement exploités pour grouper et/ou dissocier les unités discursives (groupes de vers, notamment, dans les textes en vers).

passage d'une lecture pour les oreilles à une lecture, qui, à défaut d'être parfaitement ou dans tous les cas silencieuse, s'adresse davantage aux yeux[12].

Les quelques éléments rappelés ici me semblent ouvrir des directions de recherche utiles pour la linguistique diachronique et appuyer l'idée que le développement et la structuration progressive de l'écrit en langue vernaculaire ont sans doute permis un renouvellement linguistique et l'émergence de nouveaux usages et outils. L'apparition des marqueurs de topicalisation en français pourraient également confirmer cette hypothèse. Les travaux de B. Combettes ont clairement montré que leur développement et leur grammaticalisation en français étaient intimement liés à leur usage dans le discours argumentatif (Combettes 2001 : 111, Combettes & Prévost 2003[13]). Il n'est évidemment pas facile de déterminer si ces éléments sont apparus d'abord dans le discours argumentatif écrit ou oral ou s'ils ont trouvé dans l'un des deux médiums un contexte plus favorable à leur expansion. On peut toutefois supposer que le développement de l'écrit argumentatif en langue vernaculaire a joué un rôle dans leur histoire.

Le discours scientifique offre lui aussi un terrain d'exploration particulièrement fécond pour étudier ces relations. On s'est jusqu'ici beaucoup intéressé au lexique spécialisé et technique contenu dans les textes de ce type, mais assez peu à d'autres phénomènes linguistiques. Les travaux réalisés par S. Bazin-Tacchella sur la traduction française au 15ème siècle du grand traité de chirurgie de Guy de Chauliac (*Chrirurgia Magna*), et, plus spécifiquement, ses études sur les marqueurs de topicalisation (Bazin-Tacchella 2007) et le fonctionnement de *lequel* dans ce texte (Bazin-Tacchella 2005), montrent bien qu'il s'agit là d'un champ de recherche très riche et trop peu exploré encore. Il serait intéressant d'étudier la fonction des connecteurs dans ces textes par exemple, et de la comparer à celle qu'ils assurent dans la prose narrative.

Au terme de ce parcours rapide, je voudrais défendre à nouveau une linguistique qui tienne compte non seulement des locuteurs mais aussi des conditions matérielles et pragmatiques dans lesquelles les productions langagières se réalisent. A l'écrit, ces conditions pragmatiques se matérialisent principalement par le support graphique et visuel du document physique. Dans le même ordre d'idée, il me semble nécessaire de mesurer toute l'importance, et ses conséquences pour la description linguistique, des contacts et échanges linguistiques qui ont

[12] Il est possible que le développement de la valeur topologique de ces éléments soit lié aussi au passage progressif d'une cohérence resserrée, caractéristique de l'ancien français, à une cohérence beaucoup plus large en moyen français. Combettes (2007 : 41) fait l'hypothèse que cette évolution implique une gestion de la mémoire à plus long terme.

[13] Ce second article montre également très bien les parallèles qui s'établissent entre structures françaises et latines.

marqué l'histoire du français dans tous ses aspects (et pas uniquement dans le domaine lexical). Or, au Moyen Âge, ces contacts linguistiques concernent aussi, et peut-être surtout, les variétés écrites du français.

2.2.2. Ecrit et contacts linguistiques avec le latin

La période médiévale a cela de très particulier qu'un contact linguistique intense s'y maintient de façon continue entre la langue vernaculaire et le latin. Cette situation d'échange durable a eu une influence déterminante sur la langue d'oïl, sous sa forme écrite essentiellement mais sans doute pas exclusivement, comme on le verra plus loin.

Si nul ne remet en cause l'imprégnation profonde du français par le latin à toutes les époques et pas seulement durant la période la plus reculée de son histoire (comme le montre notamment la latinisation des graphies dans les manuscrits de la fin du Moyen Âge), les avis divergent sur la nature du rapport entre ces deux langues. La notion de *diglossie*, qui repose sur une spécialisation des usages et sur la distinction entre une variété haute (le latin) et une variété basse (le vernaculaire), fait débat pour le Moyen Âge. Elle a récemment été reprise par S. Lusignan (2012) dans son étude sur le français picard au Moyen Âge mais dans une perspective renouvelée.

L'histoire du picard révèle en effet une situation paradoxale. Le sentiment des locuteurs médiévaux est très net, le latin constitue bien la norme haute (langue sacrée, langue savante, langue de la diplomatie et langue universelle), le français la variété basse (langue des illettrés, avant tout orale et variable). En ce sens, le concept de *diglossie* semble assez bien refléter la conscience linguistique de l'époque[14]. Mais ce que montrent en même temps les recherches de Lusignan, c'est qu'au contact du latin et dans certaines de ses *scripta régionales* (dont le picard offre l'un des plus beaux exemples), le français écrit se constitue peu à peu comme une variété haute de la langue vernaculaire :

[14] A ceci près toutefois, que la définition de la diglossie que donnent Ferguson et ses continuateurs repose sur la distinction entre une variété haute et une variété basse d'une même langue. Or les chercheurs admettent à peu près tous que la Réforme carolingienne scelle clairement dans l'esprit des locuteurs la scission entre le latin et le français (qui commence peu après à s'écrire). La notion de *diglossie* semble en revanche s'appliquer assez bien à la période antérieure à la Réforme carolingienne, à partir du moment où la distance linguistique entre la langue de communication usuelle et la variété haute du latin s'accentue, où le prestige, le degré de standardisation, le mode d'acquisition, etc., de ces deux variétés s'opposent. Pour une étude approfondie de ces questions, voir notamment Koch (2008).

> *« Mais entre le xii^e^ et le xv^e^ siècle, il s'est formé un registre lettré du français qui, s'il n'était pas encore l'égal du latin, ne saurait être caractérisé comme une langue basse. Celui-ci obéissait à des normes complexes et certains documents trahissent une véritable conscience grammaticale du français écrit de la part du scripteur. Une telle évolution s'est effectuée au fil d'un contact soutenu entre le français et le latin. Ce rapprochement entre les deux eut pour conséquence que, plus prestigieux, le latin influa sur l'évolution du français. Contact linguistique et changement linguistique vont souvent de pair. Ce rapport entre le latin et le français écrit a vraisemblablement conduit à une situation triangulaire où il y avait la langue de prestige par excellence, le latin, une langue dont la valeur ne cessait de s'accroître, le français lettré, et une langue basse, le français parlé par les illettrés. C'est à titre d'hypothèse que nous attribuons à ce dernier un statut de langue basse, car aucune source médiévale ne permet de percevoir quel fut le rapport exact entre le français lettré et celui parlé par la population illettrée. »* (Lusignan 2012 : 38).

Cette nouvelle approche permet de dépasser une vision simpliste qui opposerait la langue écrite haute (le latin) et la langue orale basse (le français). On savait depuis longtemps que le latin était resté une langue parlée tout au long du Moyen Âge, une langue bien vivante même si elle n'était plus la langue maternelle de personne[15]. Par ailleurs, il est clairement manifeste que le développement de l'écrit en langue française lui donne peu à peu accès à des registres auparavant réservés à la norme haute (textes scientifiques, textes juridiques, etc.). Mais ce sur quoi les recherches de S. Lusignan permettent d'insister, c'est sur le fait que l'essor du vernaculaire écrit se fait « en symbiose » avec le latin et le conduit à en suivre très étroitement les modèles existants :

> *« On peut même soutenir que la langue vernaculaire s'est constamment développée en symbiose avec la langue savante. »* (Lusignan 2012 : 119).

L'idée que la société médiévale serait partagée entre d'un côté des clercs qui n'utilisent que le latin et de l'autre une population illettrée qui ne comprend que le vulgaire est partiellement fausse, comme en attestent les remarques faites par les premiers sur les spécificités du français

[15] Contrairement à l'image qu'en donnent les clercs médiévaux, cette langue continue d'ailleurs d'évoluer au cours du Moyen Âge. On peut cependant hésiter à la qualifier de langue vivante. Banniard (1980 : 110) considère qu'à partir du 8ème siècle le latin a perdu son caractère de « langue de communication générale » et est devenu une langue morte, même si on continue de le parler pendant des siècles.

(voir *infra*) et comme en témoignent aussi les nombreux manuscrits, notes préparatoires de sermons par exemple[16], qui mêlent les deux langues et font conclure à une certaine imprégnation des moins éduqués par la langue savante[17]. Ce qui paraît évident en revanche, c'est que les clercs formés à la langue savante, ces professionnels de l'écriture qui maîtrisent suffisamment les deux langues pour passer de l'une à l'autre, sont principalement à la source des échanges linguistiques incessants.

Il est même permis de faire l'hypothèse que l'influence du latin dépasse le cadre strict de la langue écrite. L'Université de Paris, dont l'attachement au latin ne fait pas de doute, aurait eu, selon Lusignan, une influence déterminante sur l'émergence progressive d'une rhétorique en langue vernaculaire (liée à la prédication) et sur le développement d'une littérature en scripta francienne :

> *« Certains ont déjà souligné l'apparition plus tardive du français central par rapport à l'anglo-normand ou au picard. On sait peu de choses des conditions qui ont favorisé sa naissance, mais l'histoire littéraire nous désigne des œuvres écrites en milieu universitaire à titre de premiers monuments. Rien ne contredit l'hypothèse que le milieu lettré universitaire ait pu offrir un cadre favorable à la mise en forme de la scripta parisienne. »* (Lusignan 2012 : 119)

Dans ses recherches antérieures (notamment Lusignan 1987), S. Lusignan avaient déjà montré combien la conception qu'avaient les médiévaux de leur propre langue était intimement liée à l'enseignement de la grammaire latine. Si le français est considéré tout au long du Moyen Âge comme une langue orale acquise dans le cadre familial et s'oppose au latin, langue seconde apprise à l'école et régie par des règles grammaticales, la réflexion linguistique sur la langue vernaculaire progresse peu à peu à la faveur de l'enseignement du latin, qui se fait au départ en français. Même si cette activité métalinguistique ne donne pas encore lieu à la rédaction de grammaires françaises[18], elle témoigne d'une prise de conscience de certaines des spécificités

[16] L'exemple le plus célèbre est le *Sermon sur Jonas*.

[17] Plusieurs des recherches présentées dans Le Briz & Veysseyre (2010) portent sur l'imbrication du français et du latin dans les manuscrits médiévaux et sur les conclusions qu'il est possible d'en tirer sur la nature du bilinguisme médiéval. L'article de N. Bériou s'intéresse tout particulièrement aux traces écrites des sermons et confirme que « le bilinguisme y règne, surgissant au détour de maintes phrases à la mesure de la place ménagée à l'autorité scripturaire – et accessoirement à d'autres citations – dans le discours de tout prédicateur. » (Bériou 2010 : 191)

[18] Hormis à partir du 13ème siècle en Angleterre, où l'anglo-normand devient, comme le latin, une langue seconde apprise à l'école.

de la langue vulgaire (présence de l'article, de la double négation, etc.)[19]. L'influence du latin semble ainsi opérer non seulement sur la langue écrite et parlée dans la France septentrionale, mais aussi – même si cela doit être nuancé – sur l'image que les clercs médiévaux développent de leur propre langue maternelle.

D'autres chercheurs, C. Buridant en particulier (Buridant 2011, 2012 et à par.), ont bien mis en évidence aussi l'influence des traductions d'œuvres latines sur les langues romanes et sur le français en particulier. Ces traductions, qui se multiplient à partir du 13ème et surtout aux deux siècles suivants, conduisent à un renouvellement très profond du lexique. Elles façonnent également la prose en langue française, selon deux phases successives. Dans un premier temps, elles tendent à privilégier le fractionnement de la structure syntaxique (par une sorte de « dépliage » de la syntaxe latine), favorisent la coordination et la progression à thème constant ou linéaire, multiplient les binômes synonymiques, etc. Les travaux de Buridant révèlent à quel point ces premières traductions donnent à la prose française certains de ses traits les plus remarquables :

> « *C'est ainsi que dans le genre narratif en prose de l'ancien français – chroniques ou romans –, un ensemble de traits majeurs se dégagent des traductions, qui rejoignent en les accentuant les traits stylistiques de la prose originale*. » (Buridant 2012 : 16)

Ces traits perdurent jusqu'à la fin du Moyen Âge. Mais certaines translations de la seconde moitié de la période médiévale, sans doute dans le souci de suivre plus fidèlement l'original latin, tendent au contraire à complexifier la structure syntaxique. Elles hiérarchisent davantage les unités syntaxiques et préfèrent l'hypotaxe à la coordination. Ces nouvelles tendances, qui se développent au départ surtout dans les traductions de traités scientifiques et didactiques (Buridant 2011 : 119-120), se répandront peu à peu elles aussi dans les textes de la fin du Moyen Âge. Il semble donc que les innovations et les habitudes de traduction influent durant toute la période médiévale sur l'évolution de la prose française.

L'étude de ces œuvres et des prologues des traducteurs fait par ailleurs conclure C. Buridant au sentiment très vif de parenté unissant toujours le latin au français dans l'esprit des clercs médiévaux, si bien qu'il se demande s'ils n'y voyaient pas plutôt deux registres distincts d'une seule et même langue :

[19] L'étude de M. Colombo-Timelli (1996) sur les traductions françaises de l'*Ars minor* de Donat, qui enseignaient aux débutants le latin et la grammaire (latine mais souvent expliquée et illustrée par le français), appuient son analyse.

> « *Si traduire, c'est nécessairement mettre à distance la langue-source, dans un mouvement de transfert, quelle est la distance qui sépare, chez les clercs, le latin du vernaculaire ? N'est-elle pas une distance* a minima, *dans une communion de deux langues ou de deux registres de langue ? Dans les premiers textes romans autonomes en particulier, qui s'inspirent de modèles latins, ne peut-on parler de traduction implicite ?* » (Buridant à par.)

S'il paraît difficile de prétendre qu'après le 9ème siècle le latin et le français sont encore conçus comme deux formes différentes d'un même idiome, il est indéniable en même temps que ces deux langues restent consubstantiellement liées dans l'esprit des locuteurs et qu'elles évoluent ensemble. Les traditions discursives et les relations intertextuelles ne connaissent pas de frontière entre elles et doivent, par conséquent, s'envisager en diachronie dans les deux langues (voir le chapitre 2, section 2.2.).

La multiplication des traductions dans la seconde partie du Moyen Âge témoigne naturellement d'une extension de la langue vernaculaire, le latin des auteurs classiques devenant au fil du temps de plus en plus obscur à un grand nombre de lettrés (Monfrin 1964 : 18-20). Elle puise aussi sa source dans l'intérêt grandissant des laïcs pour les textes philosophiques et scientifiques de l'Antiquité païenne, les commanditaires de ces traductions étant le plus souvent de grands seigneurs moins bien formés à la langue savante. Mais ce mouvement très ample de translation conduit du même coup à une sorte de « relatinisation » du français. L'étude des néologismes introduits durant cette période montre qu'ils sont dans leur quasi-totalité construits sur une base latine et que, le plus souvent, leur sens ne pouvait être accessible aux lecteurs qu'à condition qu'ils connaissent déjà leur sens latin (Duval 2010). Il n'est d'ailleurs pas toujours aisé de déterminer dans les textes de cette période si une unité lexicale est latine ou française, les traductions et textes de vulgarisation scientifique ayant souvent pour caractéristique de mêler les deux langues et d'expliquer le sens de certains mots spécialisés par le renvoi à leur origine latine. De la même façon, certaines tournures très latines, l'ablatif absolu, par exemple, ou la proposition infinitive dite savante (non régie par un verbe de perception), passent alors dans la prose française.

Ces échanges linguistiques continus et intenses entre le latin et le français produisent donc des effets durant toute la période médiévale et à tous les niveaux de la structure linguistique : niveau graphique, niveau lexical, niveau syntaxique, niveau discursif et textuel. Dans ce dernier cas, le rôle des traditions discursives latines et de leur adaptation en français doit être particulièrement souligné (voir le chapitre 2).

Il ressort de tout ce qui précède que l'écrit en langue vernaculaire ne saurait être dissocié de l'écrit en latin, le français investissant progressivement dans la société médiévale les fonctions et les modèles façonnés par la tradition latine. On peut dès lors se demander dans quelle mesure il est possible d'étudier une langue (le français ou le latin) de manière indépendante de l'autre, même si dans la pratique on est souvent obligé de le faire. De même que dans le domaine de la variation dialectale on parle de *scripta* écrites, sortes de *koinés* qui nivellent les particularismes locaux afin d'être comprises par tous mais qui n'étaient sans doute parlées par personne sous la forme qu'on leur connaît, de même il ne paraît pas totalement absurde de considérer que les textes écrits en français reflètent un jeu de normes et de contraintes qui leur sont propres et les distinguent fortement de la langue de communication usuelle. L'élaboration progressive de ces normes ne peut être dissociée des échanges linguistiques constants avec le latin. Le français écrit durant cette période pourrait dès lors être considéré comme une sorte d'artefact qui prend peu à peu de l'ampleur et acquiert une existence et un développement propres.

Ce premier chapitre s'est attaché à montrer l'utilité d'une approche de la langue et des textes médiévaux qui soit centrée sur les phénomènes énonciatifs ou qui tienne compte de ces aspects dans l'analyse synchronique et diachronique des énoncés (but communicatif du locuteur, position du locuteur et de l'allocutaire, etc.). Il visait également à défendre une recherche linguistique qui s'appuie sur l'étude du contexte historique, socioculturel et matériel dans lequel ont été réalisés les textes nous donnant accès aux ressources linguistiques.

Je tenterai de prolonger cette réflexion dans le chapitre qui suit en abordant sous un angle complémentaire, plus méthodologique et parfois relativement technique, la question des corpus linguistiques, de leur rôle, de leurs apports et leurs limites dans l'analyse des données linguistiques, tout spécialement dans le cadre des études diachroniques.

Chapitre 2 : Linguistique diachronique et méthodologie de corpus

Ce second chapitre porte sur les aspects de la méthodologie de corpus qui me paraissent les plus centraux pour la recherche linguistique diachronique. Ce sont en tout cas ceux auxquels j'ai attaché la plus grande importance dans mes études individuelles et collectives. Les quelques réflexions qui suivent s'inspirent donc de ma pratique des corpus, comme simple utilisatrice, mais elles reposent aussi sur les expériences que j'ai acquises dans la production de corpus numériques diachroniques, ces deux volets du traitement des corpus étant de plus en plus intriqués au plan théorique et étant également de plus en plus souvent assurés par les mêmes personnes (voir le chapitre 3). Bien que mes expériences personnelles se limitent aux corpus écrits – et tous les exemples que je prendrai illustreront les questions posées par ce type de corpus – plusieurs des points abordés plus bas me semblent avoir cependant une portée plus générale.

1. Le corpus d'analyse

La notion de *corpus* a naturellement fait l'objet d'une réflexion particulièrement approfondie dans le cadre de la linguistique de corpus. J'insisterai dans ce qui suit sur les spécificités de la définition qu'elle en donne et sur les concepts de *représentativité*, de *variation externe* et *interne* et d'*équilibrage du corpus*.

1.1. Définition du corpus

Comme je l'ai dit plus haut, mes travaux de recherche se situent dans le cadre général de la linguistique ou méthodologie de corpus, telle qu'elle est définie et appliquée à l'anglais notamment par Biber (1988, 1995 et 1998) et Mac Enery (notamment Mac Enery & Wilson 2001, Mac Enery & Hardie 2012), et, pour l'étude des expressions référentielles, par Botley et Mac Enery (Botley & Mac Enery 2000, Botley 2001 et 2006). Comme un grand nombre d'études menées à l'heure actuelle, ces recherches se basent sur des données attestées, non forgées par le linguiste, mais elles se distinguent de beaucoup d'autres en ce qu'elles supposent une définition précise et rigoureuse du corpus. La définition de J. Sinclair, reprise par Habert *et al*. (1997) me servira de point de départ :

> *« Un corpus est une collection de données langagières qui sont sélectionnées et organisées selon des critères linguistiques explicites pour servir d'échantillon du langage. »* (Habert *et al.*, 1997 : 144)

Le corpus ainsi conçu ne doit pas être confondu avec un simple réservoir à attestations, ce à quoi il se réduit parfois dans certains travaux. Il s'agit d'une collection de données d'un volume suffisamment important pour pouvoir servir d'échantillon du langage, ce qui lie de manière indirecte cette méthodologie de recherche aux outils numériques[20] ; d'autre part, ces ressources linguistiques servent de base à l'analyse et circonscrivent la « langue » qui est étudiée. On verra plus loin que cette langue, ou plutôt ces variétés linguistiques, sont souvent prédéterminées grâce à des facteurs extralinguistiques. C'est pourquoi la seconde définition du *corpus* proposée par Habert (2000) me paraît à la fois plus précise (elle ajoute la mention de ces facteurs extralinguistiques) et plus restrictive que celle de Sinclair. Elle remplace en effet le terme de *langue* par le concept plus circonscrit et plus précis d'« emplois déterminés d'une langue », qui correspond mieux à la notion de *langue dans ses usages* (*language in use*), telle qu'elle est revendiquée et étudiée par les linguistes de corpus[21] :

> *« Un corpus est une collection des données langagières qui sont sélectionnées et organisées selon des critères linguistiques et extralinguistiques explicites pour servir d'échantillon d'emplois déterminés d'une langue ».* (Habert 2000 : 13)

Cette définition du corpus insiste sur le fait que les ressources qui le constituent doivent être sélectionnées et organisées selon des critères *explicites*. Ces critères peuvent évidemment varier en fonction des objectifs de recherche qu'on s'est fixés. Si l'on souhaite étudier ce qui distingue le français oral du français écrit, par exemple, cela suppose que le corpus soit constitué et organisé de manière à permettre de contraster ces deux dimensions de la variation langagière.

La définition de critères de sélection et d'organisation explicites implique également une certaine *connaissance* des données rassemblées dans le corpus. On sait que, dans la pratique,

[20] Le corpus correspond en effet à un ensemble de ressources généralement bien supérieur à ce que l'on peut traiter « à la main » et l'on constate une inflation continue dans la taille des corpus disponibles pour la recherche.

[21] A titre d'exemple, on peut citer les premiers mots de Biber (1995 : 1) : « *Variability is inherent in human language* ». Comme on le verra plus loin, les notions d'*usage* et de *variation* sont au centre de l'approche de corpus.

la facilité d'accès conditionne souvent le choix des ressources, mais cette condition est nécessaire à l'interprétation des résultats fournis par le corpus. Elle s'accompagne de la nécessité de *documenter* le corpus et le détail des données qu'il incorpore. C'est parce qu'il peut décrire l'origine, la qualité et les conditions d'usage des données qu'il exploite que le chercheur pourra évaluer de quels usages particuliers ces données rendent compte.

La méthodologie de corpus offre ainsi un cadre particulièrement adapté à la recherche diachronique, non pas uniquement parce que la distance entre les états anciens de la langue et ses usages actuels interdit au linguiste de prétendre à la compétence linguistique et l'oblige à puiser dans les documents du passé les données qu'il analyse, mais aussi et surtout parce que cette méthodologie permet de traiter au mieux la variation linguistique. Or la recherche diachronique suppose par définition l'étude de l'apparition, de la coexistence et de la disparition de certaines variantes permettant le changement linguistique. En outre, les médiévistes, plus que d'autres diachroniciens sans doute, ont de tout temps été particulièrement sensibles à l'hétérogénéité des usages transmis dans les documents médiévaux. Cette hétérogénéité saute immédiatement aux yeux de tous ceux qui, étudiants ou curieux, approchent ces textes et elle en rend l'accès parfois difficile au lecteur d'aujourd'hui. Pour toutes ces raisons, la méthodologie de corpus offre un cadre « par nature » très séduisant et utile aux médiévistes s'intéressant à la diachronie.

1.2. La représentativité du corpus

La notion de *représentativité*, qui se trouve au centre de l'approche de corpus, est étroitement liée à celle de *variation*. En établissant un lien entre les données et les usages qu'elles représentent, elle permet au chercheur d'identifier et de sélectionner les paramètres de variation sur lesquels il travaille, la variation diachronique n'étant qu'une dimension parmi beaucoup d'autres. La définition de la représentativité donnée par J. Sinclair et du rôle central qu'elle joue dans la constitution du corpus servira à nouveau de base à notre réflexion :

> *« Corpus builders should strive to make their corpus as representative as possible of the language from which it is chosen »* (Sinclair 2005 : 2)

Pour constituer un échantillon de langue utile à la description linguistique, le corpus doit être représentatif de la langue, ou des usages, dont il prétend rendre compte.

La notion de *représentativité* peut s'entendre de plusieurs façons différentes. Par défaut et sans autre précision, elle sera généralement ramenée à la langue dans son ensemble. Le corpus représentatif – on parle souvent dans ce cas de *corpus de référence* – sera celui qui permettra d'étudier la langue dans toute la variété et la richesse de ses usages, cette langue n'étant bien entendu pas envisagée comme une structure homogène et stable mais comme la somme des variations qui se rencontrent dans un idiome donné. Les contextes de variation pris en compte devraient en théorie correspondre aux contextes d'usage les plus fréquents de la langue étudiée. Le recensement de l'ensemble de la production linguistique d'une tranche temporelle étant à peu près impossible à faire, il assez rare qu'on pose clairement la question en ces termes lors de la sélection du corpus.

Nombreux sont les linguistes à considérer la réalisation d'un corpus représentatif d'une langue ou d'un état de langue comme une chimère, les possibilités de variation linguistique étant pour ainsi dire infinies et en grande partie encore inconnues des chercheurs. C'est pourtant dans cette perspective que nous avons conçu le programme *Corpus représentatif des premiers textes français*, financé par l'Agence nationale de la recherche (2007-2011). La volonté de construire un corpus aussi représentatif que possible d'une langue particulière – en l'occurrence le très ancien français (9ème-12ème siècle) – constitue bien un idéal à atteindre, et ce n'est pas parce que cette représentativité n'est jamais parfaite et qu'elle peut toujours être améliorée qu'on doit abandonner sa poursuite.

Les difficultés sont certes nombreuses, spécialement pour les périodes anciennes du français où les sources sont lacunaires, les domaines couverts par l'écrit vernaculaire s'accroissant au fur et à mesure que l'on progresse dans le temps ([doc. 5] Guillot *et al*. 2008) : on sait que les premiers textes étaient presque tous religieux, puis, qu'à partir du 12ème siècle, une littérature commence à s'écrire, presque toujours en vers et dans l'aire anglo-normande surtout. Si la recherche diachronique doit s'accommoder d'une certaine pénurie des données, les corpus de langue moderne doivent au contraire faire face à une abondance difficilement gérable et maîtrisable. Mais ces difficultés, de quelque type qu'elles soient, ne remettent pas en cause la valeur des corpus de référence, puisque ce sont eux qui permettent à la recherche de parvenir au plus haut degré de généralité possible.

Le cadre méthodologique défini par les entreprises de ce type justifie également à soi seul qu'on se donne cet objectif, ce cadre devant permettre, *in fine,* de décrire les usages spécifiques dont les données rassemblées dans le corpus sont représentatives. Que l'on vise un corpus représentant une langue dans son intégralité ou un ensemble d'usages plus limité, la méthodologie consiste à toujours définir le plus clairement et le plus finement possible les

conditions de variation délimitant le périmètre linguistique couvert par le corpus. La représentativité peut dès lors s'entendre comme une valeur relative et scalaire qui sert avant tout à connaître et à caractériser au mieux le corpus qui sert de base à l'analyse.

Cette notion relative et scalaire peut également s'appliquer aux données elles-mêmes, deux textes ou deux unités discursives pouvant différer dans leur degré de représentativité relativement à un usage prédéterminé. On peut supposer, par exemple, que certains textes sont plus représentatifs que d'autres du roman du 20ème siècle et, qu'à ce titre, ils sont de meilleurs candidats pour représenter le roman dans un corpus. On verra dans la section suivante que la représentativité des données se définit le plus souvent de manière externe grâce à des paramètres situationnels et fonctionnels qui s'appliquent aux unités discursives dans leur ensemble. Du point de vue fonctionnel et externe, il est souvent difficile de distinguer différents degrés de représentativité, même si rien n'en interdit le principe, et, dans la pratique, il me semble que cela se fait rarement[22]. Ce qui paraît finalement conférer à telle ou telle unité discursive un degré de représentativité supérieur à une autre, ce serait plutôt ses caractéristiques internes, ce sur quoi porte l'analyse linguistique proprement dite et qui se mesure grâce aux traits linguistiques qu'elle met au jour. C'est donc plutôt en bout de chaîne, une fois que l'analyse a été réalisée et par comparaison avec les autres données intégrées dans le corpus, que des unités discursives pourront être caractérisées comme étant plus représentatives ou plus prototypiques de certains usages que d'autres.

1.3. Les critères de variation externe / interne

Comme on vient de le voir, les paramètres qui permettent de décrire les unités linguistiques et les usages qu'elles représentent sont déterminés *a priori* et de façon purement externe à l'analyse proprement dite. Ces critères précisent quels facteurs de variation sont pris en compte dans le corpus, et permettent éventuellement d'en sélectionner quelques-uns pour focaliser l'étude sur ceux-ci. C'est en mettant en relation la variation externe du corpus sur

[22] Habert (2000 : 17) insiste par ailleurs sur les risques de *déformation* qu'encourt le corpus :
« *Une déformation se produit quand les caractéristiques d'un échantillon sont systématiquement différentes de celles de la population que cet échantillon a pour objectif de refléter. Un extrait de 2000 mots d'une interview de F. Mitterrand par Y. Mourousi ne permet guère d'extrapoler et d'en tirer des conclusions sur le français mitterrandien ou sur l'interaction journaliste-homme politique. Utiliser les articles de la seule rubrique Economie du* Monde*, quel que soit le volume textuel rassemblé, risque fort de déboucher sur une image déformée du français employé par ce journal.* »
B. Habert identifie un second écueil à éviter, lié à la taille du corpus et qu'il appelle l'*incertitude* du corpus :
« L'incertitude survient quand un échantillon est trop petit pour représenter avec précision la population réelle ».

lequel on travaille avec la variation interne repérée dans les données elles-mêmes que l'on progresse dans la connaissance et la description linguistique.

Pour J. Sinclair, ces critères de variation externe ont trait à la fonction communicative des unités discursives :

> *« The contents of a corpus should be selected without regard for the language they contain, but according to their communicative function in the community in which they arise. »* (Sinclair 2005 : 1)

Il s'agit d'une définition très restrictive, même si la visée communicative du discours dans le contexte de l'énonciation constitue un facteur tout à fait primordial et central pour sa caractérisation externe. Il me paraît nécessaire d'ajouter d'autres critères de nature différente à celui-ci, même si tous partagent deux caractéristiques communes :

- ils permettent toujours de décrire l'unité discursive dans son intégralité ; cette unité peut correspondre ou non au texte (il est possible de descendre à un niveau plus fin, si l'on veut étudier, par exemple, la variation entre le discours direct et le récit ou la variation entre différents types de discours direct), mais la caractérisation externe s'applique toujours à tous les éléments qui sont rassemblés au sein de l'unité discursive ;
- ces critères doivent être bien distincts des critères internes qui seront retenus lors de la phase d'analyse du corpus.

Même s'ils ne peuvent se réduire à la fonction communicative de l'unité discursive, les éléments généralement retenus comme critères externes sont avant tout de nature situationnelle et fonctionnelle. Dans les recherches qu'il a menées sur l'anglais oral et écrit ou sur les dimensions de la variation inter-linguistique entre registres, D. Biber (notamment 1988 et 1995), en cite un grand nombre : canal de la communication, destinataire, auteur, visée du discours, thèmes, etc. On peut naturellement y adjoindre la date où le discours a lieu, et pour la période médiévale la date du manuscrit qui a servi à l'établissement du texte par exemple. En réalité, les critères pris en compte peuvent varier en fonctions des époques, des sociétés dans lesquelles les discours sont produits, etc. Ils dépendent en partie aussi des questions de recherche et des variétés d'usage qu'on se propose d'étudier.

Dans ses premiers travaux, Biber utilise *genre* comme terme couvrant l'ensemble de la variation situationnelle. La définition qu'il en donne correspond à son acception habituelle :

> *« Genres are the text categories readily distinguished by mature speakers of a language ; for exemple […] novels, newspaper articles, editorials, academic articles,*

> *public speeches, radio broadcasts, and everyday conversations. These categories are defined primarily on the basis of external format. »* (Biber 1989 : 5-6)

Mais Biber insiste aussi sur le fait que les genres sont directement liés à la fonction communicative du texte. C'est ce qui explique que ce terme puisse subsumer l'ensemble des variables situationnelles. Biber précise également que les catégories génériques sont définies sur la base d'éléments non formels :

> *« Genre categories are determined on the basis of external criteria relating to the speaker's purpose and topic ; they are assigned on the basis of use rather than on the basis of form. »* (Biber 1988 : 170)

Les propriétés formelles relèvent, selon Biber, de la variation interne aux unités discursives. Le regroupement de certaines de ces propriétés formelles permet à Biber de définir des types de textes, un même type pouvant s'instancier dans des genres différents[23] :

> *« Genres characterize texts on the basis of external criteria, while text types represent groupings of texts that are similar in their linguistic form, irrespective of genre. For example, an academic article on Asian history represents formal, academic exposition in terms of the author's purpose, but its linguistic form might be narrative-like and more similar to some type of fiction than to scientific or engineering academic articles. The genre of such a text would be academic exposition, but its text type might be academic narrative. »* (Biber 1988 : 170)

Dans son ouvrage de 1995, Biber abandonne cette terminologie et préfère le terme de *registre* à celui de *genre*, mais son approche reste fondamentalement la même.
Il ne paraît pas nécessaire d'entrer ici dans les débats sur la sémantique de ces diverses notions (voir notamment Biber 1995 : 6-11 et Lee 2001), ni sur l'utilité qu'il peut y avoir à choisir un terme pour subsumer tous les facteurs de variation externe (je reviendrai dans la section 2.2. sur la définition du genre discursif). Ce sur quoi il me semble plus intéressant d'insister, c'est sur l'impossibilité de superposer parfaitement critères formels et critères

[23] Dans son article de 1989 déjà cité, Biber établit une distinction entre les facteurs extralinguistiques, qui permettent de définir les genres, et les facteurs linguistiques qui sont à la base de la typologie des textes qu'il propose.

internes. Le fait qu'un texte soit écrit en vers ou en prose, par exemple, relève bien d'une opposition formelle. Cette propriété de l'unité discursive peut cependant être définie *a priori*, préalablement et indépendamment de l'analyse linguistique, et pour l'ensemble de cette unité. Nous avons ainsi retenu ce critère comme facteur de variation externe dans le système des descripteurs de la Base de français médiéval et nous l'avons nommé critère de la « forme » du texte.

L'approche développée par Koch et Österreicher (notamment 1990 et 2001) autour de l'opposition entre oral et écrit, ou plus exactement entre modes conceptuels de la proximité et de la distance communicative, les conduit à proposer un ensemble de paramètres permettant de situer les textes sur un continuum allant du pôle de la proximité au pôle de la distance communicative. J'ai déjà fait allusion à ces paramètres dans le chapitre 1 (section 2.2.1.) et je voudrais insister ici sur le fait que certains d'entre eux sont comparables à ceux énumérés par Biber : communication privée/publique, interlocuteur intime/inconnu, coprésence/séparation spatio-temporelle, communication spontanée/préparée, dialogue/monologue, liberté/fixation thématique.

Ces différents éléments ont pour fonction de caractériser le contexte et la situation énonciative dans laquelle le discours est produit. D'autres critères concernent au contraire les traits formels du discours (émotionnalité forte/faible, ancrage référentiel dans la situation, par exemple). Or la fonction de cet ensemble de facteurs semble être assez proche de celle qu'assigne Biber aux critères externes, puisque ces paramètres permettent de situer les textes sur une échelle qui les caractérise en bloc.

Si une dichotomie absolue entre critères formels et non formels semble difficilement tenable[24], il n'en reste pas moins que le mélange des deux et l'utilisation des premiers comme critères externes peuvent créer des difficultés pratiques (comment reconnaître *a priori* un discours dont l'émotionnalité est forte, par exemple ?). C'est ce qui nous a en partie freinés dans l'intégration des paramètres définis par Koch et Österreicher dans le système des descripteurs (méta-données) de la Base de français médiéval.

Au terme de ce rapide survol, je retiendrai surtout la nécessité de bien différencier les critères de variation externe et interne comme principe méthodologique, la nature et la réalité de leur contenu pouvant toujours se discuter. Je reviendrai plus en détail dans la section 2.2 sur les

[24] Par ailleurs, rien n'interdit en principe qu'une variable interne devienne à un autre moment une variable externe du corpus : « Pour certaines études, il est possible aussi qu'un phénomène qui constitue une variable interne devienne à son tour un contenant, qui porte ainsi une variable externe » ([doc. 3] Guillot *et al.* 2007 : 130).

critères externes qui nous ont semblé les plus pertinents pour le français médiéval et que nous avons adoptés pour notre Base.

1.4. L'équilibrage du corpus

Le corpus, qu'il soit conçu comme un corpus de référence ou comme un corpus plus spécialisé, résulte toujours d'une forme d'échantillonnage par prélèvement d'un volume de données plus ou moins important sur l'ensemble de la production langagière effective. Lorsque les données rassemblées dans le corpus sont de nature hétérogène, l'équilibrage des différents types de ressources agrégées (définis grâce aux critères externes dont il vient d'être question) permet de ne pas biaiser l'analyse en accordant une importance prépondérante à certains d'entre eux au détriment des autres. Il s'agit là d'un principe très général et bien admis, qui peut, dans certains cas, être pondéré grâce à des outils statistiques permettant de minorer les effets négatifs du déséquilibre.

Dans la pratique, il s'avère pourtant très difficile de parvenir à un équilibre parfait. Cela est particulièrement vrai pour les corpus diachroniques, dont on sait qu'ils sont tributaires des ressources produites dans le passé et parvenues jusqu'à nous. Le corpus des premiers textes français, par exemple, contient un grand nombre d'œuvres en vers, la grande majorité des textes antérieurs au 13ème siècle étant produits sous cette forme. La plupart des textes en prose intégrés au corpus pour corriger ce biais sont des traductions d'œuvres religieuses latines, ce qui tend à accroître le poids des traductions et des textes religieux dans le corpus. Le gain obtenu d'un côté génère parfois un déséquilibre ailleurs dans le corpus…

Les critères de variation étant multiples, et d'une certaine façon sans limite (d'où la nécessité d'en faire une liste qui soit gérable et d'un niveau de granularité acceptable), l'équilibre total semble impossible et les recherches empiriques doivent composer avec l'existant. C'est la raison pour laquelle il me paraît important d'aborder cet aspect du corpus dans la présentation qu'on en donne, et c'est ce que je me suis toujours efforcé de faire dans mes publications. Même s'il peut sembler fastidieux de donner des informations aussi précises et concrètes, ces données sont pourtant nécessaires à la compréhension et à l'évaluation du travail de recherche et l'on peut regretter que la pratique n'en soit pas davantage généralisée et normée. Ce devrait être le cas, en particulier, dès lors qu'on adopte une approche comparative et contrastive du corpus de recherche.

2. L'approche contrastive

La plupart des études diachroniques, parce qu'elles comparent des états de langue successifs, impliquent une approche contrastive des données. Plusieurs des recherches que j'ai menées en diachronie mais aussi en synchronie adoptent cette perspective et visent à apparier les phénomènes observés à leur contexte d'usage en caractérisant et en comparant ces différents contextes (voir notamment ma thèse de doctorat et [doc. 3] Guillot *et al.* 2007, [doc. 6] Guillot 2009, [doc. 8] Guillot 2010a, [doc. 10] Rainsford *et al.* 2012, [doc. 13] Guillot *et al.* à par a et [doc. 14] Guillot *et al.* à par. b). Cette approche me paraît tout à fait centrale dans l'étude du système linguistique, que l'analyse qu'on en fait soit de nature qualitative ou quantitative. Nous avons tenté d'en décrire les grands principes dans notre article collectif de 2007 et l'avons illustrée à l'aide d'une étude sur le déterminant anaphorique *ledit*.

2.1. Présentation de l'approche

L'approche contrastive mobilise les variables externes dont il a été question plus haut. Elle utilise ces variables pour construire les contextes de comparaison des phénomènes linguistiques étudiés. L'intérêt de cette méthode est qu'elle se sert de contrastes établis grâce à des critères prédéfinis comme d'un cadre d'analyse, pour parvenir, de proche en proche, à une connaissance aussi précise que possible des objets et des phénomènes linguistiques étudiés.

Cette méthodologie implique qu'on sélectionne en général un petit nombre de contrastes pertinents pour comparer les usages, en neutralisant au mieux les variations autres que celles sur lesquelles on travaille. La plupart des études diachroniques, par exemple, visent à neutraliser tous les facteurs de variation autres que le temps. En pratique, on s'aperçoit que cet objectif est à peu près impossible à réaliser complètement (de même qu'il est très difficile de pondérer parfaitement le corpus, et pour les mêmes raisons). Cela suppose en effet qu'on dispose de suffisamment de données pour ne jouer que sur un critère à la fois en contrastant des sous-corpus parfaitement comparables à un critère près. Dans le cas du français médiéval, on peut supposer que les variables externes les plus prégnantes recouvrent *a minima* les paramètres suivants : date, forme (prose/vers), dialecte, domaine, genre et auteur. Notre étude sur l'oral représenté ([doc. 14] Guillot *et al* à par. b) montre que l'opposition récit *vs* oral représenté est également une dimension de variation essentielle, dont il faudrait tenir compte dans ce type d'approche. Compte-tenu de la disponibilité des ressources actuelles pour le Moyen Âge, il est impossible de neutraliser la quasi-totalité de ces variations. La seule

alternative est de trouver le meilleur compromis en précisant de la façon la plus honnête possible les limites des résultats obtenus.

2.2. Le cadre typologique

Nos recherches sur les variables externes nous ont amenés à construire et à utiliser des typologies discursives élaborées collectivement. Grâce à l'aide de nos collègues médiévistes impliqués dans le projet *Corpus représentatif des premiers textes français*, nous avons pu définir un système de méta-données adapté à l'ensemble de la période médiévale. Les grands principes de ce système sont développés dans la *Présentation des descripteurs du projet CORPTEF* ([doc. 7] http://corptef.ens-lyon.fr/spip.php?rubrique60) et le *Manuel de description des textes de la BFM* (http://bfm.ens-lyon.fr/article.php3?id_article=301)[25].

Ce système a été présenté à de multiples occasions et il été utilisé par l'équipe du *Dictionnaire étymologique de l'ancien français* (Université de Heidelberg) pour les lettres A et B de sa bibliographie en ligne (http://www.deaf-page.de/bibl_neu.htm), qui sert de référence aux médiévistes travaillant sur le français. Ce système a également inspiré d'autres projets de recherche (notamment le projet *Modéliser le changement : les voies du français* dirigé par France Martineau à l'Université d'Ottawa et le projet de *Grande grammaire historique du français* dirigé par Christiane Marchello-Nizia, Bernard Combettes, Sophie Prévost et Tobias Scheer). Outre son utilité pour les études qui sont menées dans notre équipe, ce système est pour nous une première avancée dans la définition et la diffusion de normes de description des textes médiévaux qui soient stables et partagées par une large communauté de chercheurs (voir le chapitre 3)[26].

Il ne me paraît pas utile de présenter ici en détail toutes les variables de notre système et leurs diverses valeurs possibles. Je me contenterai d'en énumérer les principales : identifiant de l'édition du texte, auteur de l'œuvre, date de composition de l'œuvre et du manuscrit, dialecte du texte (de l'auteur et du scribe), forme de l'œuvre (*vers/prose*), domaine discursif,

[25] Le document intitulé *Présentation des descripteurs du projet CORPTEF* expose le système dans ses grandes lignes. Le *Manuel de description des textes de la BFM* explique comment chaque champ doit être renseigné dans la base Access qui contient tous nos descripteurs et définit les normes de codage des métadonnées de la Base de français médiéval.

[26] Cette volonté de normalisation et d'échange explique que notre système de descripteurs ait été conçu dans l'optique de l'approche présentée ici, mais aussi dans un but purement documentaire. Les descripteurs de notre Base servent également à identifier, à (re)trouver ou à échanger des textes, ce qui explique qu'ils comportent des informations bibliographiques (titre de l'œuvre, éditeur, date, lieu et maison d'édition) ou juridiques (conditions d'usage des textes), ainsi qu'un identifiant unique pour chaque édition du texte. Ces métadonnées sont d'un autre type que les variables externes mobilisées dans les contrastes, mais elles sont intégrées à notre système de descripteurs au même titre que les autres.

(*didactique, historique, littéraire, juridique, religieux, sources documentaires*), genre discursif (*bestiaire, chanson de geste, mémoire, roman, sermon, traité*, etc.)[27], thème (limité aux œuvres didactiques : *nature, géographie, trivium, quadrivium, médecine, droit, théologie,* exempla*, encyclopédie, politique*), relation (*traduction, remaniement, commentaire, œuvre originale*).

Les catégories du domaine et du genre discursifs méritent qu'on s'y arrête davantage. Dans notre système, le domaine discursif correspond à la destination principale du texte et au domaine d'activité auquel il se rattache. Notre conception du domaine est très proche de ce que Sinclair appelle la *fonction communicative* du texte : lorsque la fonction du texte est de divertir le lecteur, le domaine est littéraire, lorsqu'elle est d'enseigner et d'instruire, le domaine est didactique, lorsqu'elle est d'édifier, le domaine est religieux, lorsqu'elle est de consigner les événements du passé, le domaine est historique, et lorsqu'elle est de réguler la vie sociale par des règles, le domaine est juridique. Le domaine des sources documentaires (ou actes de la pratique) rassemble des textes qui peuvent, par certains côtés, se rapprocher des précédents mais dont la finalité est plus pratique que normative (chartes, actes divers, comptes, registres, censiers, etc.)[28].

Le genre discursif correspond quant à lui à une nomenclature très répandue (d'où les nombreux répertoires et manuels de référence qui utilisent les catégories génériques, comme par exemple le *Manuel bibliographique de la littérature française du Moyen âge* de Bossuat, le *Grundriß der romanischen Literaturen des Mittelalters* dirigé par Jauss et une partie des volumes de la *Typologie des sources du moyen âge occidental* dirigés par Génicot), mais qui est en même temps plus difficile à définir. Je reviendrai plus loin sur la définition du genre et me contenterai de noter ici que notre système éclate dans deux catégories bien distinctes, le domaine et le genre discursifs, ce que Biber rassemblait sous le terme unique de *genre*, d'une façon qui me semble trop imprécise. Les différents manuels cités plus haut suivent une démarche comparable à la sienne et n'utilisent pas une terminologie très claire dans la définition de ces catégories.

Le fait que nous proposions de séparer le domaine et le genre discursifs ne doit pas masquer les liens étroits qui unissent ces deux dimensions de la description des textes. Il est évident que certains genres, pour ne pas dire la plupart, relèvent d'un domaine discursif unique : les

[27] Il est difficile d'en donner une liste finie. Les genres sont relativement nombreux, même pour la période médiévale, et leur liste tend à s'allonger au fur et à mesure que de nouveaux textes sont intégrés au corpus.

[28] D'où le terme répandu chez les historiens d'*actes de la pratique*. Dans son panorama des sources de l'histoire médiévale, O. Guyotjeannin insiste sur leur caractère appliqué :

> « *Un cran au-dessous de la littérature juridique, les actes de la pratique ont été longtemps vus comme l'application de la norme […].* » (Guyotjeannin 1998 : 176)

chansons de geste et le roman appartiennent au domaine littéraire, le traité au domaine didactique, les chroniques au domaine historique, etc. Les quelques cas d'appartenance multiple sont relativement isolés (on peut citer le théâtre qui peut être religieux ou profane au Moyen Âge et, dans le dernier cas, littéraire). Mais ces associations privilégiées ne remettent pas en cause, selon nous, l'utilité de distinguer deux niveaux de description différents. Certains chercheurs adoptent une autre démarche et proposent des systèmes hiérarchisés qui pourraient permettre de formaliser ces relations : le genre pourrait, dans ce cas, être considéré comme une sous-catégorie du domaine discursif. Mais, outre qu'on vient de voir qu'il existe quelques genres communs à deux domaines, cette solution présente l'inconvénient de mêler sans les distinguer des types d'informations qui nous paraissent être différents. Notre système de descripteurs se distingue ainsi de beaucoup d'autres en ce qu'il comporte un très grand nombre de catégories clairement distinctes mais peu hiérarchisées[29].

L'examen critique de mes propres travaux de recherche (individuels ou collectifs) et des études menées par les collègues qui utilisent notre Base montre qu'en réalité toutes les variables externes de notre système ne sont pas utilisées de façon constante et uniforme. On observe, en particulier, que les genres et les domaines discursifs sont beaucoup plus souvent mobilisés que les dialectes dans ces études. Cette observation semble conforter une hypothèse de Biber (1995). Dans son ouvrage sur les variations de registres, Biber établit une distinction très nette entre ce qu'il appelle les registres et les dialectes. Dans sa terminologie, les registres (qui remplacent en quelque sorte les genres de son ouvrage de 1988) recouvrent les facteurs de variation liés au contexte situationnel (but de la communication, circonstances de l'énonciation, relations entre les partenaires de la communication, caractère planifié ou non du discours, etc.) et les dialectes se limitent à la dimension sociale et dialectale de la variation linguistique (variations diastratique et diatopique). L'une des thèses fortes de son ouvrage est que la variation en registres prime de manière forte sur la variation en dialectes. Ses effets sur la production langagière seraient bien supérieurs à ceux des facteurs sociaux et régionaux, plus conventionnels et dont l'action serait plus limitée :

> *« When speakers switch between registers, they are doing different things with language – using language for different purposes and producing language under different*

[29] La typologie de Biber, par exemple, subdivise les genres (23 au total dans son ouvrage de 1988) en différents sous-genres. Si l'on prend l'exemple de la presse, on observe que ce genre regroupe les sous-genres de la politique, des sports, de la société, etc. Dans le cas de l'oral préparé, Biber distingue les sermons, les exposés académiques, les discours politiques, etc. Ces sous-genres mêlent en réalité des informations de nature très diverse. Le genre de Biber correspond pour nous au domaine, et les sous-genres correspondent tantôt à ce que nous appelons des genres, tantôt à ce que nous définissons comme des thèmes.

> *circumstances. Many language choices are functionally motivated, related to these differing purposes and production circumstances, and thus there are often extensive linguistic differences among registers. In contrast, dialect differences are largely conventional and therefore less fundamental in nature. »* (Biber 1995 : 2)

Même si l'opinion exprimée dans cet extrait me paraît bien dogmatique et correspond sûrement mieux à la situation de l'anglais ou de l'américain actuels qu'à celle du français médiéval, il reste que la question de l'importance relative des différents paramètres de variation mérite d'être posée. L'un des apports de l'approche développée par Biber dans ce livre est qu'il étudie les variations de registres dans une perspective translinguistique et transculturelle, en s'appuyant sur l'idée que leurs caractéristiques sont communes à toutes les cultures. Et l'on peut certainement mettre en relation l'usage consistant à contraster les genres et les domaines discursifs avec le fait qu'il s'agit de catégories très générales et très largement partagées.

Il est probable en même temps que si la plupart des études que j'ai consultées s'appuient sur la variation en genres et en domaines davantage que sur la variation dialectale, par exemple, c'est parce que ces études portent avant tout sur des phénomènes syntaxiques et sémantiques, et non sur les aspects phonétiques et/ou graphiques de la langue. Comme on l'a dit plus haut, la pertinence des variables dépend en partie du type et du niveau d'analyse choisis. Il semble que le paramètre dialectal influe assez peu sur la variation syntaxique, sans qu'on puisse dire avec exactitude si ce postulat reflète vraiment la réalité ou s'il découle partiellement du fait que les études de syntaxe médiévale tiennent rarement compte de ce type de variation[30]…

Ces quelques réserves étant posées, on peut tout de même noter que les typologies en genres et en domaines discursifs ont la caractéristique remarquable qu'elles tendent à transcender non seulement les époques (c'est le cas aussi de la variation dialectale) mais aussi et surtout les langues elles-mêmes. Koch et Österreicher (2001 : 601-603) insistent sur cet aspect de ce qu'ils nomment les *traditions discursives*, pour certaines héritées du latin, et qui sont en grande partie communes à toutes les langues romanes[31]. Ces traditions discursives sont le fruit d'une longue tradition historique et sont parfaitement conventionnalisées, contrairement à ce

[30] A l'exception notable de R. Ingham, qui montre au travers de nombreuses études sur la syntaxe de l'anglo-normand que ce dialecte suit globalement les évolutions syntaxiques du français continental et reste assez imperméable à l'influence de l'anglais. Si ces études intègrent bien la variation dialectale, leurs conclusions tendent finalement à minorer le facteur dialectal dans la variation et l'évolution syntaxique.

[31] On peut noter que l'*Inventaire systématique des premiers documents des langues romanes* (Frank 1997) est organisé en fonction d'une typologie en domaines et genres, ce qui montre bien que ces catégories sont communes aux langues romanes et permettent de les comparer.

que la citation de Biber donne à entendre. Si ces traditions opèrent à un niveau très général, c'est parce qu'elles reposent sur des conventions culturelles qui dépassent le cadre d'une communauté linguistique donnée. Ainsi s'explique qu'elles soient communes au vaste territoire couvert d'abord par la Romania puis par les langues romanes et même par d'autres langues[32].

Les traditions discursives sont ainsi intimement liées pour le français écrit au développement de la langue écrite et à la fixation de ses normes. Elles forment les moules qui seront peu à peu investis par la langue vernaculaire et jouent, à ce titre, un rôle tout à fait prépondérant sur ses règles d'usage. Le bilinguisme des clercs et la place longtemps dominante du latin à l'écrit favorisent d'autant plus la transmission de ces canons translinguistiques sur la longue durée.

La stabilité des traditions discursives explique en partie aussi que tous les corpus diachroniques de français que je connais et auxquels j'ai participé (le corpus de la *Grande grammaire historique du français* ou le corpus en cours de construction du projet *Presto* par exemple) s'appuient principalement sur une typologie en genres et en domaines. Bien que l'on sache que ces catégories ne sont pas entièrement stables – on peut se demander ce qu'il y a de commun entre un roman du 12ème siècle, souvent en vers, et un roman du 20ème siècle –, et même si certains genres disparaissent (la chanson de geste par exemple) et que d'autres surgissent au fil du temps (les genres de la presse notamment), cette typologie en genres et en domaines permet bien d'établir une sorte de cadre d'analyse panchronique.

Les paramètres du genre et du domaine discursifs se distinguent donc par leur caractère relativement universel et, en même temps, par la difficulté maintes fois éprouvée d'en donner une liste close et parfaitement consensuelle. Il est possible de discuter à l'infini d'une liste de genres ou du choix d'affecter tel ou tel genre à une œuvre donnée. Mais, comme on l'a dit plus haut, le but premier de la démarche contrastive n'est pas de parvenir à une catégorisation parfaite des contextes d'usage, mais plutôt d'utiliser des catégories prédéfinies pour établir des comparaisons internes. En ce sens, les genres et les domaines discursifs constituent des typologies très utiles, parce qu'elles sont le produit d'une longue et large tradition et d'un savoir partagé sur les données textuelles[33].

[32] Koch (1997) donne comme exemple les canons poétiques des troubadours occitans, français et italiens qu'on trouve aussi chez les Minnesänger en moyen haut allemand.

[33] Nous avons tenté de suivre cette tradition dans la définition des genres et des domaines de la Base de français médiéval en nous inspirant très largement du *Manuel bibliographique de la littérature française du Moyen âge* de Bossuat, repris et prolongé par Monfrin et Vielliard. Cette bibliographie, qui sert de référence aux médiévistes, philologues, historiens et linguistes travaillant sur des textes français, est organisée par domaines et par genres (même si sa terminologie est un peu différente et si ces deux paramètres sont souvent mêlés, comme je l'ai souligné plus haut).

S'il semble bien que les genres et les domaines discursifs constituent de tout temps et en tous lieux des paramètres fondamentaux pour l'étude et la prise en compte de la variation linguistique, il est probable en revanche que d'autres paramètres varient en importance au fil du temps. Tel est cas en français de la variation dialectale, bien sûr, mais aussi de la variation en forme (*vers/prose*). L'opposition entre le vers et la prose est très centrale durant tout le Moyen Âge, mais elle devient de plus en plus secondaire, au fur et à mesure que le développement extraordinaire de la prose réduit la part du vers à presque rien. Il est évident que des facteurs multiples, liés à l'évolution des sociétés, à la place occupée par l'écrit, etc., relativisent l'importance des différentes dimensions de la variation linguistique. Dans son ouvrage sur la variation sociale en français, F. Gadet attache une importance particulière à « la saillance pour la communauté ». Elle propose une hypothèse sur l'évolution du poids relatif de ces variations :

> *« Sans doute l'ordre d'apparition et de perte des types de variation reflète-t-il la saillance pour la communauté. Nous ferons ainsi l'hypothèse que le français serait passé d'une domination diatopique (19ème siècle) à un primat du diastratique, jusqu'à un primat actuel du diaphasique. »* (Gadet 2003 : 16)

S'il paraît évident que la variation diatopique est particulièrement forte au Moyen Âge – ce qui explique que les premières études sur le français médiéval aient porté avant tout sur les systèmes phonétique et graphique et que ce soit toujours ce sur quoi les introductions des éditions de textes médiévaux insistent au premier chef[34] –, la recherche sur les corrélations entre types de variation et sur la prégnance de certains d'entre eux à différents moments de l'histoire du français reste en grande en grande partie à mener.

Comme je l'ai souligné déjà, la plupart des recherches actuelles privilégient dans la pratique un ou deux paramètres de variation. Et dans le champ de la diachronie, bon nombre de travaux se contentent d'une partition du corpus en domaines et/ou un nombre limité de genres discursifs. C'est ce que nous avons fait dans notre étude sur *ledit* ([doc. 3] Guillot *et al*. 2007) et *or(e)* par exemple ([doc. 6] Guillot 2009). Or j'ai indiqué plus haut que le domaine correspondait à la fonction communicative du texte mise en avant par Sinclair comme principal facteur de variation externe.

[34] La plupart des œuvres médiévales sont anonymes et nos connaissances sur les auteurs qui sont expressément nommés dans les textes sont de toute façon si fragmentaires qu'un des buts des introductions est de dater et de localiser les textes. Les éléments sur lesquels les éditeurs s'appuient pour ce faire sont principalement les traits dialectaux (traits phonétiques/graphiques et parfois aussi lexicaux).

Mais le choix de s'en tenir au domaine repose certainement aussi sur le fait que ce critère correspond à un niveau de granularité assez élevé et permet de limiter le nombre de parties contrastées : nous avons identifié six domaines seulement pour la période médiévale, mais notre nomenclature en genres comporte actuellement plus de 40 valeurs différentes. Son application est donc bien plus complexe et plus lourde que celle des domaines, à moins de se limiter à la comparaison de quelques genres uniquement (ce à quoi on se limite souvent, comme je l'ai signalé plus haut). Comme on l'a souligné à plusieurs reprises, la rareté des ressources numériques disponibles pour le français médiéval interdit à peu près complètement de prétendre couvrir tous les genres possibles. Une manière de résoudre ce problème très pratique est de définir des genres suffisamment généraux pour en limiter drastiquement le nombre. Mais comme je l'ai dit, la catégorie du genre repose sur une longue tradition historique. Il faut donc composer avec des critères divers pour parvenir à un équilibre qui tiennent compte de ces traditions sans conduire à une inflation des catégories qui les rende difficilement gérables. C'est ce que nous nous sommes efforcés de faire pour notre Base de français médiéval, sans prétendre être parvenus à un système idéal.

2.3. Les unités de description

J'ai insisté plus haut sur le fait que les critères de description externe doivent porter sur l'intégralité de l'unité discursive. Tous les exemples que j'ai donnés jusqu'ici ont pris comme unité le texte, ce texte pouvant se définir comme une instanciation particulière de l'œuvre médiévale (comme on le sait, la culture manuscrite du Moyen Âge fait qu'à une même œuvre peuvent correspondre plusieurs versions différentes, parfois en très grand nombre).

Ce choix apparemment très simple pose en réalité quelques difficultés pratiques, notamment dans le cas des recueils composites mêlant ce qu'on peut considérer comme des sous-parties d'un même texte ou comme des textes à part entière. C'est le cas des *Lais* de Marie de France par exemple, ensemble de courts récits en vers écrits à peu à la même époque par le même auteur mais bien délimités les uns des autres. C'est le cas également du *Roman de la Rose*, dont la tradition littéraire considère qu'il constitue une seule et même œuvre, commencée dans les années 1230 par Guillaume de Lorris et continuée une quarantaine d'années plus tard par Jean de Meun[35].

[35] On s'appuie pour cela principalement sur le fait que Jean de Meun indique de manière très explicite au début de son texte qu'il poursuit l'œuvre de son prédécesseur Guillaume de Lorris.

J'ai cependant signalé deux recherches menées sur l'oral représenté au Moyen Âge qui opèrent des contrastes sur des unités discursives de niveau inférieur correspondant à des sous-parties d'un même texte. En théorie, rien ne s'oppose à ce qu'on choisisse comme unités discursives des divisions internes aux textes, les résultats des deux études que je viens de mentionner montrant de façon évidente que les variations intra-textuelles peuvent être bien supérieures aux variations intertextuelles. L'analyse contrastive de l'oral représenté et du récit (qui recouvre de manière large tout ce qui n'est pas de l'oral représenté) dans un corpus relativement étendu nous a ainsi révélé que l'opposition entre ces deux types discours était très forte, qu'elle semblait primer sur les autres facteurs de variation (notamment le domaine discursif) et rester relativement stable au cours du Moyen Âge[36].

Dans mes recherches sur la sémantique des démonstratifs au Moyen Âge il m'est également arrivé à plusieurs reprises de mentionner le caractère polémique ou argumentatif d'une séquence textuelle en essayant de montrer en quoi l'usage du démonstratif servait le but communicatif de l'auteur dans cette séquence (en particulier dans le *Roman de la Rose* et les *Miracles* de Gautier de Coinci, voir notamment [doc. 11] Guillot 2012a). Mais je n'ai jamais utilisé de façon systématique une typologie des types de discours intra-textuels, comme en propose par exemple J-M. Adam à travers la notion de *séquence discursive* (narrative, descriptive, argumentative, explicative et dialogale, notamment Adam 1999 : 82). Si la pertinence et l'utilité de telles catégories paraissent évidentes, celles-ci posent d'importantes difficultés dans la pratique : de quelle manière peut-on repérer, délimiter et typer efficacement au moyen d'une valeur (séquence descriptive par exemple) un segment discursif au sein d'un texte ?

Si l'on s'en tient à la seule question de la délimitation de ces séquences, on observe qu'elle implique une forme d'analyse et d'interprétation des éléments internes au texte. Tel n'a pas été le cas de la délimitation du discours direct dans les deux études que j'ai citées plus haut : le codage du discours direct s'est appuyé sur les marques de ponctuation (les guillemets introduits dans les textes par les éditeurs modernes) et a été réalisé de manière semi-automatique grâce à ces bornes graphiques. Cela ne signifie évidemment pas que le bornage du discours direct est exempt de toute forme d'interprétation et qu'il ne repose pas sur des choix parfois contestables. Mais la méthode contrastive se fonde de préférence sur des unités prédécoupées et pré-catégorisées pour en faire ensuite l'analyse interne. Si nous disposons un

[36] Ces recherches se sont basées sur les catégories morphosyntaxiques des mots. Elles auraient sans doute donné des résultats bien différents si l'on avait considéré uniquement leur forme graphique, beaucoup plus sensible à la variation dialectale que ne l'est leur catégorie morphosyntaxique (voir plus haut). Les tendances mises au jour dans notre étude auraient probablement été beaucoup moins nettes dans ce cas.

jour d'indices suffisamment clairs permettant une pré-construction des différents types de séquences discursives dans les textes, nous serons certainement plus à mêmes d'appliquer notre méthode à ces nouveaux objets. Et il ne fait pas de doute que l'étude de ces séquences, de leurs contrastes internes et de leur évolution donnerait de précieux renseignements sur la langue et ses usages.

La méthode dont je viens d'exposer les grandes lignes permet d'analyser les phénomènes linguistiques en établissant des contrastes sur la base de critères externes. Les typologies textuelles jouent un rôle déterminant dans cette approche puisqu'elles permettent la création de tels contrastes.

L'une des caractéristiques de cette méthode est qu'elle permet une sorte de va-et-vient entre les catégories définies *a priori* et les nouvelles catégories qui peuvent être définies *a posteriori* après analyse. En ce sens, la méthode comparative offre aussi l'avantage de permettre de construire de nouvelles typologies fondées cette fois sur les variations internes aux unités comparées. Les nombreux travaux de D. Biber sont parmi ceux qui illustrent le mieux cette démarche et c'est dans son sillage que nous avons voulu développer nos propres recherches diachroniques :

> « *Dans cette méthode, nous proposons donc de considérer les variables génériques non seulement comme des prétextes à la comparaison* a priori *des phénomènes linguistiques, mais aussi comme une représentation de la connaissance générique que nous avons acquise* a posteriori *sur les textes, après analyse comparée des différentes réalisations linguistiques observées dans ces textes*. » ([doc. 3] Guillot *et al*. 2007 : 131)

Je voudrais montrer aussi qu'une telle démarche peut aider à définir et à délimiter des zones de variation intra-discursive. Le cas de l'oral représenté me servira d'exemple. Le balisage du discours direct dans plusieurs textes médiévaux a permis de réaliser plusieurs études sur ce qui caractérise ce type de discours par opposition au reste. Parmi ces différentes études, les recherches de C. Marchello-Nizia (2012) apportent des renseignements très précieux sur les éléments linguistiques qui se trouvent à la frontière entre oral représenté et récit. Elles montrent que les éléments qu'on analyse généralement comme des introducteurs du discours direct (verbes de paroles, incises) partagent en réalité beaucoup de propriétés avec le discours direct lui-même (emploi massif du présent de l'indicatif, ordre des mots particulier avec le

verbe en première position, etc.), ce qui conduit C. Marchello-Nizia à considérer que ces éléments font eux-mêmes partie de la séquence d'oral représenté. L'ensemble de ces remarques permet de revenir sur le mode de délimitation des unités discursives étudiées. Elles pourraient, à terme, nous permettre d'utiliser les éléments mis au jour dans les recherches sur l'oral représenté pour un meilleur balisage de ces séquences.

3. L'annotation linguistique des corpus

J'aborderai cette question dans les limites de mon expérience personnelle. Celle-ci a concerné, d'une part, l'élaboration d'un modèle d'étiquetage morphosyntaxique du français médiéval (jeu Cattex2009[37]) et d'un modèle d'annotation syntaxique de l'ancien français (projet ANR-DFG SRCMF)[38], d'autre part, une annotation partielle et totalement adaptée à ma recherche personnelle sur l'évolution sémantique du démonstratif en français. Ces différents types et niveaux d'annotation ont été exploités dans plusieurs de mes recherches individuelles et collectives.

Les deux types d'annotation que je viens de définir se distinguent à plus d'un titre. Dans un cas, il a fallu concevoir et appliquer un système de catégories s'appliquant à tous les mots des textes, dans le but d'ouvrir de nouvelles possibilités de recherche avec le corpus annoté. Le système de catégories a, de ce fait, été défini en fonction des usages les plus larges possibles[39]. Dans le second cas, l'annotation linguistique a directement contribué au processus de recherche. Elle a permis de formaliser un ensemble de traits et d'usages linguistiques en fonction des hypothèses de départ et de quantifier ces usages dans le corpus. Une partie des catégories annotées ont repris, en les adaptant, les étiquettes préconstruites du jeu Cattex2009. Les autres informations encodées ont été définies au fur et à mesure de l'avancée de la recherche.

---

[37] Ce jeu a été créé en collaboration avec Sophie Prévost et Alexei Lavrentiev. Il est accessible en ligne : http://bfm.ens-lyon.fr/article.php3?id_article=176. Il est complété par une présentation des principes d'étiquetage Cattex (http://bfm.ens-lyon.fr/article.php3?id_article=173&var_mode=calcul) et par le *Manuel de référence Cattex2009* (http://bfm.ens-lyon.fr/article.php3?id_article=323), qui indique de façon relativement précise comment les étiquettes ont été utilisées dans les textes.

[38].Le site actuel du projet contient une partie de la documentation : http://srcmf.org/.

[39] Le jeu d'étiquettes Cattex2009 a été conçu pour l'ensemble de la période médiévale, mais le modèle d'annotation syntaxique de SRCMF a été élaboré et appliqué à l'ancien français uniquement. Il serait très certainement possible d'utiliser ce modèle sur le moyen français, même s'il est probable que l'évolution linguistique conduirait à le modifier partiellement et qu'elle en compliquerait l'application concrète. Ce modèle a été conçu pour répondre à des usages variés, et nous nous sommes toujours efforcés d'éviter au maximum les partis pris théoriques qui en restreindraient l'intérêt et l'application à des approches particulières.

Il s'agit dans les deux cas d'une annotation qu'on peut appeler experte (réalisée par des chercheurs, des spécialistes), et je m'intéresserai dans ce qui suit aux questions posées par la création et l'application d'un systèmes de catégories davantage qu'au processus d'annotation proprement dit ou à son évaluation *a posteriori*.

3.1. Les buts et les fonctions de l'annotation linguistique

Le développement actuel de grands corpus enrichis ouvre la voie à des recherches inédites par l'ampleur des données analysées et la finesse et la complexité des catégories mobilisées et souvent combinées au cours de l'analyse. Pour autant, ces corpus sont parfois critiqués par certains collègues, qui contestent en général le choix et la pertinence des informations annotées.

L'un des griefs souvent reprochés à l'annotation linguistique est qu'elle surimpose aux données langagières des catégories qui reposent sur l'analyse et l'interprétation du linguiste :

> *« But what is corpus annotation ? It can be defined as the practice of adding* ***interpretative***, ***linguistic*** *information to an electronic corpus of spoken and/or written language data. »* (Leech 1997 : 2)

Ainsi s'explique qu'une annotation puisse toujours être contestée et discutée à l'infini et qu'une certaine défiance en limite parfois le développement et l'usage chez les linguistes. Cette défiance se nourrit aussi de ce qu'on sépare souvent, de manière artificielle, annotation et représentation du texte. Cette dernière ne ferait pas intervenir le jugement subjectif du chercheur et ne viendrait pas altérer l'authenticité des données. Seules les données brutes et, par conséquent, totalement objectives, devraient être soumises à l'examen du linguiste. La citation suivante montre que l'idée d'une telle dissociation est assez répandue, même si Leech prend bien soin de préciser qu'il n'est pas toujours facile d'opérer une distinction aussi nette et s'il en restreint l'application aux textes écrits :

> *Second, we assume a distinction between the 'annotation' and the 'representation' of a text – a distinction which may be easy or less easy to apply. For a written text, generally these two kinds of information are relatively easy to separate* (Leech 1997 : 3).

Nous verrons dans le chapitre 3 que la création de ressources numériques, même écrites, implique une forme d'interprétation de ces ressources. La fiction de données qui seraient totalement brutes et objectives n'est réellement tenable ni en théorie ni en pratique. Cet état de fait est particulièrement évident aux yeux des médiévistes, pour qui les ressources textuelles sont par définition médiatisées – elles sont le produit de l'activité scientifique d'un philologue éditeur de texte – mais aussi pour les chercheurs qui travaillent sur l'oral et utilisent des transcriptions dont les choix multiples font toujours débat.

Il n'est pas rare aussi que les tenants de l'approche *corpus-driven* (par opposition à ceux qui se réclament de la méthode *corpus-based*) remettent en cause la pertinence et l'utilité de l'annotation, sous prétexte qu'elle repose sur des catégories prédéfinies en fonction d'hypothèses théoriques qu'on surimpose aux données. Cette méthode présenterait le défaut de partir d'une théorie préconstruite pour analyser les données, au lieu de faire émerger de ces données elles-mêmes les catégories et la théorie permettant de les analyser[40]. Le regroupement de plusieurs formes sous une catégorie unique (le lemme par exemple) ferait ainsi courir le risque de passer à côté de divergences ou de particularités qui seraient pourtant pertinentes et même essentielles pour l'analyse. Seul l'examen des formes linguistiques dans leur contexte d'occurrence précis permettrait d'établir les regroupements utiles, la méthode prévoyant qu'on considère *a priori* chaque occurrence lexicale comme unique :

> *« There is a good case for arguing that each distinct form is potentially a unique lexical unit, and that forms should only be conflated into lemmas when their environments show a certain amount and type of similarity ».* (Sinclair 1991 : 8)

Mais les objections de ce type ne me paraissent pas contredire l'utilité de l'annotation linguistique, puisque celle-ci n'a jamais prétendu dispenser le chercheur de l'examen des formes étudiées dans leur contexte d'occurrence. La fonction principale de l'annotation n'est pas – ou pas seulement – de proposer une analyse préconstruite des unités à étudier, mais de permettre le repérage de ces unités, et éventuellement, leur tri et/ou leur filtrage. Nous verrons, dans la section suivante, que cette opération de repérage est particulièrement sensible

[40] La citation suivante donne une bonne idée de l'opinion que les défenseurs de l'approche *corpus-driven* ont des travaux réalisés dans le cadre de l'approche *corpus-based* :

> « *The corpus is considered useful because, on occasions, it indicates where minor corrections and adjustments can be made to the model adopted and, of course, it can also be valuable as a source of quantitative evidence. In this case, however, corpus evidence is brought in as an extra bonus rather than as a determining factor with respect to the analysis, which is still carried out according to pre-existing categories ; although it is used to refine such categories, it is never really in a position to challenge them as there is no claim made that they arise directly from data* » (Tognini-Bonelli 2001 : 66).

et importante dans la recherche diachronique et qu'elle s'effectue souvent de façon progressive, en veillant à restreindre les données sélectionnées par touches successives, que l'on utilise des catégories pré-annotées ou pas.

Je me bornerai à noter ici qu'une pré-annotation permet parfois de sélectionner des formes à côté desquelles on serait peut-être passé sans cette aide. Le cas des démonstratifs médiévaux me servira d'exemple. Il est assez rare que les ouvrages de synthèse (grammaires ou dictionnaires) enregistrent toutes leurs graphies possibles, et tout spécialement la forme *se* (pour *ce*), qui est assez rare mais pas exceptionnelle. Un corpus lemmatisé ou annoté en morphosyntaxe permet d'inclure ces occurrences, ce qui est à peu près impossible sans annotation préalable. Outre qu'on ne pense pas toujours à cette graphie particulière, le fait qu'elle soit identique à celle du réfléchi et de la conjonction de subordination *se* empêche qu'on en tienne compte en pratique, à moins de passer un temps considérable à sélectionner une à une les occurrences recherchées. Les exemples de ce type sont certainement plus abondants pour le français médiéval que pour le français moderne, où les variations graphiques sont bien moindres, mais il en existe toujours. Il n'est peut-être pas anodin qu'une bonne partie de ceux qui, à l'instar de Sinclair, critiquent l'annotation linguistique et les regroupements qu'elle opère, travaillent sur une langue à la morphologie aussi pauvre que l'anglais moderne. L'étude des langues qui possèdent une riche morphologie et/ou une importante variation graphique a, au contraire, tout à gagner d'un pré-repérage des paradigmes.

Je voudrais défendre ici une approche très pragmatique de l'annotation. Il ne fait pas de doute qu'une telle procédure, associée aux outils numériques de traitement et d'analyse, enrichit les processus de recherche et produit, dans certains cas, des résultats qu'il aurait été impossible d'obtenir autrement. Outre ces avantages qui sont évidents aux yeux de tous, il me semble que l'apport de l'annotation linguistique est d'autant plus clair que celle-ci respecte deux grands principes :

1) L'annotation est d'autant plus utile à la recherche que les catégories qu'elle utilise sont partagées par une large communauté de chercheurs. Les informations et les catégories encodées enregistrent une forme d'état de la recherche au moment où l'annotation a lieu. En ce sens, l'annotation permet une sorte de capitalisation du savoir linguistique. Ainsi s'explique probablement que les premières expérimentations aient porté sur le niveau morphosyntaxique des unités, les parties du discours constituant depuis des siècles une sorte de socle commun. Même si ces catégories sont sans cesse discutées et remises en question,

tout chercheur est à même de les utiliser parce qu'il en maîtrise relativement bien le système. Il n'en va pas de même des catégories sémantiques, par exemple, dont le contenu et la forme sont dans l'ensemble moins stables et consensuels.

Toute la difficulté d'une bonne annotation consiste ainsi à trouver l'accord le plus large en ce qui concerne les catégories et la manière de les affecter (cf. la question des standards et des normes au chapitre 3), tout en couvrant la plus grande variété de ressources et d'usages possibles. Il importe en effet de construire un système qui s'adapte à tous les cas de figure, en particulier à ceux qui sont rares et/ou inattendus. L'un des grands principes de la linguistique de corpus étant le principe de complétude (Leech 1992), il est essentiel que le système d'annotation soit utilisé sur un ensemble de données qui n'a pas été présélectionné en fonction des attentes générées par ce système. Dans la pratique, une telle démarche conduit presque toujours à faire face à des situations non prévues. On verra dans la section suivante (3.2.) que le respect de ce principe révèle souvent des zones de discordance entre le modèle préconstruit et les données auxquelles il s'applique et que c'est là aussi l'un des grands apports de l'annotation linguistique.

2) Comme je l'ai indiqué plus haut, la fonction principale de l'annotation est de permettre le repérage, l'extraction, le tri des données et, parfois, la combinaison des informations à exploiter lors de la phase d'analyse. Il importe de faire en sorte que le repérage, en particulier, soit le moins mauvais possible, qu'il évite le bruit (la prise en compte de données non pertinentes) et le silence (l'oubli de données pertinentes). Mais il me paraît surtout nécessaire de distinguer ce repérage de l'analyse proprement dite. L'annotation linguistique constitue un préalable ou, au mieux, une étape dans l'analyse des données linguistiques, et rien n'empêche que les catégories qui ont rendu possible l'extraction des unités linguistiques à étudier soient totalement oubliées ou remises en cause par la suite. Mindt (1991 : 194), par exemple, considère que l'annotation n'est utile que si l'on accepte de prendre le risque de

> « *redefine linguistic classes, regroup cases or reclassify items* ».

Mes expériences personnelles m'ont montré que la phase d'analyse conduit le plus souvent à opérer deux types de modification au moins à l'annotation de départ :

a) Il est très fréquent que l'analyse linguistique conduise à prendre en compte des informations supplémentaires à celles qui ont été encodées dans le corpus. Ces informations sont parfois annotées à leur tour, de façon dynamique, au cours du processus de recherche.

C'est ce que j'ai fait, par exemple, dans l'étude que j'ai consacrée à l'évolution sémantique du démonstratif, pour laquelle j'ai créé un ensemble de 17 étiquettes qui se sont progressivement ajoutées aux catégories morphosyntaxiques du jeu Cattex2009[41].
Cette pratique est d'autant plus courante qu'il est exceptionnel que le corpus soit annoté à tous les niveaux d'analyse possibles (lemme, morphologie, syntaxe, sémantique, pragmatique, etc.). Le chercheur est donc obligé d'ajouter une part d'information au corpus. Si les catégories annotées se superposaient parfaitement aux catégories d'analyse, l'analyse n'apporterait rien de plus.

b) Il arrive souvent aussi que les catégories annotées soient retravaillées pendant l'analyse et qu'elles donnent lieu à des regroupements et/ou à des scissions internes. Ces modifications sont souvent liées, elles aussi, à la prise en compte d'informations supplémentaires ou à des changements de niveau de description. J'illustrerai ce constat à l'aide de deux exemples personnels.

La recherche que nous avons menée sur les éléments préverbaux dans un corpus de textes des 12ème et 13ème siècles a été réalisée grâce à l'annotation morphosyntaxique (parties du discours) et syntaxique (fonctions et relations de dépendance) des mots de ces textes. L'extraction des données étudiées a fait intervenir différents filtres construits à partir des informations annotées : position des unités par rapport au verbe dans la proposition avec restriction aux énoncés comportant au moins deux éléments préverbaux, filtrage des pronoms clitiques, de la négation et des pronoms adverbiaux *en*/*y*, qui ne doivent pas être comptabilisés comme des éléments pleins occupant une position antérieure à la zone verbale. L'ensemble des structures syntaxiques sélectionnées grâce à ces filtres ont ensuite été analysées, mais après un éclatement partiel et une reconfiguration des catégories de départ. Des regroupements ont été opérés, notamment grâce à la réunion des éléments initiaux dissociés dans le modèle mais traditionnellement considérés comme étant extra-phrastiques ou extra-propositionnels (incises, incidentes, apostrophes et interjections). Une analyse et une annotation morphologique et sémantique de quelques éléments ciblés (les circonstants, et parmi eux, les adverbes, subordonnées et SN à valeur temporelle ou spatiale, les adverbes en

[41] Ces étiquettes sont hiérarchisées. Elles comprennent, par exemple, anaphore fidèle, anaphore non fidèle, deixis discursive semi-fidèle, deixis discursive non fidèle avec N sous-spécifié, deixis discursive non fidèle avec N évaluatif, etc.

–ment, etc.) ont été ajoutées, ce qui nous a conduits à scinder la vaste catégorie des circonstants issue de l'annotation initiale.

La recherche que nous avons réalisée en 2011 sur ce qui différentie l'oral représenté du récit a également exploité les étiquettes morphosyntaxiques Cattex2009 dans un corpus de textes médiévaux relativement étendu. Les résultats fournis par le décompte des parties du discours et par les outils statistiques de la plateforme TXM (calcul des spécificité) ont été précisés grâce au filtrage et à l'examen plus approfondi de certaines formes ou constructions (par exemple, les différentes constructions dans lesquelles l'infinitif est employé), grâce à des regroupements d'étiquettes qui étaient distinctes dans le jeu Cattex2009 (les interrogatifs directs, qu'ils soient pronoms, adjectifs ou adverbes), et, dans certains cas, grâce à la combinaison de ces deux opérations (nous avons, par exemple, regroupé différentes parties de discours marquant la personne – pronoms personnels et possessifs – puis avons à nouveau éclaté cet ensemble personne par personne).

Il apparaît donc que le travail d'analyse implique la manipulation des données et des catégories qu'elles comportent. On note, en particulier, que les exemples décrits ci-dessus nécessitent souvent le retour au contexte et l'examen approfondi des formes et de leur entourage. La recherche se développe grâce à la navigation entre différents niveaux de description et grâce à une redéfinition constante des catégories d'analyse. En bout de chaîne, ce travail permet parfois de proposer des catégories inédites.

Les quelques remarques qui précèdent ne prétendent évidemment pas rendre compte de toutes les méthodologies de recherche possibles, mais elles mettent en évidence une certaine façon d'exploiter les corpus annotés. D'autres études procèdent différemment et se basent presque exclusivement sur les catégories qui sont encodées dans le corpus, sans les retravailler ni les enrichir d'informations supplémentaires. Dans ce cas, l'apport de l'analyse peut provenir, d'une part, des données quantitatives qui sont fournies par l'exploitation d'un corpus annoté, d'autre part, de la combinaison originale d'annotations multiples. Dans nos deux recherches sur l'oral représenté, par exemple, nous avons exploité à la fois l'encodage du discours direct (ou oral représenté) dans les textes et l'étiquetage Cattex2009 de tous les mots dans les unités textuelles. Nous avons, par ailleurs, fait appel à divers outils statistiques (calcul des spécificités et analyses factorielles des correspondances) permettant de faire ressortir les fréquences les plus significatives et de faire apparaître les variations internes les plus saillantes.

Les recherches de ce type ne se focalisent pas sur l'étude d'un phénomène précis mais elles exploitent l'ensemble du corpus et son annotation pour permettre aux outils numériques de

dégager des données les corrélations ou les cooccurrences les plus remarquables. La perspective de recherche s'inverse alors avec celle des études visant à décrire un objet linguistique particulier. Les contrastes établis sur la base de critères externes ne sont plus utilisés comme un moyen de décrire des phénomènes linguistiques, mais ce sont les phénomènes linguistiques qui servent à caractériser les parties contrastées. Ce qui passe alors au premier plan, ce sont les corrélations qu'on peut établir entre les caractéristiques linguistiques d'un type de discours[42]. Les travaux de Biber sont menés dans ce cadre et s'attachent à montrer que les relations qu'on peut établir entre catégories sur- et sous-représentées sont particulièrement significatives (d'où les notions de *dimension* et d'*approche multidimensionnelle* introduites et développées dans ces divers travaux, cf. Biber 1989).

3.2. L'apport heuristique de l'annotation

Comme je l'ai indiqué plus haut, l'application concrète d'un système de catégories amène souvent à faire des découvertes et nous renseigne surtout grâce aux zones que le modèle couvre mal.

J'illustrerai ce point au travers de l'exemple des possessifs du français médiéval. La plupart des grammaires distinguent, pour le singulier, les formes toniques (du type *mien*) et les formes atones (du type *mon*). La terminologie employée varie d'une grammaire à l'autre, mais on peut résumer la situation en disant que les formes atones sont considérées comme des déterminants (Zink 1989 parle de « prédéterminant », Moignet 1984 des « articles possessifs et Ménard 1994 les considère comme des « adjectifs »), les formes toniques fonctionnant comme des adjectifs ou des pronoms (Zink les nomme « possessifs prédicatifs », Moignet utilise le terme d'« adjectifs possessifs »[43] et Ménard y voit des « pronoms » et des « adjectifs »)[44].

Les formes atones s'emploient toujours en l'absence d'un autre déterminant, comme en (2), (3) et (4), les formes toniques s'utilisent avec l'article défini comme pronoms possessifs (en (4)) ou comme SN substantivés (en (5)). Les formes toniques sont parfois employées aussi

[42] Ce type d'étude accorde une place tout à fait centrale à la fréquence des phénomènes linguistiques instanciés dans les parties contrastées : « Frequency plays a central role in the analysis, since each dimensions represents a constellation of linguistic features that frequently co-occur in texts » (Biber 2010 : 246).

[43] Moignet compare les formes toniques des possessifs à des adjectifs qualificatifs.

[44] Plusieurs travaux de Wunderli (notamment Wunderli 1977) portent sur les possessifs en français médiéval et proposent une terminologie et une analyse du système un peu différentes. Mais on verra plus loin que ces travaux rendent également compte de manière imparfaite de la distribution des formes.

avec l'article ou un autre déterminant en fonction adjectivale (en (6)) ou sans aucun déterminant en fonction d'attribut du sujet (en (7)) :

(2) Pent a **sun** col **un soen** grant escut let (*Chanson de Roland*, v. 3149)
Il pend à son cou un grand écu large qu'il a

(3) […] s'asist **sa qeue** entre **ses james**… (*Roman de Renart*, branche 1, v. 288)
il s'assit, la queue entre les jambes

(4) **Sa fome** regarde derriere […]
Tout autresi fera **la moie** […] (*Roman de Renart*, branche 1, v. 2060 et 2065)
Sa femme regarde en arrière […] la mienne fera la même chose

(5) **li mien** li ert abandoné (*Roman d'Eneas*, v. 614)
mes biens seront mis à sa disposition.

(6) Li amiralz **la sue** gent apelet (*Chanson de Roland*, v. 3396)
L'émir appelle son peuple

(7) la force n'est mie **soue** (*Roman de Renart*, branche 1, v. 1952)
la force n'est pas à lui / de son côté

La plupart des grammaires signalent également la possibilité de trouver la forme tonique du possessif, non précédée de l'article, devant un nom :

(8) **Pur sue amor** altretel funt li altre (*Chanson de Roland*, v. 3123)
Par amour pour lui, les autres font de même.

Sont rassemblées dans cet ensemble un petit groupe d'expressions figées, relativement peu nombreuses et peu fréquentes, qui assurent une fonction adverbiale dans la proposition : « moie colpe » (*par ma faute*), « tue merci » (*grâce à toi*), « de moie part » (*de ma part*), etc.
Ce rapide survol a pour but de montrer que la syntaxe des formes atones et toniques ne se confond pas. C'est, on peut le remarquer, l'un des rares secteurs grammaticaux du français où une répartition aussi nette semble exister dès le départ, les formes semblant être déjà

fonctionnellement spécialisées comme déterminants ou comme adjectifs/pronoms. Dans sa *Morphologie*, G. Zink le dit très clairement :

> *« Le morphème se modifie selon la place et la fonction que lui assigne la phrase : possessif conjoint, atone et proclitique, prédéterminant du nom à la manière d'un article, de forme ténue ; possessif prédicatif, tonique et autonome, qualifiant et (pro)nominalisable, de forme étoffée ».* (Zink 1989 : 115).

Quelques grammaires constatent en même temps des confusions possibles par suite de l'évolution phonétique des diphtongues dans certaines régions (dans l'aire anglo-normande en particulier). Les diphtongues se réduisant, les formes toniques rejoignent les formes atones, mais continuent de s'utiliser dans les contextes qui leur sont habituellement réservés (*men* est utilisé à la place de *mien*, *ton* à la place *tuen*, etc). C. Buridant est sans conteste l'auteur qui souligne le plus explicitement que « le départ entre formes toniques et formes atones peut être brouillé par le traitement des diphtongues, qui peuvent se réduire » (Buridant 2000 : 151, § 118).

L'étiquetage morphosyntaxique de la *Passion de Clermont* nous a permis de constater que ces confusions se rencontrent dans ce texte particulièrement ancien (sans doute composé vers l'an Mil). Comme on pouvait s'y attendre, on trouve des formes sans diphtongues utilisées comme des formes toniques, notamment après l'article défini :

(9) sant Johan, **lo son** cher amic. (*Passion*, v. 108)
saint Jean, son ami très cher

(10) **los sos** affanz vol remembrar
per que cest mund tot a salvad. (*Passion*, v.3-4)
je veux vous rappeler toutes ses peines par lesquelles il a sauvé tout ce monde.

Buridant (2000 : 149, § 116) signale cette évolution particulière de la diphtongue [oe], réduite à [o] dans les textes de la *Passion* et de la *Vie de saint Léger* (rédigée à la même période et dans le même manuscrit). Ce qui peut étonner, c'est que ces confusions, bien que non systématiques, soient si fréquentes dans la *Passion* (8 possessifs sans diphtongue sur un total de 20 occurrences de la séquence article défini + possessif + nom).

Dans sa *Nouvelle grammaire de l'ancien français*, C. Buridant va plus loin encore et voit dans ces confusions apparentes une tentative avortée de réorganiser tout le système à partir des formes atones, d'où la présence « de formes atones prédéterminées en fonction d'adjectif, correspondant à des tentatives embryonnaires de constituer un système possessif employant les formes atones dans toutes les fonctions (Buridant 2000 : 155, § 122). Et il ajoute plus loin qu'on rencontre aussi « des formes atones en fonction de pronom, répondant à une tendance à employer ces formes dans toutes les positions » (Buridant 2000 : 156, § 123).
Mais ce que les grammaires ne disent jamais, c'est qu'on rencontre très souvent aussi la confusion inverse, c'est-à-dire l'emploi d'une forme lourde (avec diphtongue) à la place d'une forme atone en fonction de déterminant du nom. L'étiquetage morphosyntaxique des possessifs dans la *Passion de Clermont* nous a permis de repérer cette autre anomalie apparente :

(11) sanz Pedre sols segwen lo vai,
quae **sua fin** veder voldrat. (*Passion*, v. 167-168)
saint Pierre, qui veut voir sa fin, est seul à le suivre.

(12) Argent ne aur non i donat
mas que son sang et **soa carn**[45] (*Passion*, v. 385-386)
Il ne leur a donné ni or ni argent, mais uniquement son sang et sa chair

(13) dunc lor gurpit **soe chamisae**
chi sens custurae fo faitice (*Passion*, v. 267-268)
alors il leur abandonna sa chemise qui était faite sans couture

Un rapide sondage dans la Base de français médiéval montre que ce cas n'est pas isolé et qu'on trouve des occurrences similaires dans plusieurs autres textes :

(14) é Deu salvad **suen pople** á cel jur (*Quatre livres des rois*, p. 26)
et Dieu sauva son peuple ce jour là

(15) remembrer de **suen saint testament** (*Psautier de Cambridge*, p. 283)

[45] On note, dans cet énoncé, la coordination de la forme *son* avec la forme *soa*.

rappeler son saint testament

(16) De **soen escrin** qu'ele out od sei. (Adgar, *Le Gracial*, p.197, v. 112)
de son écrin qu'elle avait avec elle

(17) plus que de **sue desverie** (Sanson de Nanteuil, *Proverbes Salemon*, livre 1, p. 48, v. 1576-1577),
plus que de sa folie

(18) par **soen humble contenement** (Henri de Lancastre, *Livre de seyntz medicines*, p.33)
par son comportement humble

On observe que toutes ces constructions sont très différentes des expressions figées mentionnées plus haut (*por sue amor*, etc.) et l'on voit mal, dans l'état actuel des connaissances, ce qui peut justifier la présence de formes toniques dans ces différents énoncés[46].

Les découvertes de ce type ne sont pas rares. Elles dévoilent un apport de l'annotation qu'on peut considérer comme étant secondaire (le but premier de l'annotation n'est généralement pas de repérer ce genre de choses), mais qui n'est pas mineur, puisque ces découvertes ouvrent de nouvelles pistes de recherche.

Ces situations obligent en même temps à faire des choix souvent difficiles et discutables, puisqu'il faut utiliser un système qui correspond imparfaitement aux données qu'il prétend décrire. Dans le cas présent, il fallait décider de donner la priorité aux facteurs morphologiques, en étiquetant comme faibles toutes les formes dépourvues de diphtongue et comme fortes toutes celles qui en contiennent une, ou en donnant la préséance aux arguments syntaxiques, en adaptant l'étiquette à la fonction syntaxique du possessif indépendamment de sa forme. C'est le parti que nous avons pris dans l'étiquetage de la *Passion* et des autres textes de la Base de français médiéval, ce principe étant conforme à notre pratique en général : nous catégorisons comme déterminants ou comme pronoms (et parfois comme adjectifs, pour *tout* par exemple) les démonstratifs et les indéfinis, dont les formes sont indifférenciées et polyfonctionnelles en français médiéval. L'application de cette règle pour l'étiquetage des

[46] Wunderli (1977 : 45) signale, sans donner d'exemple, cet emploi sporadique des formes toniques sans article. Il semble considérer cet usage comme très marginal et n'en propose pas d'explication satisfaisante. On peut noter que tous les textes cités (à l'exception notable de la *Passion de Clermont*) sont anglo-normands.

possessifs offre l'avantage de la clarté et de la cohérence avec les principes qui fondent l'ensemble de notre système. Ce choix a aussi le mérite de rapprocher le paradigme des possessifs de ceux des autres marqueurs grammaticaux, dont ils sont finalement peut-être moins éloignés qu'on ne le pense habituellement[47]…

J'ai voulu défendre ici une approche pratique et réaliste de l'annotation linguistique, les réserves sur son utilité et sur la pertinence des catégories annotées me paraissant relever d'un faux débat : ce qui compte à mes yeux, ce n'est pas tant d'avoir des catégories « parfaites » que de permettre au chercheur de trouver ce qu'il recherche, même si c'est parfois au moyen de stratégies qui lui sembleront détournées, ce qui suppose de lui proposer un système cohérent, stable et appliqué de manière constante (d'où l'importance de la documentation du système et des principes d'annotation, que soulignent tous ceux qui pratiquent ou exploitent l'annotation linguistique).
Je voudrais insister, pour clore cette section, sur deux points. Tout d'abord, l'annotation linguistique correspond en réalité à l'activité habituelle et centrale du linguiste, qui catégorise les données qu'il analyse :

> *« Nonetheless, it is important to note that, setting scale aside, corpus annotation is largely the process of providing – in a systematic and accessible form - those analyses which a linguist would, in all likelihood, carry out anyway on whatever data they work with. »*
> (McEnery & Hardie 2012 : 13)

Ce qu'apporte de plus l'annotation telle qu'elle se pratique de nos jours dans les corpus numériques, c'est qu'elle permet de traiter un volume de données d'une ampleur inégalée et qu'elle contraint le processus de catégorisation à la plus grande systématicité. J'ai pu éprouver à de multiples reprises que l'élaboration d'un système d'annotation oblige à formaliser et souvent à hiérarchiser de manière cohérente les traits et les catégories linguistiques qu'on mobilise lors de l'analyse. En ce sens, ce processus peut être considéré comme une aide à la création d'un système de catégories scientifiques rigoureuses et adaptées aux données.
D'autre part, il est possible aussi d'utiliser les procédures d'annotation pour tester la validité d'hypothèses théoriques de recherche. C'est l'une des voies que suggère G. Leech, après avoir

[47] On rappellera ici que la distinction entre formes atones et formes toniques du possessif est de toute façon limitée aux trois personnes du singulier (sauf en picard).

examiné en détail les rôles respectifs du linguiste et de la machine dans différents processus de recherche et d'enrichissement linguistique. Il propose, en particulier, de tester la validité d'une grammaire linguistique grâce à l'annotation de données attestées au moyen de cette grammaire :

> *« A testing algorithm can use observed corpus data as a means of evaluating the coverage of a grammar or the performance of a parser. »* (Leech 1991 : 23)

Il s'agit d'un domaine que j'ai peu exploré jusqu'à présent, mais qui est sûrement très prometteur pour l'avenir et qui, surtout, permet de répondre à des objections telles que celles de S. Auroux contre la possibilité pour les Sciences du langage de pouvoir se doter de processus de recherche expérimentaux :

> *« La recherche d'attestations dans les textes (quelles que soient la sophistication et l'utilisation de moyens techniques coûteux, voire informatiques), la constitution d'un corpus (aussi longue et compliquée que soit par exemple la constitution d'un corpus des inscriptions étrusques) ne relèvent pas directement des protocoles expérimentaux. A cela deux raisons : i) elles ne sont pas en relation directe avec une hypothèse explicite à tester ; ii) elles ne correspondent pas à la production d'un phénomène. A la rigueur, si on parvient à satisfaire la condition (i) elles peuvent entrer dans un raisonnement empirique »* (Auroux 1998 : 183)

Ce n'est en effet pas la seule recherche d'attestations qui répond à un tel objectif, mais on peut espérer que l'annotation linguistique nous permette de progresser dans cette voie, comme le signalaient déjà Habert et Zweigenbaum au début des années 2000 (2002 : 101) :

> *« Ces corpus enrichis permettent le test d'hypothèses sophistiquées mais aussi permettent de mettre en évidence des phénomènes ou des corrélations inattendues. »*

4. Les outils d'analyse

Les outils informatiques de recherche et d'analyse permettent de réaliser une multitude d'opérations qui participent aux différentes phases de la recherche. Certains d'entre eux permettent de repérer, d'extraire, de trier et de comptabiliser des unités linguistiques ou des suites d'unités dotées ou non de propriétés. D'autres permettent d'enrichir le corpus grâce à l'annotation linguistique de l'ensemble ou d'une partie de ses unités. D'autres encore mobilisent des calculs statistiques sur les fréquences pour mettre en évidence les caractéristiques significatives des données textuelles et, dans certains cas, représenter sous forme de cartographies synthétiques et visuelles les éléments qui s'apparentent ou s'opposent au sein d'un corpus.

Je m'intéresserai, dans ce qui suit, aux premiers et aux derniers de ces outils et ne traiterai pas des opérations d'annotation. Mon objectif sera de montrer dans quels types d'études ces différents outils sont employés et quel a été leur apport dans mes différents travaux de recherche. Pour cela, je distinguerai deux types de recherches, d'une part les études qui portent sur un objet linguistique précis, d'autre part les études qui décrivent les spécificités d'un type de discours. On va voir que les outils que j'ai utilisés dans les deux cas sont assez différents, qu'ils recouvrent la première catégorie pour les premiers, et la troisième pour les seconds, l'élément invariant de toutes ces recherches étant qu'elles reposent toujours sur une approche contrastive et/ou typologique.

4.1. Outils permettant l'étude d'un phénomène linguistique ciblé

La plupart des études que j'ai menées jusqu'à présent ont porté sur un paradigme de formes (les démonstratifs, le déterminant anaphorique *ledit*, les éléments préverbaux) ou sur une forme unique (l'adverbe déictique *ore*, dont les graphies médiévales sont relativement variées). Qu'elles soient diachroniques ou synchroniques, ces études visaient toujours une analyse sémantique et/ou syntaxique de ces unités. Elles reposaient de façon centrale sur l'examen des occurrences en contexte. Mais je voudrais insister ici sur l'importance de la phase préalable de sélection des occurrences.

Les recherches portant sur un phénomène linguistique ciblé supposent évidemment qu'on soit à même de repérer dans le corpus les occurrences du (ou des) paradigme(s) ou de l'unité étudiés. Cette opération de repérage s'effectue généralement à partir des propriétés formelles

du paradigme ou de l'unité et elle peut être médiatisée par les outils numériques mis à notre disposition. C'est là un usage, pour ainsi dire, très « basique » des outils de recherche. Les fonctionnalités « index » et « concordance »[48] (ou « contexte ») de la plateforme TXM permettent de réaliser une telle opération en sélectionnant un ensemble d'occurrences grâce à une équation de recherche[49]. L'« index » permet d'obtenir la liste des formes répondant à cette équation, avec leur fréquence dans le corpus. La « concordance » permet d'obtenir le détail des occurrences avec leur contexte adjacent.

L'une des spécificités de l'approche diachronique est certainement qu'elle accorde une importance toute spéciale à l'apparition ou à la disparition d'une forme ou d'une construction à un moment de l'histoire d'une langue ou d'une variété linguistique. Mais, dans un grand nombre de cas aussi, la recherche diachronique s'intéresse aux changements de fréquence et à l'évolution des conditions d'emploi de l'objet linguistique étudié. L'approche diachronique rejoint en cela la plupart des études synchroniques[50].

Dans tous les cas, il importe principalement de repérer la totalité des occurrences du phénomène observé ([doc. 4] Pincemin *et al*. 2008). Trois raisons au moins expliquent qu'on attache un soin particulier à n'en oublier aucune. D'une part, l'analyse des fréquences impose que soient bien comptabilisées toutes les occurrences dans le corpus. D'autre part, le principe de complétude veut que l'analyse linguistique des occurrences en contexte prenne elle aussi en compte toutes les attestations de l'unité étudiée. Enfin, la datation d'un phénomène ou d'une construction implique qu'on soit à même d'en repérer le plus finement possible la ou les premières occurrences. Ainsi s'explique que le processus de repérage de ces occurrences fasse l'objet d'une attention extrême et qu'il s'effectue le plus souvent par essais successifs en restreignant progressivement le nombre des formes sélectionnées. La fonctionnalité « index », qui permet d'obtenir une simple liste de formes et de fréquences, est particulièrement adaptée pour mettre au point l'équation de recherche qui ajuste au mieux la liste des formes sélectionnées à la liste des formes recherchées.

Il est donc plus prudent, particulièrement pour une langue dont les variations morphologiques et graphiques sont aussi vastes et imprévisibles qu'elles le sont en ancien français, de partir d'une équation de recherche extrêmement large. Dans le cas des démonstratifs par exemple, je me suis limitée au départ à une expression donnant comme seule condition que le mot pouvait

[48] On parle généralement de concordance KWIC (Key Word In Context).

[49] Mes recherches les plus anciennes ont utilisé l'outil Weblex, développé par S. Heiden à l'ENS de Lyon, mais ses fonctionnalités de recherche étaient pour l'essentiel identiques à celles de la plateforme TXM.

[50] Elle continue de se distinguer des études synchroniques en ce qu'elle accorde une importance primordiale au facteur temporel, ce qui la conduit à contraster la description des usages en fonction de tranches temporelles prédéfinies.

commencer (ou pas) par « i » et qu'il était suivi d'un « c » puis d'un nombre d'éléments dont le nombre et la nature étaient indéterminés : « i ?c.* ». Le résultat d'une telle requête est une liste de formes relativement importante, ce qui permet de repérer dans cette liste la quasi-totalité des formes démonstratives instanciées dans le corpus[51]. Il est ensuite possible de fabriquer l'expression de requête souhaitée, en contrôlant que toutes les formes pertinentes sélectionnées grâce à l'expression la plus large sont bien incluses dans les résultats de l'expression la plus restreinte. Il est courant que la liste d'arrivée contienne encore un certain nombre d'intrus et l'on observe que le bruit est généralement mieux toléré que le silence, pour les deux raisons que j'ai invoquées plus haut.

L'expérience montre pourtant qu'il n'est pas toujours matériellement possible de procéder de la sorte et que le silence est, dans certains cas, préférable au bruit. Dans le cas de ma recherche sur les démonstratifs, l'équation de requête donnée plus haut comme étant la plus large possible exclut pourtant dès le départ la graphie *se*. Comme je l'ai souligné déjà, inclure cette forme produit un bruit considérable en raison de l'homographie avec la forme *se* du réfléchi et de la conjonction de subordination, toutes deux particulièrement fréquentes dans les textes médiévaux. La graphie *se* du démonstratif étant au contraire relativement rare, le coût de son inclusion aurait été bien supérieur à celui de son exclusion des formes et des contextes sélectionnés.

L'opération de repérage puis de sélection des occurrences implique souvent, comme on vient de le voir, des choix stratégiques et méthodologiques. Dans le cas d'une recherche sur des morphèmes grammaticaux comme les démonstratifs, dont la fréquence est de toute façon très élevée dans tous les textes, on peut faire le pari qu'il est relativement peu coûteux d'exclure une forme rare. Tel n'est pas le cas, en revanche, si l'on travaille sur des objets peu instanciés dans le corpus. Il sera alors nécessaire soit de faire le choix inverse, soit d'indiquer de manière très explicite les limites des analyses réalisées sur le corpus.

On observe donc qu'une opération apparemment aussi triviale que la recherche d'occurrences repose en réalité sur un raisonnement et sur des choix méthodologiques qui ne sont pas sans conséquences sur les résultats de la recherche. Il est pourtant rare que cette phase du processus de recherche et ces choix soient explicitement exposés dans les publications scientifiques des linguistes. On peut noter à l'inverse le soin avec lequel les psycholinguistes décrivent généralement les conditions dans lesquelles se déroulent leurs expérimentations. Il ne serait probablement pas inutile d'adapter ces pratiques à la linguistique descriptive. C'est

[51] On verra dans le paragraphe suivant que cette expression de requête exclut pourtant l'une des formes possibles du paradigme.

peut-être là une condition nécessaire au développement d'une recherche linguistique qui soit plus expérimentale.

Je ne reviendrai pas ici sur l'utilité de l'annotation linguistique dans ce processus de repérage. C'est grâce à l'étiquetage morphosyntaxique d'un texte que j'ai pu, par exemple, identifier la graphie *se* du démonstratif en français médiéval (voir *supra*). Certaines recherches sont très lourdes à réaliser, voire presque impossibles, en l'absence du pré-repérage réalisé grâce à l'annotation, les recherches sur l'ordre des mots ou des constituants, par exemple, sur l'expression du sujet ou sa non expression, etc.

Outre qu'ils permettent d'élaborer des stratégies de recherche des occurrences à traiter, les outils de recherche et d'analyse offrent également des possibilités de tri de ces occurrences, parfois grâce à des clés de tri multiples (sur la forme ou les propriétés de l'unité recherchée, sur la forme et les propriétés de son entourage). Les fonctionnalités de tri sont bien sûr utiles pour élaguer et nettoyer plus rapidement et avec plus de sûreté les données non pertinentes pour la recherche, mais elles permettent aussi de mettre en évidence les régularités linguistiques qui révèlent la prégnance de certains phénomènes et serviront le plus souvent de base à l'analyse linguistique.

L'approche *corpus-driven* préconise de se fonder sur ces régularités pour faire émerger des données elles-mêmes les catégories permettant de les analyser. Mais il est bien des cas aussi où ce sont les phénomènes rares, les énoncés apparemment « non réguliers », qui résistent à l'analyse et obligent, précisément, à remettre en cause des catégories élaborées pour rendre des comptes des phénomènes les plus courants. Il me semble que c'est là un reproche que font souvent les linguistes de corpus (ou linguistes de l'attesté) aux linguistes de bureau (ou linguistes du possible[52]) : leurs théories linguistiques s'accordent mal avec la réalité des usages dans toute leur variété. L'exploitation des corpus et des données réelles révèle souvent une diversité d'emploi qu'on ne soupçonnait pas et fait émerger des usages non prévus, voire rejetés comme non grammaticaux par les théories linguistiques construites sur des énoncés forgés[53]. Il est assez courant en même temps que ces énoncés apparemment déviants ne soient pas les plus fréquents. Il semble donc important que le linguiste soit capable de repérer, de décrire et d'expliquer les régularités les plus saillantes tout en rendant compte en même temps des usages plus marginaux.

[52] L'expression est reprise à B. Habert (voir notamment Habert 2004 : 12), qui propose cette formulation à la suite des travaux de Fillmore (1992) de Milner (1989) et Corbin (1980).

[53] Il me semble que les reproches qu'on peut faire à ces théories sont doubles : d'un côté, elles reposent le plus souvent sur des énoncés inventés qui ne se rencontrent jamais dans la réalité et paraissent totalement « anormaux » aux linguistes de terrain ; d'un autre côté, elles méconnaissent la grande variété et la grande richesse des usages réels.

C'est ce que je me suis efforcé de faire dans ma recherche sur l'évolution sémantique des démonstratifs en français, en examinant la totalité des occurrences fournies par le corpus. L'analyse des données m'a permis de repérer des usages qu'on savait exister mais dont la fréquence était généralement sous-estimée dans les théories élaborées jusque-là (fréquence d'emploi de la série CIST dans l'anaphore et dans la deixis discursive, par exemple[54]). Elle a confirmé par ailleurs, grâce à des données quantitatives précises, les tendances et les régularités qui avaient déjà été repérées (emploi massif de CIL QUI/DE, exclusion de CIST de l'emploi mémoriel, par exemple). Mais elle a permis aussi d'identifier et d'analyser des énoncés relativement rares et, pour cette raison, peu ou pas représentés dans les études traitant des démonstratifs médiévaux (les occurrences de CIST QUI / DE par exemple). Ce faisant, le but de la recherche a toujours été de proposer des hypothèses qui rendent compte aussi bien des usages les plus rares que des plus fréquents.

L'attention particulière portée aux phénomènes rares ou exceptionnels est une constante de la linguistique diachronique. Les évolutions linguistiques commencent généralement par apparaître dans des contextes restreints et limités en nombre d'occurrences, et ce n'est qu'avec le temps qu'elles peuvent s'étendre et, éventuellement, donner naissance à de véritables changements[55]. Le diachronicien a donc ceci de particulier qu'il attache une importance toute spéciale aux faibles fréquences et qu'il est même souvent amené à les rechercher et à les traquer dans les données[56]. Ainsi s'explique, d'une part, l'attrait que représentent les gros corpus pour le diachronicien ([doc. 5], Guillot *et al*. 2008) et, d'autre part, le soin avec lequel le chercheur s'attache à repérer toutes les occurrences du phénomène étudié, sans en oublier aucune (voir ci-dessus).

---

[54] La théorie de G. Kleiber, la plus répandue à ce jour pour l'analyse du système des démonstratifs médiévaux, suppose en effet que CIST ne peut être employé comme anaphorique qu'en présence de conditions particulières (ce qui n'est pas le cas de la forme non marquée CIL) : « on notera […], que lorsque le SN démonstratif anaphorique renvoie, du fait de son substantif, à un référent qui ne se présente que comme étant identifiable immédiatement, c'est *cist* qui paraît l'emporter » (Kleiber 1987 : 26). Sans l'affirmer de manière explicite, cette théorie conduit à prédire un usage relativement limité de CIST dans l'anaphore et les exemples de CIST anaphorique cités dans l'article de Kleiber sont, de fait, peu nombreux.

[55] Voir les recherches de Kroch sur le changement linguistique et la courbe en S qu'elles ont permis de mettre en évidence (Kroch 1989). Habert & Fuchs (2004), dans leur introduction au numéro du *Français moderne* intitulé *Traitement automatique et ressources numérisées pour le français*, s'intéressent eux aussi aux phénomènes peu instanciés et à l'analyse qu'on peut en faire. Ils citent le changement linguistique comme l'une des causes possibles de ces phénomènes :

> « *Ces phénomènes peuvent relever de causes distinctes : 1) le lapsus ; ii) la violation intentionnelle des règles ; le jeu sur le langage ; iii) la variation interne […] ; iv) le changement* » (Habert &é Fuchs 2004 : 91).

[56] La linguistique diachronique doit aussi savoir comment interpréter l'absence d'attestation, sans qu'il lui soit jamais possible de déterminer avec une parfaite certitude si cette absence de réalisation est due à une véritable impossibilité ou aux hasards de la transmission et de la collecte des données (voir notamment Marchello-Nizia 1995 : 23 et Prévost 2011 : 7-16).

Ce rapide tour d'horizon s'est volontairement limité aux fonctionnalités qui semblent les plus simples et les plus immédiatement utiles au linguiste diachronicien. Il est un point sur lequel je n'ai pas insisté, c'est l'apport des outils numériques à la démarche contrastive et comparative. Ces nouveaux outils permettent en effet de créer et d'utiliser les espaces de comparaison pertinents pour l'analyse, et ils offrent parfois des méthodologies originales pour faire apparaître les similitudes et les dissimilitudes caractérisant ces contextes. Ce sont ces aspects des outils de recherche et d'analyse qui seront développés dans la section suivante.

4.2. Outils permettant l'étude d'un type de discours

J'illustrerai mon propos avec nos travaux de recherche portant sur l'oral représenté au Moyen Âge. Comme je l'ai indiqué à plusieurs reprises déjà, ces recherches ont été rendues possibles par le pré-balisage d'un ensemble de textes relativement nombreux et diversifiés. Les séquences d'oral représenté ont été délimitées grâce aux guillemets insérés dans le texte médiéval par les éditeurs modernes.

Nos recherches avaient pour but de dégager les spécificités de ce type de discours par contraste avec tout ce qui n'est pas de l'oral représenté. Les données qu'on regroupe grossièrement dans ce second ensemble sont très certainement assez hétérogènes et le terme de récit, qu'on utilise parfois en opposition à discours, ne nous a pas paru approprié pour les caractériser. Nous avons donc préféré nous en tenir à une formulation « en creux », qui définit cet ensemble de données par sa seule exclusion de l'oral représenté. C'est cet oral représenté qui est au centre de notre recherche, davantage que la partie du corpus avec laquelle il est comparé.

Les outils qui nous ont permis de mener ces recherches sont implémentés dans la plateforme TXM. Il s'agit du calcul des spécificités, d'une part, de l'analyse factorielle des correspondances, d'autre part. Nous avons tout d'abord employé ces outils à seule fin de faire émerger les éléments caractéristiques des deux types de discours contrastés grâce aux critères externes (guillemets) dans l'ensemble du corpus. Les étiquettes morphosyntaxiques annotées dans les textes nous ont permis de repérer quelles parties du discours étaient limitées à un type de discours (les interrogatifs directs et les termes d'adresse sont restreints à l'oral représenté) et, ce qui est bien plus fréquents, quelles parties du discours étaient sous- ou sur-représentées dans l'un de ces deux types relativement à l'ensemble des occurrences du corpus.

L'un des intérêts majeurs de tels outils est qu'ils permettent de combiner et de synthétiser un nombre de facteurs de variation très important, bien supérieur à ce qu'un humain est capable de traiter à soi seul. Les 58 étiquettes du jeu Cattex2009, par exemple, nous ont permis de faire ressortir les variations les plus saillantes dans les deux parties contrastées. Les travaux de C. Biber sont l'une des meilleures illustrations de ce type de démarche, par le nombre de parties opposées et par la variété des annotations multi-niveaux mobilisées pour construire les faisceaux de traits permettant de typer chacune des parties. Nos recherches sur l'oral représenté ont une visée plus limitée. Elles étaient pour nous une manière d'explorer des outils statistiques de synthèse et de nouvelles méthodologies de recherche.

Je voudrais insister ici sur deux points principaux. D'une part, ces recherches ont été permises grâce à l'outillage mis à notre disposition par la plateforme TXM, mais aussi et surtout grâce à l'aide apportée par nos collègues S. Heiden et B. Pincemin[57] dans la construction des parties contrastées et l'interprétation des résultats. Si les nouveaux outils de ce type peuvent donner l'impression de produire des tableaux et des graphiques synthétiques de manière relativement autonome, l'interprétation de ces éléments repose sur un savoir et des pratiques qui sont malheureusement encore assez peu répandus dans la recherche en linguistique. La coopération avec les informaticiens, les statisticiens et parfois aussi les chercheurs des autres sciences humaines et sociales plus avancés permet d'éviter deux écueils opposés : la crainte et la fuite devant des outils encore peu connus, qu'on ne sait pas comment utiliser et qui paraissent fonctionner de manière opaque comme des boîtes noires ; l'usage non contrôlé de ces mêmes outils et la confiance aveugle qu'ils feront apparaître directement, ou magiquement, ce qui est invisible à l'œil nu du chercheur. Si les outils d'analyse statistique et de synthèse constituent bien une aide et une méthode d'analyse, ils ne la remplacent pas.

D'autre part, l'interprétation des tableaux produits par le calcul de spécificités ou des graphiques construits par l'analyse factorielle des correspondances implique le recours possible, à tout moment, aux données qui ont été synthétisées grâce aux outils. Le maintien d'un lien constant entre les données linguistiques et les résultats produits par les outils d'analyse, mais aussi par les outils de requête abordés dans la section précédente, semble une condition nécessaire à un usage raisonné de ces différents outils à des fins de recherche scientifique[58]. En ce sens, les recherches qui sont ciblées sur un phénomène précis et celles

[57] Nos deux collègues représentent la communauté très large en France et à l'étranger de la *Textométrie*, dont les travaux se développent depuis les années 1970 dans le champ des Sciences humaines et sociales et dont la plateforme TXM est très emblématique (voir notamment le site du projet à l'origine de cette plateforme : http://textometrie.ens-lyon.fr/).

[58] Ce lien permet d'éviter notamment l'effet « boîte noire » mentionné plus haut.

qui visent à caractériser un type de discours par des traits ou des faisceaux de traits linguistiques ne diffèrent pas fondamentalement. Elles se rejoignent dans les modes de navigation auxquels elles font le plus souvent appel entre outils et données de différents types.

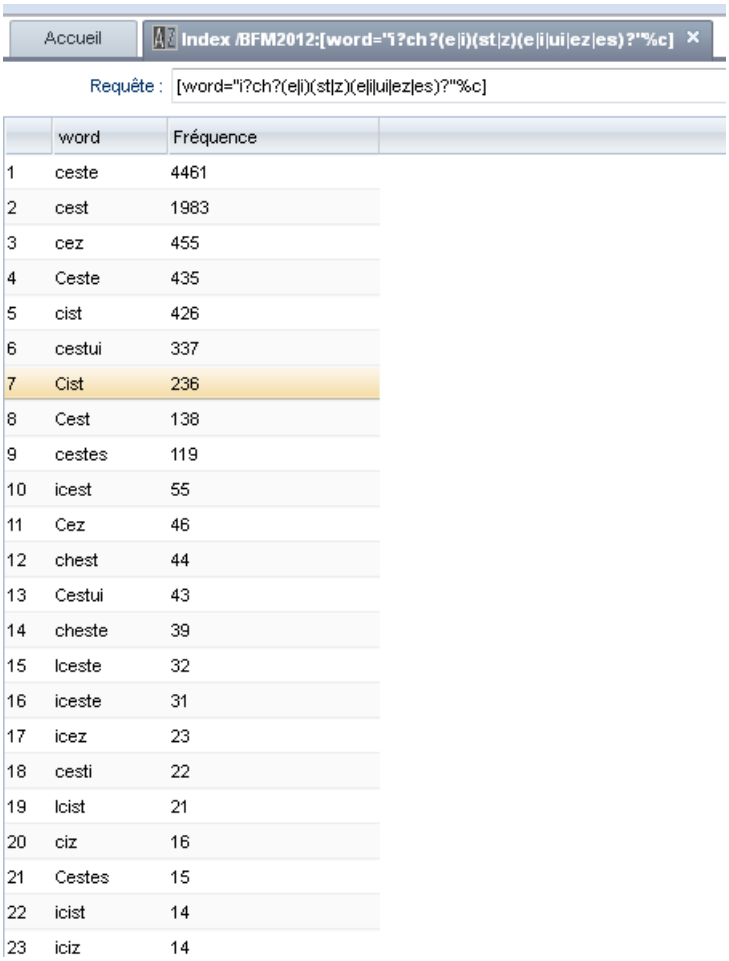
4.3. La navigation entre outils et niveaux d'analyse différents

L'un des principes importants de la méthode textométrique implémentée dans la plateforme TXM est que les outils de recherche et d'analyse veillent à établir des liens hypertextuels permettant l'examen des contextes d'apparition des occurrences traitées. Ce parti pris repose sur le postulat explicitement revendiqué dans cette approche que « l'interprétation des calculs se fonde sur des indicateurs chiffrés mais aussi sur l'examen systématique des contextes, maintenant facilité par des liens hypertextes pertinents » (voir la rubrique *Qu'est-ce que la Textométrie / Présentation* du site Textométrie[59]).

Ces liens permettent de passer, par exemple, d'une ligne d'un « index » de formes à la « concordance » correspondante, ou d'une ligne d'une « concordance » à la page de l'« édition » du texte qui la contient. La Figure 1 reproduit le résultat de la fonctionnalité « index » dans la plateforme TXM. L'expression de requête visait à repérer toutes les graphies possibles de la série CIST des démonstratifs. La commande « index » fournit la liste des formes qui répondent à l'expression de requête et indique les fréquences associées :

[59] http://textometrie.ens-lyon.fr/spip.php?rubrique80.

Figure 1 : Index des occurrences de CIST dans la BFM2012



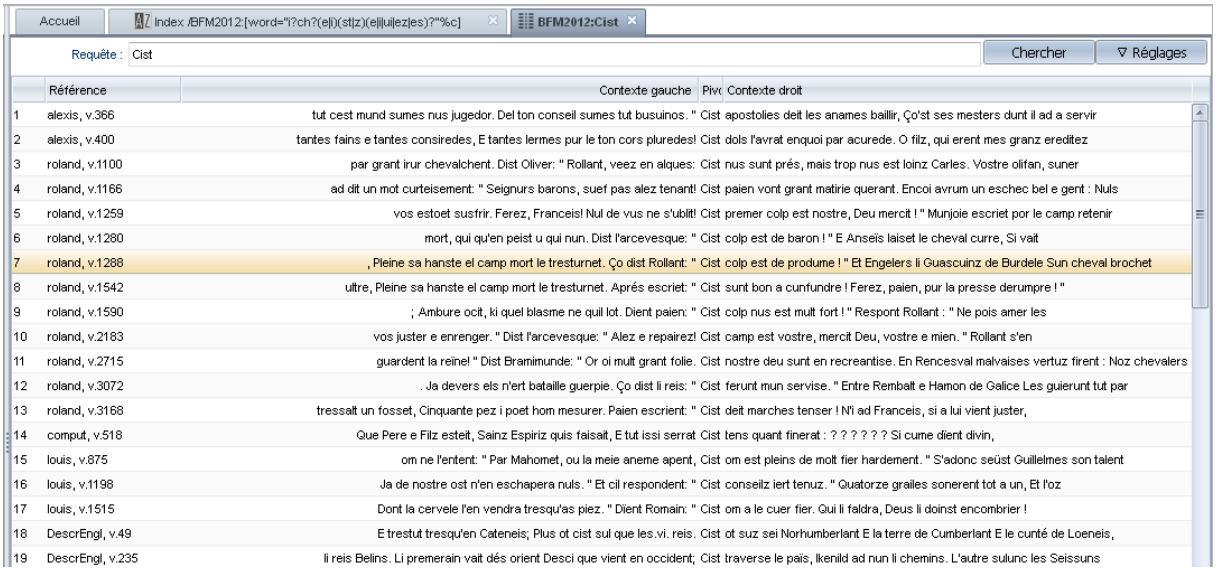
Accueil | Index /BFM2012:[word="i?ch?(e|i)(st|z)(e|i|ui|ez|es)?"%c]

Requête : [word="i?ch?(e|i)(st|z)(e|i|ui|ez|es)?"%c]

	word	Fréquence
1	ceste	4461
2	cest	1983
3	cez	455
4	Ceste	435
5	cist	426
6	cestui	337
7	Cist	236
8	Cest	138
9	cestes	119
10	icest	55
11	Cez	46
12	chest	44
13	Cestui	43
14	cheste	39
15	Iceste	32
16	iceste	31
17	icez	23
18	cesti	22
19	Icist	21
20	ciz	16
21	Cestes	15
22	icist	14
23	iciz	14

L'outil TXM offre la possibilité, en cliquant sur l'une des lignes du tableau (dans la Figure 1, la septième ligne surlignée comportant la forme *Cist*), d'afficher chaque occurrence de la forme dans son contexte (sous la forme d'une « concordance ») :

Figure 2 : Concordance de la forme CIST sélectionnée dans l'Index



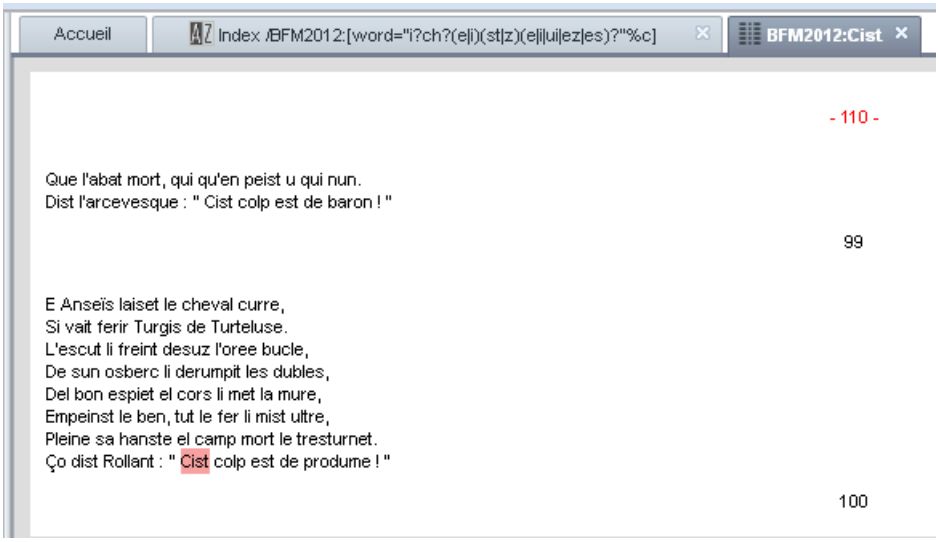
Accueil | Index /BFM2012:[word="i?ch?(e|i)(st|z)(e|i|ui|ez|es)?"%c] | BFM2012:Cist

Requête : Cist — Chercher — Réglages

	Référence	Contexte gauche	Piv	Contexte droit
1	alexis, v.366	tut cest mund sumes nus jugedor. Del ton conseil sumes tut busuinos. "	Cist	apostolies deit les anames baillir, Ço'st ses mesters dunt il ad a servir
2	alexis, v.400	tantes fains e tantes consiredes, E tantes lermes pur le ton cors pluredes!	Cist	dols l'avrat enquoi par acurede. O filz, qui erent mes granz ereditez
3	roland, v.1100	par grant irur chevalchent. Dist Oliver: " Rollant, veez en alques:	Cist	nus sunt prés, mais trop nus est loinz Carles. Vostre olifan, suner
4	roland, v.1166	ad dit un mot curteisement: " Seignurs barons, suef pas alez tenant!	Cist	paien vont grant matirie querant. Encoi avrum un eschec bel e gent : Nuls
5	roland, v.1259	vos estoet susfrir. Ferez, Franceis! Nul de vus ne s'ublit!	Cist	premer colp est nostre, Deu mercit ! " Munjoie escriet por le camp retenir
6	roland, v.1280	mort, qui qu'en peist u qui nun. Dist l'arcevesque: "	Cist	colp est de baron ! " E Anseïs laiset le cheval curre, Si vait
7	roland, v.1288	, Pleine sa hanste el camp mort le tresturnet. Ço dist Rollant: "	Cist	colp est de produme ! " Et Engelers li Guascuinz de Burdele Sun cheval brochet
8	roland, v.1542	ultre, Pleine sa hanste el camp mort le tresturnet. Aprés escriet: "	Cist	sunt bon a cunfundre ! Ferez, paien, pur la presse derumpre ! "
9	roland, v.1590	; Ambure ocit, ki quel blasme ne quil lot. Dient paien: "	Cist	colp nus est mult fort ! " Respont Rollant : " Ne pois amer les
10	roland, v.2183	vos juster e enrenger. " Dist l'arcevesque: " Alez e repairez!	Cist	camp est vostre, mercit Deu, vostre e mien. " Rollant s'en
11	roland, v.2715	guardent la reïne! " Dist Bramimunde: " Or oi mult grant folie.	Cist	nostre deu sunt en recreantise. En Rencesval malvaises vertuz firent : Noz chevalers
12	roland, v.3072	. Ja devers els n'ert bataille guerpie. Ço dist li reis: "	Cist	ferunt mun servise. " Entre Rembalt e Hamon de Galice Les guierunt tut par
13	roland, v.3168	tressalt un fosset, Cinquante pez i poet hom mesurer. Paien escrient: "	Cist	deit marches tenser ! N'i ad Franceis, si a lui vient juster,
14	comput, v.518	Que Pere e Filz esteit, Sainz Espiriz quis faisait, E tut issi serrat	Cist	tens quant finerat : ? ? ? ? ? ? Si cume dient divin,
15	louis, v.875	om ne l'entent: " Par Mahomet, ou la meie aneme apent,	Cist	om est pleins de molt fier hardement. " S'adonc seüst Guillelmes son talent
16	louis, v.1198	Ja de nostre ost n'en eschapera nuls. " Et cil respondent: "	Cist	conseilz iert tenuz. " Quatorze grailes sonerent tot a un, Et l'oz
17	louis, v.1515	Dont la cervele l'en vendra tresqu'as piez. " Dient Romain: "	Cist	om a le cuer fier. Qui li faldra, Deus li doinst encombrier !
18	DescrEngl, v.49	E trestut tresqu'en Cateneis; Plus ot cist sul que les.vi. reis.	Cist	ot suz sei Norhumberlant E la terre de Cumberlant E le cunté de Loeneis,
19	DescrEngl, v.235	li reis Belins. Li premerain vait dés orient Desci que vient en occident;	Cist	traverse le païs, lkenild ad nun li chemins. L'autre sulunc les Seissuns

Selon un principe similaire, l'activation d'une ligne de la « concordance » (dans la Figure 2, la ligne 7 surlignée) permet d'accéder directement à la page correspondante dans l'«édition » du texte. L'occurrence sur laquelle la recherche a porté est colorée :

Figure 3 : Edition de la page qui contient l'occurrence de *Cist* sélectionnée dans la concordance



La plateforme TXM, comme beaucoup des outils du même type, permet également d'utiliser de manière successive et, dans la même session de travail, outils de requête (comme la concordance KWIC), outils d'édition (l'affichage d'une page de texte par exemple) et outils d'analyse (du type du calcul des spécificités ou de l'analyse factorielle des correspondances). La facilité de navigation entre ces diverses fonctionnalités crée des liens structurels entre les différents types d'outils, qui peuvent être mobilisés de façon complémentaire et s'adapter ainsi aux différentes phases de la recherche.

Ces possibilités de navigation existent également entre les multiples niveaux d'analyse des unités linguistiques. Elles permettent de travailler tantôt au niveau de la catégorie morphologique ou syntaxique, tantôt au niveau de la liste des formes, etc.[60]. Sans que l'on prenne toujours le temps d'historiciser ces différentes phases de navigation et de recherche et sans qu'on soit même parfois bien conscient des allers et retours incessants qu'on effectue entre outils et niveaux d'analyse différents, la recherche s'enrichit et progresse de façon dynamique[61]. La circulation rendue possible par l'interface de la plateforme et les hyperliens

[60] Il est bien entendu toujours possible de combiner les formes avec les propriétés qui ont été annotées sur ces formes dans les textes.

[61] Comme je l'ai souligné à propos de l'annotation linguistique, il me semble que l'une des constantes de la recherche en corpus est le (re)modelage des catégories d'analyse des données, ce processus impliquant des regroupements et « dégroupements » multiples au fur et à mesure que les hypothèses s'affinent.

favorise l'interprétation des résultats et l'analyse des données et elle s'intègre pleinement dans le processus de la recherche.

On peut se demander si ces facilités de navigation n'induisent pas par elles-mêmes des évolutions importantes dans le champ de l'analyse des données linguistiques. La possibilité offerte au chercheur d'associer analyse quantitative, en utilisant des outils statistiques plus ou moins perfectionnés, et analyse qualitative, grâce à l'examen des contextes de réalisation des objets repérés et comptabilisés, le conduit à combiner ces deux types d'approche, pourtant longtemps apparus comme opposés (et qui le sont encore dans le discours de certains linguistes).

Les quelques éléments abordés dans cette section peuvent donner l'impression que les nouveaux outils de la recherche numérique n'ont que des avantages et qu'ils s'adaptent à merveille à la recherche scientifique en Sciences du langage, qu'elle soit menée en synchronie ou en diachronie.

Cette présentation, sans doute trop partiale et sommaire, reflète un certain engouement de la linguistique pour des méthodes et des techniques dont elle découvre peu à peu les spécificités et les apports. Là encore, mon parcours personnel n'est peut-être que le reflet d'une évolution plus générale marquant l'ensemble de notre discipline. Quelle que soit l'issue de cette évolution ou de cette simple tendance, la réflexion suscitée par les outils numériques de recherche et d'analyse a le mérite de placer sur le devant de la scène la question des méthodologies propres à la recherche scientifique. Cette réflexion permet en outre de confronter les méthodologies des Sciences du langage à celles d'autres disciplines.

Mais ces avancées méthodologiques ne doivent pas masquer la lourdeur des processus permettant le développement des données numériques et de ces nouvelles technologies (constitution de corpus, annotation linguistique, développement d'outils de recherche et d'analyse). Les initiatives actuelles en faveur de la mise en place d'infrastructures de recherche pour les Sciences humaines et sociales témoignent d'une prise de conscience de l'importance de ces ressources pour l'essor de la recherche en France mais aussi de la nécessité d'en organiser le développement à l'échelle nationale, leur coût, leur diffusion et leur pérennité à long terme conditionnant une bonne partie de leur devenir.

Chapitre 3 : Linguistique diachronique et développement des infrastructures de recherche

Le contexte dans lequel se développe la recherche linguistique en France et à l'étranger est marqué par l'accroissement des ressources numériques, qu'il s'agisse des données linguistiques ou des outils qui permettent de les traiter, et par leur accessibilité toujours plus directe et plus rapide. La multiplication de programmes de recherche de tous ordres, dont l'objectif principal ou l'un des objectifs secondaires est la production de corpus[62], explique en partie la multiplication de ces ressources. Cet essor des ressources numériques peut sembler partiellement anarchique mais il témoigne probablement des mutations profondes qui sont en cours dans le monde de la recherche en Sciences humaines et sociales et en Sciences du langage en particulier.

Ce chapitre sera centré sur l'émergence, l'organisation et les développements actuels de nouvelles infrastructures de recherche par les chercheurs eux-mêmes. J'aimerais montrer que ces infrastructures et les projets collectifs dans lesquels s'insèrent un nombre toujours plus grand de linguistes ont des répercussions importantes sur les méthodologies et les pratiques de recherche, spécialement dans le domaine de la diachronie. Ces infrastructures permettent de faire émerger de nouvelles façons de travailler qui, au bout du compte, infléchissent la recherche scientifique elle-même.

La plupart des points qui seront abordés dans ce chapitre reflètent les réflexions qui sont menées collectivement dans notre équipe de recherche, et je tenterai de les mettre en perspective avec mon propre parcours de recherche. Je puiserai mes exemples dans les réalisations et les projets scientifiques auxquels j'ai participé, ces initiatives ayant généralement conduit à la création de corpus diachroniques (projet BFM, projet ELICO, projet CoRPTeF, projet SRCMF, projet PRESTO). Ce développement des ressources diachroniques met en lumière des questions qui sont très générales et bien connues mais qui doivent se régler d'une manière ou d'une autre dans le cadre d'expériences concrètes.

[62] Je pense naturellement au programme « Corpus et outils de la recherche en Sciences humaines et sociales » de l'Agence nationale de la recherche, mais aussi à tous les programmes de recherche dont la finalité affichée n'est pas de produire un corpus mais dans lesquels la constitution du corpus occupe de fait une place centrale (en raison en particulier des coûts humains et financiers qu'elle engendre).

1. Le maniement des ressources linguistiques

J'ai insisté à plusieurs reprises sur le fait que le chercheur construit son corpus en fonction de ses propres objectifs de recherche. Mais, dans bien des cas aussi, il produit lui-même les ressources qu'il exploite. Le joli néologisme de « corpiste » proposé par B. Habert rend bien compte de cette évolution importante des métiers de la recherche en Sciences humaines et sociales :

> *« nécessité pour le « corpiste » d'assurer tut ou partie des tâches découlant du recours au corpus : normalisation, correction, choix de logiciels, traitements ; etc., au prix parfois d'incohérences et de « bricolages » pas toujours documentés. »* (Habert 2000 : 16)

1.1. La préparation et le formatage des données primaires

Les activités de préparation et de formatage des données primaires ont bien évidemment une finalité pratique : elles sont constitutives de la création et de l'organisation des ressources linguistiques et permettent l'intégration de ces ressources dans les outils de recherche et d'analyse. Ces opérations sont partiellement conditionnées par les logiciels et par les contraintes qu'ils imposent aux données qu'ils sont capables de traiter. Mais ce qui était peut-être perçu au départ comme un ensemble de tâches ingrates et techniques, comportant un risque d'asservissement de la recherche aux exigences de la machine, devient peu à peu un espace de réflexion investi par les chercheurs eux-mêmes.

La préparation et le formatage des données offrent l'occasion de développer une attitude réflexive sur les ressources linguistiques qu'on analyse et ces opérations obligent à se mettre d'accord sur la façon dont on veut représenter, coder et structurer les données primaires. Ces activités très concrètes posent, de fait, le même type de difficultés que celles qui caractérisent l'annotation linguistique et elles s'en rapprochent sur bien des points. La forme qu'on donne aux unités linguistiques, la façon dont on les délimite et dont on les structure reflètent une série de choix, qui peuvent être, selon les cas, plus ou moins conscients, plus ou moins motivés. Il y a plusieurs manières, par exemple, de traiter les coupures de mots et de gérer les unités lexicales complexes telles que « parce que » en français moderne ou médiéval. De même, on peut décider d'intégrer au pas la ponctuation dans les données écrites ou transcrites. Dans un texte médiéval, on peut dissocier les voyelles *-i-* et *-u-* des consonnes *-j-* et *-v-* ou

décider de suivre les habitudes des copistes qui distinguent rarement ces sons et adaptent en général le choix des lettres à leur position dans le mot plutôt qu'au son qu'elles transcrivent, etc.

Le fait qu'un nombre croissant de chercheurs formatent et codent leurs données primaires eux-mêmes leur permet de manipuler ces données et il est probable que cette activité influe sur le rapport qu'ils entretiennent avec leur propre objet d'étude. Cet objet étant appréhendé comme nécessairement construit, il impose un certain relativisme quant aux résultats qu'on en obtient. Mais le traitement des données primaires permet aussi d'expliciter les choix qui sont effectués dès ce niveau de la création du corpus (comme à tous les autres), ce qui conduit peut-être à une plus grande rigueur dans la définition, et les limites, de l'objet d'étude. Il s'agit là d'une forme de manipulation des données, d'un genre très différent de celle à laquelle se livrent les linguistes travaillant sur des énoncés forgés, qui confère à ces choix de codage et de formatage une place et une certaine importance dans le processus de la recherche-même.

Ce travail de codage et de formatage n'est pas sans lien avec une activité très ancienne, celle de l'édition de textes anciens ou modernes. L'activité philologique d'établissement du texte a de tout temps mené à se poser des questions comparables à celles auxquelles nous confronte la production de données numériques à l'heure actuelle. J'en citerai trois grands types, à titre d'exemple. J'essaierai de montrer au travers de ces quelques illustrations les nombreux points de passage entre philologie médiévale et approches linguistiques actuelles, les questions abordées ci-dessous permettant de rapprocher et de faire dialoguer des communautés dont les rapports étaient très distendus depuis de nombreuses années. Là n'est pas le moindre de leurs apports.

Il peut être nécessaire à ce stade de distinguer clairement deux types de corpus numériques diachroniques, ceux qui se basent sur des éditions modernes de textes médiévaux – c'est le cas de la Base de français médiéval –, et ceux qui recourent au manuscrit médiéval pour proposer une édition originale du texte – c'est le cas de l'édition numérique de la *Queste del saint Graal*. On verra dans les sections qui suivent que le producteur de ressources numériques est conduit dans les deux cas à faire des choix du même type, si bien que les deux situations se rejoignent, la production de ressources numériques pouvant s'assimiler à l'édition de ces ressources. Ce qui diffère dans les deux cas, c'est la position que le linguiste a, ou n'a pas, à adopter face à un éditeur à la source des données qu'il manipule.

1.1.1. La segmentation en unités

Une première série de choix est relative aux unités linguistiques et graphiques et à la relation qu'on peut établir entre ces unités de différent niveau. Comme je l'ai dit plus haut, le marquage des frontières lexicales pose la question du traitement des mots composés et des unités complexes. A cela s'ajoutent des problèmes plus spécifiques à la période médiévale. Si dans un grand nombre de cas les normes graphiques et grammaticales nous ont habitués à un découpage consensuel pour le français moderne, la situation est en effet beaucoup moins stable pour le français du Moyen Âge. La plupart des éditeurs de textes adaptent au texte médiéval les règles du découpage moderne, mais avec une constance variable, en particulier lorsqu'une unité moderne provient de deux unités qui étaient tout d'abord autonomes et qui ont pu le rester pendant une période plus ou moins longue avant de se souder (c'est le cas de l'adverbe *jamais* ou de la préposition *parmi*, etc.).

On peut évidemment remettre en cause de tels choix et décider, par exemple, de prêter une attention plus grande à la segmentation graphique suivie par le copiste du manuscrit servant de base à l'édition. Une autre solution – et c'est celle que nous avons a adoptée dans certains cas, peu nombreux et bien repérés – est de dissocier la forme graphique du mot de sa forme linguistique. A une même unité linguistique peuvent très bien correspondre deux ou plusieurs unités graphiques (on peut considérer *parmi* comme une unité linguistique unique, une préposition, et admettre qu'elle s'écrive en deux mots), et inversement, une unité graphique peut parfois regrouper deux unités linguistiques (c'est l'une des conséquences des phénomènes d'enclise particulièrement répandus aux débuts du français : *au* > *a* + *le*, *du* > *de* + *le*, *jou* > *je* + *le*, *nel* > *ne* + *le*, etc.). La question de la délimitation des frontières de mots a ainsi le mérite d'obliger à préciser ce qui relève du niveau graphique et ce qui relève du niveau linguistique de ces unités.

La définition des unités et segments linguistiques dépasse naturellement le cadre du lexique. Le marquage des limites de phrases, en particulier, pose toujours des problèmes très délicats. Il est vrai que ces problèmes sont beaucoup plus apparents pour les périodes anciennes du français (on sait que le concept de *phrase* est assez récent) que pour les écrits modernes, qui peuvent donner l'impression qu'il s'agit là de questions mineures. Mais les choix sur lesquels repose le marquage de ces unités, qui sont à la fois syntaxiques et graphiques, ressurgissent lors de la transcription écrite de données orales. Le bornage des unités de rang inférieur, syntagmes et propositions, se posent en des termes comparables et oblige à se faire une doctrine sur l'usage, ou le non usage, de la ponctuation dans le corpus.

Les habitudes des philologues éditeurs d'œuvres médiévales tendent, là encore, à appliquer aux textes anciens les pratiques actuelles, ce qui facilite la compréhension rapide de ces textes et permet un grand nombre de recherches linguistiques[63] mais donne parfois une image déformée de la syntaxe médiévale (comme le montrent les travaux de B. Combettes sur les constructions détachées, par exemple). Dans le domaine des corpus numériques en revanche, les pratiques sont plus diverses. Une opinion assez répandue est qu'il est préférable d'éviter tout usage de la ponctuation pour laisser au linguiste une plus grande liberté dans l'analyse de la structure syntaxique des énoncés, la ponctuation induisant une lecture et une pré-analyse de cette structure. Cette thèse, qui, dans sa forme la plus extrême, confère au texte la forme d'une simple suite de mots, me semble admise par la plupart des chercheurs travaillant sur l'oral, dont les travaux de C. Blanche-Benveniste et de l'équipe du Gars peuvent apparaître comme emblématiques :

> *« Lorsqu'il s'agit d'étudier la langue parlée, et surtout d'en étudier la syntaxe, on peut envisager de donner un texte sans ponctuation : celle-ci ne viendrait qu'après l'analyse syntaxique […]. La ponctuation, si on la met trop tôt, préjuge de l'analyse syntaxique et impose un découpage sur lequel il est difficile de revenir. »* (Blanche-Benveniste & Jeanjean 1987 : 139)

Certains corpus numériques de textes médiévaux (celui construit par A. Dees dans les années 60, pour ne citer que le plus célèbre et le plus utilisé) suivent le même principe. Cette attitude peut bien sûr se justifier par le refus de plaquer une pré-analyse moderne sur un état de langue ancien, la ponctuation ayant, comme on l'a dit, le « défaut » d'être interprétée comme une segmentation à la fois graphique et syntaxique. Mais il me semble que ce choix résulte plus souvent encore d'une défiance des linguistes envers les pratiques et les habitudes des philologues à l'origine de la ponctuation des éditions[64].

La position que nous avons adoptée pour notre Base de français médiéval est différente. Nous avons choisi de conserver la ponctuation des éditions de textes médiévaux en raison des avantages énumérés plus haut. Les producteurs des éditions scientifiques ayant fait l'effort d'établir une version du texte, nous pensons qu'il est légitime de s'appuyer sur leur

[63] Le repérage des phrases permet, par exemple, d'étudier les éléments qui se trouvent en position initiale, immédiatement après une ponctuation forte.

[64] Il est possible aussi que le choix de nettoyer les textes de toute ponctuation ait été motivé au départ par des contraintes techniques. Ces contraintes ont aujourd'hui disparu, si bien que les choix reflètent uniquement la volonté du chercheur.

interprétation de ce texte. La ponctuation n'est de toute façon qu'un élément parmi beaucoup d'autres de leur propre lecture du manuscrit médiéval. A ce titre, elle reste bien sûr sujette à discussion et peut être remise en cause de façon plus ou moins sporadique, en particulier lorsqu'une annotation syntaxique se superpose aux données primaires, mais sans que cela implique de la rejeter par principe et de façon systématique.

Ce qui nous semble en revanche bien plus fondamental, c'est d'infléchir les pratiques des philologues vers un respect plus systématique du manuscrit qui sert de base à l'édition. Les quelques études menées récemment sur la ponctuation médiévale (Lavrentiev 2009, Mazziotta 2009) montrent que ce système, même s'il est assez différent du nôtre et qu'il laisse une large place à la variation, nous renseigne sur les pratiques médiévales de l'écrit. La ponctuation médiévale peut être un complément utile à l'analyse linguistique des énoncés et il serait dommage de s'en priver. Elle constitue par ailleurs en elle-même un objet de recherche particulièrement riche.

Les choix sur lesquels repose la représentation des données primaires, spécialement dans le domaine de la segmentation des unités linguistiques, permettent ainsi de renouveler les échanges entre linguistes, qui étaient jusqu'à présent surtout des utilisateurs des éditions de textes, et ceux dont le métier a toujours été de produire ces éditions. On peut souhaiter qu'il en résulte, d'une part, une attitude plus réflexive et plus critique des linguistes diachroniciens sur les données qu'ils analysent, d'autre part, une approche renouvelée de l'édition des textes médiévaux grâce aux apports de la linguistique et aux questions que les linguistes posent aux textes.

L'un des apports du numérique est également qu'il permet de gérer plus finement et de manière indépendante plusieurs niveaux de représentation et d'analyse des données. De même qu'il est possible de distinguer l'unité graphique du mot de l'unité lexicale à laquelle il correspond, de même, on peut très bien envisager de distinguer différents types de découpage des unités de plus grande taille. Rien n'interdit, par exemple, d'intégrer dans le texte numérique différents niveaux de ponctuation (la ponctuation du manuscrit, la ponctuation de l'éditeur). La segmentation de la ponctuation peut très bien s'ajouter à un découpage syntaxique des unités linguistiques, etc. La souplesse du numérique permet au chercheur de dissocier autant de niveaux de représentation et d'analyse qu'il le souhaite. Il peut ainsi adapter ses choix à ses propres objets de recherche et également tenir compte d'autres usages que les siens[65].

[65] C'est l'une des visées de notre édition électronique de la *Queste del saint Graal*. La production de données numériques offre l'occasion au chercheur de s'intéresser à d'autres usages du corpus, à ceux du grand public par

Tout ce qui vient d'être dit sur la segmentation en phrases et l'usage de la ponctuation dans les corpus numériques vaut bien entendu pour les unités de rang supérieur. Les pratiques de l'édition de texte gagneraient, dans ce domaine aussi, à suivre avec plus de fidélité ce qui se trouve dans les manuscrits, en accordant plus d'attention aux pieds de mouche[66] par exemple. Comme on l'a vu plus haut, la recherche diachronique pourrait utilement mettre à profit l'étude de la structuration textuelle et de la mise en page du manuscrit médiéval.

1.1.2. Les graphies

Le second grand volet des choix d'encodage et de formatage concerne la forme graphique des données primaires. Là encore, le rapprochement de l'édition de textes médiévaux et de la transcription de données orales permet de mettre en lumière des points de discussion qui sont communs à ces deux activités et probablement assez universels mais qui prennent une acuité particulière lorsqu'il s'agit de traiter ces données particulières.

L'une des difficultés majeures pour les textes anciens, et spécialement pour ceux qui sont antérieurs à l'adoption des normes orthographiques régissant les imprimés, est la variation graphique. On sait que cette variation est inhérente aux textes, et plus largement à la culture du Moyen Âge. Cette tendance forte à l'hétérographie est directement tributaire aussi de l'importance de la variation dialectale à cette même période de l'histoire du français, les graphies enregistrant des réalisations phoniques qui pouvaient être assez différentes, même si elles le font de manière imparfaite et en réduisant dans des proportions qui peuvent varier les particularismes locaux pour donner accès à une variété écrite du français assez largement partagée.

Les choix en matière graphique posent donc à leur manière la question du rapport de la graphie et de la prononciation. Les linguistes travaillant sur l'oral sont confrontés aux mêmes questions (transcription phonétique ou orthographique, normes orthographiques adoptées). Dans le champ des études médiévales, le consensus le plus large veut qu'on respecte les graphies du manuscrit dans toute leur variété. Mais, comme je l'ai signalé plus haut, ce

---

exemple. Je n'aborderai pas davantage ce point, même si cette ouverture vers d'autres publics me semble importante pour l'évolution de la recherche. L'essor du numérique conduit en effet les chercheurs à faire preuve d'une attitude plus collective et plus communautaire en produisant des données qui serviront à leurs collègues (voir la section 4 de ce chapitre), mais il offre peut-être aussi un nouveau moyen de sortir le monde de la recherche de son relatif isolement.

[66] Ces marques graphiques, qui sont à l'origine du signe de fin de paragraphe dans les outils numériques modernes, scandent de façon plus ou moins régulière les folios des manuscrits médiévaux. Elles jouent le rôle de bornes graphiques.

principe a des conséquences importantes sur les stratégies et les possibilités de requête. Il rend la sélection d'un paradigme de formes beaucoup plus difficile et décuple d'autant l'intérêt d'une annotation morphosyntaxique ou d'une lemmatisation du texte. Et ce principe peut toujours être remis en cause par le chercheur en fonction de ses propres objectifs de recherche.

La seconde grande question à laquelle doivent répondre les éditeurs de corpus numériques concerne le traitement des abréviations médiévales. La plupart des éditions de textes médiévaux proposent une résolution de ces abréviations et, le plus souvent, elles ne signalent pas clairement quels éléments graphiques sont transcrits à partir du manuscrit et quels éléments sont ajoutés par l'éditeur. Pour certaines études linguistiques, il peut pourtant être très utile, voire indispensable, de disposer de cette information. Il serait hasardeux, par exemple, d'étudier la distribution des marques casuelles dans un texte médiéval sans savoir quel pourcentage de formes abrégées contient le manuscrit et quels ont été les principes et les lieux de résolution des abréviations par l'éditeur. L'émergence des corpus numériques et la part qu'ils prennent à leur constitution permet aux chercheurs diachroniciens de prendre la pleine mesure de ces choix et des conséquences qu'ils peuvent avoir sur l'analyse des données.

1.1.3. La délimitation du texte

J'aborderai très rapidement pour clore cette section le problème de la délimitation du texte proprement dit, de sa séparation d'autres éléments qu'on peut hésiter à lui rattacher ou à l'en extraire. Je pense notamment aux peintures et miniatures médiévales, mais aussi aux gloses, titres et rubriques diverses qui peuvent être en plus ou moins grand nombre dans les manuscrits médiévaux. Ma courte expérience en matière d'édition de texte m'a montré que toutes sortes d'éléments au statut incertain peuvent s'immiscer dans le manuscrit sans qu'on sache toujours très bien quelle place leur accorder dans le document numérique.

Le sixième livre de la traduction du *De casibus* de Boccace par Laurent de Premierfait au début du 15ème siècle, par exemple, contient un dialogue fictif entre la figure allégorique de Fortune et l'auteur du livre. Plusieurs des manuscrits que j'ai consultés pour mon édition de ce texte contiennent des mentions du type « Fortune parle » et « l'auteur parle », qui sont soit insérées dans le texte-même – et dans ce cas la mention est mise en évidence grâce à des caractères rubriqués (manuscrit BnF fr. 226) –, soit en marge du texte (manuscrit Arsenal 5193). Dans son étude sur la signalisation du discours rapporté dans les manuscrits

médiévaux, E. Llamas-Pombo (2010) étudie les manuscrits du *Roman de la Rose* et de l'*Ovide moralisé* et montre qu'il n'est pas rare qu'y soit également représentée la voix de l'auteur (ou du narrateur) par des rubriques du type *Ovide acteur*, *Ovide parle*. Ce type d'insertion n'est pas aussi marginal qu'on pourrait l'imaginer. Il oblige l'éditeur à assigner un rôle à des éléments qu'on peut considérer comme étant internes au texte mais qui assurent en même temps une fonction de balisage et de structuration du document manuscrit, à la manière de certaines marques de mise en page modernes.

Sans prétendre à l'exhaustivité, ce rapide survol vise surtout à illustrer au travers de quelques exemples concrets de quelle manière la création de ressources numériques renouvelle la prise de conscience par les chercheurs de la matérialité des données qu'ils exploitent et de la façon dont celles-ci sont construites. Cette familiarité avec les données et leur mode de constitution est particulièrement utile au linguiste diachronicien travaillant sur des périodes aussi éloignées et étrangères à la nôtre que l'est la période médiévale et sur des manuscrits dont les caractéristiques matérielles sont parfois très différentes de celles des imprimés modernes. J'espère avoir montré aussi, grâce aux quelques points abordés ici, qu'il est impossible d'établir une séparation très nette entre la représentation des données et leur annotation linguistique.

L'étude de la langue du Moyen Âge menée grâce aux sources manuscrites qui se sont conservées jusqu'à aujourd'hui a, d'une certaine manière, l'avantage de favoriser la mise à distance de l'objet d'étude. C'est donc tout naturellement que, dans ses premiers travaux sur l'oral, l'équipe du Gars a rapproché son travail sur les normes de transcription de l'activité des philologues éditeurs de textes médiévaux (voir notamment la préface de J. Monfrin à l'ouvrage de Blanche-Benveniste & Jeanjean). Mais, au-delà des rapports assez étroits qu'on peut établir entre deux communautés qui sont, par la force des choses, obligées d'inventer ou de créer de façon consciente les données qu'elles manipulent, il me semble que les points abordés ici soulèvent des questions qui concernent la recherche en linguistique en général et qui sont peu à peu intégrées par les chercheurs, qu'ils soient directement ou pas producteurs des ressources sur lesquelles ils travaillent.

1.2. La normalisation des formats

La normalisation des formats répond à des objectifs multiples : d'une part, recenser l'existant grâce à des catégories communes, d'autre part échanger les ressources avec d'autres partenaires et les rendre compatibles avec une large gamme d'outils, et enfin, pérenniser ces ressources autant que faire se peut. Dans ces différents aspects, l'élaboration de normes communes et de standards favorise et conditionne partiellement le développement d'une recherche cumulative ([doc. 1] Heiden & Guillot 2003).

Les coûts engendrés par le traitement des données textuelles et par leur équipement informatique expliquent également l'importance de la circulation et de la sauvegarde de ces ressources. Des échanges et de l'inter-opérabilité des corpus dépend la possibilité de les réutiliser et de les enrichir par phases successives. Plusieurs initiatives nationales (en particulier, les Très Grandes Infrastructures de Recherche) et internationales (notamment, pour le Moyen Âge, le Consortium international pour les corpus de français médiéval) visent à répondre à ces objectifs. L'une de leurs principales tâches est la définition et la diffusion de normes et de recommandations ou guides de bonnes pratiques à l'adresse de ceux qui créent ou utilisent des ressources numériques (voir notamment le guide des bonnes pratiques pour les corpus oraux dirigé par O. Baude en 2008).

Cet effort de normalisation peut s'appliquer à différents niveaux. Il peut concerner la description externe des ressources grâce aux métadonnées qui permettent de les caractériser en bloc. L'adoption de normes communes permet dans ce cas d'identifier et de recenser les données linguistiques dont chacun dispose et de favoriser les échanges. Elle permet également la création de bases bibliographiques communes et de portails rassemblant les métadonnées décrivant les ressources accessibles en différents points.

La normalisation des métadonnées repose sur la création de jeux de catégories communes (Dublin Core, etc.), mais aussi sur l'adoption d'un identifiant unique et partagé pour chaque ressource numérique. Dans le domaine des corpus médiévaux, la bibliographie en ligne du *Dictionnaire étymologique de l'ancien français* propose un sigle unique pour chaque édition de texte ancien. Ces sigles étant déjà habituellement repris et cités par les médiévistes, nous les avons adoptés comme identifiants des éditions incluses dans notre Base de français médiéval.

J'ai présenté plus haut le travail réalisé dans le cadre du projet CoRPTeF sur les descripteurs de textes. Cette initiative collective a permis la réalisation d'une fiche de description

présentée par le Consortium international pour les corpus de français médiéval comme étant la fiche minimale adaptée à la description du texte médiéval. Cette fiche est accessible en ligne (http://ccfm.ens-lyon.fr/spip.php?article13) et a vocation à être partagée par toute la communauté des médiévistes. Elle est complétée, pour certains de ces champs, par les typologies élaborées dans le cadre du projet CoRPTeF. J'ai déjà insisté sur la liste des domaines et des genres de la BFM, qui a inspiré plusieurs autres programmes de recherche et qui constitue une sorte de socle commun librement adaptable par chacun. Si ces typologies n'ont pas de statut affirmé de norme ou de standard, elles servent souvent de base à ceux qui se lancent dans la constitution de corpus de français médiéval.

Ces normes propres aux corpus médiévaux permettent les échanges au sein d'une communauté assez vaste, puisqu'elle rassemble des linguistes, des philologues, des littéraires et des historiens. Les normes établies dans un cadre plus large encore ouvrent les usages et les échanges à d'autres disciplines et d'autres publics. Une initiative en cours du Consortium CAHIER (http://www.cahier.paris-sorbonne.fr/) vise ainsi à constituer un entrepôt OAI recensant les ressources produites par l'ensemble de ses membres. Cet entrepôt permettra que ces ressources soient connues et recensées dans les catalogues des grandes bibliothèques publiques par exemple.

L'activité de normalisation et de standardisation concerne naturellement aussi les ressources linguistiques elles-mêmes. Elle s'applique en premier lieu au codage des textes et aux choix de représentation des données textuelles. La *Text Encoding Initiative* (TEI) constitue certainement à ce jour l'une des plus belles réussites en matière de standardisation des pratiques de codage. Depuis 1987, ce consortium international œuvre à la création et à la diffusion de recommandations décrites dans des Guidelines (http://www.tei-c.org/index.xml) qui sont utilisés par un très large éventail de disciplines académiques et de bibliothèques.

Nous avons pour notre part adopté certaines de ses balises lors d'un échange de textes avec l'UMR ATILF en 2002, et, depuis cette date, notre codage s'est enrichi d'année en année, en suivant les évolutions de nos besoins et de nos objectifs de recherche (voir notre *Manuel d'encodage XML-TEI des textes de la Base de Français Médiéval* (Heiden *et al.*), http://bfm.ens-lyon.fr/article.php3?id_article=158). Nous avons récemment intégré aux textes de la BFM, par exemple, le codage des limites du discours direct. L'une des caractéristiques des recommandations de la TEI est qu'elles ne se présentent pas comme un catalogue clos et indivisible, mais plutôt comme un système cohérent dont les éléments sont détachables et librement utilisables en fonction des besoins de chacun. Les évolutions de ce système sont

proposées et validées par les chercheurs réunis dans le consortium[67]. Elles émanent des producteurs et utilisateurs de données numériques-mêmes, et non d'autorités extérieures plus ou moins habilitées à imposer des normes. Là se trouve probablement l'un des atouts et des succès de cette initiative.

Les balises de la TEI recouvrent un très vaste ensemble d'informations, qui sont pour certaines encodées dans l'en-tête des textes (diverses métadonnées du type de celles évoquées aux paragraphes précédents) et pour d'autres dans le corps-même de ces textes. Les balises permettent en particulier de structurer le document numérique en encodant les différents niveaux de division présents dans le texte : prologue, livres, chapitres, paragraphes, pages, ligne ou vers, mais aussi les prises de paroles des divers locuteurs dans le théâtre par exemple. Elles permettent également de rendre compte des corrections et interventions de l'éditeur scientifique, généralement représentées dans le texte grâce à des mises en formes typographiques : utilisation des italiques pour indiquer les résolutions des abréviations, de crochets ou de parenthèses pour indiquer un changement de manuscrit, une correction de l'éditeur, etc. Les informations ainsi encodées peuvent concerner des unités de taille diverse : groupes de mots (dans le cas d'éléments structurels par exemple), mots, parties de mots.

L'un des principes du balisage TEI de la BFM est qu'il vise à décrire le texte tel qu'il se présente dans l'édition-source et, parfois aussi, à rendre compte des corrections proposées par notre équipe. Les balises de la TEI ont surtout l'énorme avantage qu'elles permettent de gérer différents niveaux et différents choix de représentation des données. Elles sont donc particulièrement adaptées à la création d'éditions multi-facettes permettant de visualiser et d'exploiter le texte sous différents angles, sous une forme normalisée ou plus proche de celle du manuscrit par exemple. Mon objectif n'est pas d'entrer ici dans le détail de ces procédés techniques[68], mais plutôt de montrer que les outils mis à disposition par le numérique, dont les normes de codage font d'une certaine façon partie, permettent de développer des modes de représentation et d'édition du texte qui étaient hors de portée du support papier et qui servent de manière très directe le processus de la recherche linguistique.

Le travail de normalisation touche naturellement aussi l'annotation linguistique des textes. Comme j'ai essayé de le défendre plus haut, l'annotation paraît d'autant plus utile qu'elle fait appel à des catégories partagées. Mais c'est probablement dans ce domaine que la normalisation est la plus difficile, chaque courant théorique, chaque type d'exploitation

[67] La version la plus récente de la TEI à ce jour est la version P5 (depuis 2007). C'est celle que nous suivons dans le codage des textes de la Base de français médiéval.

[68] J'ai la chance de travailler avec A. Lavrentiev, grand expert des balises TEI, de leur forme, de leur sémantique et des usages que les outils et leschercheurs peuvent en faire pour la recherche.

linguistique et presque chaque linguiste ayant ses propres besoins et ses propres options d'annotation. Les difficultés rencontrées dans ce domaine expliquent sans doute qu'il dépasse le champ d'action propre de la TEI, qui ne propose pas, pour l'instant, de jeux d'étiquettes morphosyntaxiques ou de modèle syntaxique commun par exemple.

L'expérience montre qu'il y a une tension permanente entre le désir de créer des catégories qui s'adaptent, d'une part, à la langue qu'on étudie (peut-on réellement utiliser les mêmes catégories en français moderne et médiéval ?), d'autre part, aux objectifs de recherche qu'on s'est fixés et le souhait par ailleurs d'établir des liens et des passerelles avec les catégories utilisées par d'autres (jeux d'étiquettes pivots, etc.).

La recherche diachronique offre de ce point de vue des difficultés supplémentaires, puisqu'elle conduit à représenter et à accorder une place, d'une manière ou d'une autre, au changement linguistique. Les corpus diachroniques doivent permettre d'étudier les mutations, tout en offrant des catégories relativement stables ou en trouvant le moyen d'articuler différents systèmes liés à des tranches temporelles prédéfinies. La plupart des expériences auxquelles j'ai participé ont plutôt pris le parti d'adapter aux différents états du français un jeu de catégories unique et relativement panchronique[69]. Le projet de lemmatisation du français médiéval auquel nous travaillons actuellement dans le cadre du programme PRESTO, par exemple, nous conduit à privilégier l'emploi de lemmes modernes permettant de créer et d'exploiter un corpus couvrant la quasi-totalité de l'histoire du français. Il semble que les chercheurs travaillant en diachronie tendent, pour l'instant et de manière très empirique, à privilégier les systèmes à large couverture afin d'éviter la multiplication des jeux de catégories et le passage de l'un à l'autre.

Ce travail de normalisation est douloureux, fait de compromis jamais satisfaisants mais nécessaires. Il me semble que les expériences actuelles tâtonnent encore. Mais je crois qu'on peut voir chaque avancée dans le domaine de la normalisation, vue comme la stabilisation à un instant t de catégories de compromis, comme un palier dans le développement d'une recherche plus empirique.

Il est probable qu'au travers des diverses activités que je viens de décrire et qui sont de plus en plus souvent assumées par les chercheurs eux-mêmes (préparation et formatage des

[69] Cette même tendance se manifeste aussi bien au niveau des métadonnées (avec des typologies en genres identiques à toutes les périodes par exemple) qu'au niveau des catégories d'annotation linguistique des éléments internes au texte.

données primaires, normalisation et standardisation des formats), les frontières entre chercheurs et ingénieurs évoluent et se transforment. Les chercheurs entrant en contact direct avec les données, les ingénieurs ne sont plus vus comme de simples intermédiaires ou prestataires. Dans le même temps, certains informaticiens se spécialisent en ingénierie linguistique et développent une grande expertise dans les méthodologies de la recherche linguistique. Un véritable dialogue s'instaure, qui ne repose plus aussi clairement qu'auparavant sur une répartition des tâches spécialisées (même s'il reste beaucoup de choses que seuls les ingénieurs savent faire), mais sur des approches différentes et complémentaires.

J'ai pour ma part beaucoup appris au contact de mes collègues et je leur dois beaucoup. Il me semble que c'est dans cette diversité d'approches que réside le principal apport du numérique et de ses outils, et non dans le fait d'offrir de simples techniques. Et si la linguistique parvient un jour à devenir une discipline plus expérimentale, je crois qu'elle le devra en grande partie à l'action concertée des ingénieurs et des chercheurs.

Je n'ai pas abordé la question des ressources logicielles et de leur rôle dans le développement de nouvelles infrastructures de recherche. Dans ce domaine aussi, les échanges avec les informaticiens permettent à la fois d'influer sur le développement des logiciels et d'y participer (en définissant le plus clairement possible nos besoins, en rédigeant des manuels d'utilisation par exemple), mais aussi de découvrir et d'exploiter de nouvelles méthodologies de recherche et d'analyse (dans mon cas, des outils statistiques comme le calcul des spécificités ou les analyses factorielles des correspondances). Je ne développerai pas davantage ce point ici et m'intéresserai à présent aux questions que pose actuellement la diffusion des ressources numériques, qu'il s'agisse des corpus proprement dits ou de l'outillage qui leur est associé.

2. La politique de diffusion des ressources

J'axerai mon propos sur la protection juridique des productions scientifiques. Cet aspect de la recherche a longtemps été minoré, voire méprisé, par les Sciences humaines et sociales mais il a pris récemment une importance nouvelle sous l'impulsion du développement du numérique et des évolutions en cours dans le champ de l'édition scientifique.

La protection juridique des données scientifiques concerne les publications académiques et la littérature grise, aussi bien que les corpus et outils numériques. Dans tous ces domaines, le monde universitaire et les pouvoirs publics ont récemment pris conscience – parfois à leur

corps défendant – de l'importance des questions juridiques pour l'avenir de la recherche en France comme à l'étranger. Il me semble qu'au milieu de toutes les évolutions en cours – et elles sont nombreuses et relativement rapides – deux grands principes doivent être retenus : le principe de l'ouverture d'une part, le principe de la paternité d'autre part.

2.1. Le principe de l'ouverture

Il s'agit là d'une condition nécessaire au développement de la recherche scientifique. Si les bouleversements engendrés par le numérique obligent à inventer de nouveaux modes de protection, ils ouvrent également des possibilités de diffusion jusque là inégalées, grâce en particulier à l'accès en ligne. Dans ces conditions, l'ouverture des données et la libre circulation du savoir deviennent d'autant plus décisives pour l'essor de la recherche. La protection juridique des ressources et des résultats de la recherche ne peut en aucun cas les entraver.

Ce contexte a permis le développement récent des mouvements de l'Open-Access (ou libre accès à l'information scientifique et technique) et de l'Open-Source (qui défend la publication et le partage du code informatique des logiciels). La volonté de promouvoir l'ouverture et l'action militante des chercheurs face à un monde de l'édition devenu parfois hostile ont également eu pour conséquence de créer de nouvelles solidarités et de nouveaux réseaux. On peut espérer que de cette action collective naîtront dans un avenir proche de nouvelles formes de relation et de collaboration avec le monde de l'édition privée.

C'est du moins ce que j'ai retenu de ma propre expérience, marquée tout d'abord par un affrontement brutal avec l'une des maisons spécialisées dans l'édition scientifique de textes médiévaux. Une fois l'accord trouvé, mais sur des bases anciennes et peu satisfaisantes du point de vue académique, nous avons cherché à développer de nouvelles pratiques et une nouvelle politique de diffusion qui, tout en reconnaissant les droits des personnes à la source des données diffusées, permettent en même temps au plus grand nombre d'en profiter. Il est clair pour nous à présent qu'une recherche qui produirait ou qui serait réalisée à partir de ressources privées ou difficilement accessibles n'a que peu d'intérêt et peu de chance de se disséminer. La maîtrise du statut juridique des ressources (linguistiques et logicielles) et leur ouverture la plus large possible acquièrent peu à peu une importance égale à celle de leur qualité intrinsèque.

Le développement du numérique et des nouveaux moyens de communication a ainsi bouleversé des modes de relation bien établis avec ceux dont le métier était jusque-là de publier et de diffuser la recherche scientifique, qu'ils soient éditeurs publics ou privés. Les chercheurs ayant, d'une certaine façon, un temps d'avance dans le numérique et les outils de recherche associés, ils tendent de plus en plus à éditer eux-mêmes les ressources qu'ils produisent en se passant des éditeurs commerciaux. Les frontières semblent se brouiller, les métiers se redéfinissent tant bien que mal et des tensions se créent, mais il est désormais bien établi que les chercheurs (et les pouvoirs publics) ne supportent plus que leur communauté soit spoliée de ses propres productions.

On peut espérer que cette forme d'auto-émancipation du monde de la recherche rendra prochainement possible un nouveau dialogue avec le milieu de l'édition scientifique, mais sur des bases toutes différentes et qui restent en grande partie à trouver[70].

2.2. Le principe de la paternité

Les progrès qui marquent actuellement la diffusion et la circulation des ressources s'accompagnent d'avancées tout aussi significatives dans le développement d'outils permettant de retracer les différentes phases de leur création. Les procédures numériques se complexifiant et se spécialisant, il devient nécessaire de définir des acteurs et des rôles, comme on le faisait déjà pour le livre manuscrit du Moyen Âge (« mains » des copistes, rubricateur, enlumineur, relieur, etc.). Le codage, le balisage et l'enrichissement linguistique du texte peuvent donner lieu à des manipulations multiples. Il n'est pas rare qu'une même opération ou que deux opérations proches soient assurées par deux personnes différentes (c'est souvent le cas pour l'annotation linguistique), ce qui accroît d'autant les acteurs du document numérique. Le codage de ces couches d'intervention successives obéit aux mêmes règles que les autres formes de codage, et la gestion multicouches ou multi-niveaux offerte par le numérique permet de rendre compte clairement des ajouts de chacun.

Le développement de nouvelles licences spécialisées dans la diffusion de contenus numériques (du type des licences Creative Commons) permet par ailleurs d'attacher au document électronique le détail de ses conditions d'utilisation spécifique. Parmi ces conditions figure en première ligne la clause de paternité. L'exploitation d'un objet numérique peut ainsi être étroitement associée à la reconnaissance explicite par celui qui

[70] C'est l'un des objectifs du mouvement de l'Open-Access cité plus haut.

l'utilise des noms et qualités de ceux qui sont à sa source. Des règles de citation précises complètent parfois le dispositif.

Outre qu'elles obligent à distinguer différents rôles, ces nouvelles pratiques amènent en même temps à s'interroger sur la portée et la valeur des différents types d'intervention. La production du livre papier, même si elle faisait intervenir elle aussi différents corps de métier, ne donnait pas lieu à un rappel aussi précis des tâches, ce qui, d'une certaine façon, simplifiait les choses. Seul le travail de l'auteur du texte était protégé, dans la mesure où il était le seul à comporter une part de création affirmée. Ce système ancien doit désormais être adapté au monde numérique, toutes les tâches sommairement décrites ci-dessus n'apportant pas la même valeur ajoutée et ne pouvant donner lieu à la même protection juridique. Dans la pratique, les exigences de citation se limitent jusqu'ici en général à l'auteur du texte-source ou de l'édition numérique. Mais on peut se demander dans quelle mesure d'autres tâches font aussi intervenir une part de création.

L'annotation linguistique, par exemple, suppose souvent une expertise qui pourrait être mieux reconnue. Mais, plus que l'annotation effective des textes, qui suit en général des principes déjà préétablis, c'est peut-être le choix et l'élaboration du système de catégories annotées qui pourrait être considéré comme une véritable production scientifique[71]. C'est assez rarement le cas en pratique. Il est, pour l'heure, encore difficile de s'entendre sur les limites du travail de création et sur la façon de le mesurer.

Les points abordés ici ne sont pas aussi déconnectés de la recherche scientifique proprement dite que certains collègues le pensent parfois. Il est vrai qu'ils posent des questions qui semblent parfois difficilement solubles, en l'état actuel de la législation. Et il arrive que le dialogue avec les professionnels spécialistes de ces questions (avocats spécialisés dans la propriété intellectuelle et services juridiques des organismes de recherche, souvent peu au fait des pratiques et des spécificités des Sciences humaines et sociales) soit ardu et rugueux. En outre, les procédures permettant de gérer concrètement et sur le long terme la protection des ressources numériques sont souvent lourdes à mettre en place. Mais il me semble, d'une part, qu'il est désormais impossible de ne pas répondre à ces questions, d'autre part, qu'elles

[71] De la même façon, on pourrait militer pour que la documentation technique, qui contient les informations sur les jeux de catégories et la façon de les appliquer, soit traitée comme une publication scientifique, au même titre qu'un article de recherche qui, éventuellement, s'appuie sur ces catégories.

donnent l'occasion de réfléchir, sous un angle particulier et souvent négligé, à nos méthodologies et pratiques de recherche.

3. L'organisation communautaire et les échanges scientifiques

Plusieurs initiatives nationales pour créer des espaces de dialogue, d'échange et de coopération entre linguistes ont récemment vu le jour, et c'est autour des corpus que ces espaces se sont créés, au sein des Très Grandes Infrastructures de Recherche et plus spécifiquement dans les Consortiums pour les corpus écrits et oraux. Même si ces superstructures peuvent paraître, par certains côtés, pompeuses et verbeuses, même si elles s'enchevêtrent parfois de manière opaque, elles reflètent pourtant une certaine volonté du monde de la recherche de s'organiser et de se mettre d'accord sur des socles communs.

Ce sur quoi j'aimerais insister surtout, c'est sur le fait que des questions qui peuvent sembler très concrètes, terre à terre ou triviales (la protection juridique, les normes de codage, l'annotation, les outils de la recherche, etc.) aident en réalité à structurer les communautés centrées sur une discipline. On observe d'ailleurs que les IR corpus se sont tous organisés par disciplines, alors que beaucoup – la plupart ? – des questions traitées par chacun d'eux sont transversales. On peut espérer qu'ils permettront par la suite d'établir des ponts et des rapprochements entre communautés diverses.

Les deux activités principales de ces organisations collectives sont, d'une part, le recensement des ressources existantes, d'autre part, la normalisation des pratiques. Mais ce qui paraît faire leur force principale, c'est le fait que leurs activités émanent des chercheurs eux-mêmes. La standardisation des ressources n'est désormais plus vue comme la nécessité de suivre des règles imposées de l'extérieur ou d'en haut. Elle émerge plutôt de pratiques mises en commun et définies de manière collective (d'où une certaine « mode » des « guides des bonnes pratiques » rédigés par les praticiens eux-mêmes). Deux exemples concrets me serviront d'illustration.

3.1. Les normes d'édition des textes médiévaux

L'édition de textes antiques et médiévaux est une activité très ancienne, qui a donné lieu à de nombreux débats en France comme à l'étranger. Dans le domaine français, les normes d'édition se sont pour l'essentiel fixées entre la fin du 19ème siècle et les années 1930 autour

des figures de Paul Meyer, puis de Mario Roques et Joseph Bédier (notamment Roques 1926 et Bédier 1928)[72]. Divers manuels et ouvrages de synthèse (*Conseils pour l'édition des textes médiévaux* 2001-2002, Lepage 2001) s'adressent aux littéraires, linguistes et historiens qui sont amenés à publier des textes médiévaux et donnent la liste plus ou moins complète et consensuelle des grands principes à suivre.

L'édition de textes anciens repose sur un ensemble de connaissances (de la langue médiévale, des manuscrits et des techniques scribales, etc.) et sur une longue tradition établie à l'échelle nationale, les pratiques variant en fonction des pays et des langues des textes édités (Duval 2006). Des règles relativement bien admises fixent, par exemple, la façon d'utiliser les diacritiques (trémas, accents, cédille), l'usage des majuscules, de la ponctuation, etc. dans les éditions modernes de textes anciens. Toutes ces règles ont bien évidemment été conçues pour des éditions papier et on verra plus loin que certains choix ont été motivés par des considérations très pratiques directement liées à ce support.

Les débats les plus vifs ont porté, au tournant du 20ème siècle, sur le statut à accorder à chaque témoin dans une tradition manuscrite qui peut en réunir un grand nombre et sur le degré de liberté accordé à l'éditeur dans le panachage de plusieurs sources manuscrites. Aux tenants d'une approche interventionniste, qui voulaient reconstruire le manuscrit original de tel ou tel auteur médiéval à partir d'une compilation de plusieurs sources, s'opposaient les adeptes de la fidélité à l'authenticité des données attestées. L'intervention décisive de J. Bédier a plutôt donné raison aux seconds et un compromis a été trouvé : après avoir étudié la tradition manuscrite du texte et choisi son manuscrit de base, l'éditeur le suit pour l'essentiel et rejette en note les leçons des autres manuscrits retenus. Il a pour principe de corriger le moins possible – ou de manière tempérée – le manuscrit principal, ce qui laisse tout de même la porte ouverte à toutes sortes d'appréciations subjectives et permet des divergences importantes dans la façon dont l'éditeur réussit à « s'auto-tempérer ».

En réalité, malgré un accord apparent sur l'essentiel des règles éditoriales, les philologues ont toujours pris une certaine liberté avec ces principes. De même qu'ils s'autorisent à rejeter les leçons du manuscrit de base, de même ils adaptent leurs usages des trémas et accents, de la ponctuation, etc. au texte qu'ils éditent[73], mais aussi peut-être à leurs propres habitudes ou à

[72] Les normes définies par M. Roques pour ses collaborateurs des Classiques français du Moyen Âge ont été adoptées en 1925 par un congrès de romanistes. Elles ont été publiées dans le numéro 52 de la Romania en 1926 (243-249). Elles font référence dans le domaine français.

[73] L'adaptation au texte – avec l'adaptation au public visé – fait partie des grands principes mis en avant dans le manuel de référence publié par l'Ecole des chartes en 2001 :

> « *L'éditeur doit en effet toujours prendre en compte la nature et la tradition du texte, mais aussi la finalité de l'édition et les intérêts de ses lecteurs : il ne traitera pas de la même façon un original du haut Moyen*

leurs préférences personnelles (Bédier parle du « goût » de l'éditeur). L'étude menée par F. Duval (2006) sur un grand nombre d'éditions de textes français montre une hétérogénéité certaine, qu'on ne peut d'ailleurs pas toujours mesurer à la lecture des introductions, souvent trop laconiques sur le détail des choix éditoriaux.

Il est probable que cette hétérogénéité dérive aussi des normes existantes, qui restent muettes sur certains points. On observe, par exemple, que les principes définis par Mario Roques en 1925 sont assez détaillés sur l'usage de l'accent et des trémas, mais ne disent rien du découpage graphique en mots ; ce même découpage est traité en une page et demie dans le manuel de l'Ecole des chartes (fascicule 1 : 41-42). Par ailleurs, certaines décisions ont été explicitement présentées comme des choix uniquement pratiques : la place du tréma sur les suites de voyelles en hiatus est réglée par le fait que les imprimeurs ne disposent pas des lettres ö et ä dans tous les caractères d'imprimerie. Les trémas seront donc toujours placés sur les voyelles e, u et i (Roques 1926 : 245).

Il ne s'agit pas de revenir ici sur des faits qui sont bien connus mais de se demander dans quelle mesure le développement actuel des nouvelles technologies rouvre des débats qu'on pouvait penser dépassés. Il me semble que les possibilités techniques actuelles permettent de reprendre les questions anciennes sur de nouvelles bases, et cela pour quatre raisons au moins :

1) Les procédures actuelles de création du document numérique obligent à définir les choix éditoriaux de manière très précise et très explicite à chaque phase du codage (en caractères, en unités lexicales, en unités de plus grande taille). Les normes et pratiques du numérique révèlent ainsi d'autant mieux l'hétérogénéité, le caractère peu explicite et incomplet des principes anciens.

2) Le numérique permet de faire des choix multiples ou, plus exactement, de coder simultanément des choix auparavant considérés comme antagonistes. Il est désormais possible de suivre différentes logiques en même temps, de réaliser une édition à la fois normalisée et diplomatique par exemple. On peut éventuellement aussi prédéfinir différents niveaux de représentation qui regroupent des choix concordants. Dans le cas de notre édition de la *Queste del saint Graal*, trois niveaux distincts ont été définis par Alexei Lavrentiev (normalisé, diplomatique et fac-similaire), chaque niveau se caractérisant par un faisceau de choix ou de traits propre. Ce travail de regroupement, qui permet de programmer différentes vues sur un

---

Âge ou le témoin rare d'un état ancien de langue, et un acte routinier de la fin du Moyen Âge » (*Conseils pour l'édition des textes médiévaux*, fasc. 1 : 18)

De fait, on peut se demander s'il est possible ou souhaitable d'employer des principes identiques pour toutes les catégories de textes. Il s'agit là d'un point très important pour l'avenir des normes d'édition.

même texte, a l'avantage d'orienter l'utilisateur vers un type d'usage prédéterminé. Il permet par la même occasion au chercheur de réfléchir aux usages de l'édition électronique, aux méthodologies de recherche et de lecture des textes.

3) Comme on vient de le voir, l'édition électronique peut conjuguer différents types d'usages et s'adapter à tous types de publics, ce qui n'était pas le cas des éditions papier antérieures.

4) Les nouvelles technologies du numérique permettent peut-être d'aborder certains des grands problèmes de l'édition scientifique de façon différente, d'une part, parce qu'elles rendent possible une séparation plus nette entre différents niveaux de représentation et d'analyse (distinction des niveaux graphique et linguistique par exemple), d'autre part, parce qu'elles offrent des solutions techniques pour transposer avec une grande précision les caractéristiques physiques du manuscrit médiéval. Elles favorisent ainsi ce qu'on pourrait appeler une approche bédiériste « dure » et permettent de progresser dans la fidélité au témoin de base (ce que renforce la possibilité d'associer l'image du manuscrit à l'édition). De même, ces technologies mettent à notre disposition de nouveaux moyens de comparaison des manuscrits, qui pourraient probablement renouveler les méthodes d'analyse de la tradition manuscrite d'une œuvre.

Mais ce qui me paraît importer plus que tout, c'est l'occasion que nous donne l'essor actuel du numérique de régler ces questions de manière collective. Il est en effet assez probable – en tout cas très souhaitable – que la réflexion sur les normes éditoriales de la philologie numérique dépasse le cadre limité des linguistes d'un côté, des philologues et littéraires de l'autre, ou encore des historiens, pour permettre les échanges et des choix sinon totalement communs, du moins suffisamment cohérents les uns avec les autres. Le numérique permettra peut-être ainsi de réconcilier ou en tout cas de rapprocher diverses communautés. C'est une chose qui me tient particulièrement à cœur.

3.2. La création d'un corpus bilingue latin / français

Je voudrais aborder en quelques mots à présent une collaboration en cours avec plusieurs collègues latinistes autour d'un projet de publication collective sur le passage du latin tardif au français. On sait que cette période est relativement mal connue et encore très insuffisamment décrite. La complexité et l'opacité des sources documentaires qui nous sont parvenues et dont on peine à savoir quel rapport exact elles peuvent entretenir avec la langue

parlée et comprise par tous explique en partie cette relative désaffection pour une période pourtant si riche pour les langues romanes en général et le français en particulier.

L'originalité de notre projet tient à deux choses : d'une part, la volonté de rassembler des latinistes et des spécialistes du français – ce qui, aussi surprenant que cela puisse paraître, ne se produit que très rarement –, d'autre part, l'exploration commune d'un corpus bilingue français/latin permettant d'étudier les deux langues dans leurs relations réciproques durant une bonne partie du Moyen Âge (4ème-12ème siècle), et tout spécialement aux débuts du français. Les objectifs scientifiques d'un tel projet sont triples : étudier le passage d'un système linguistique à l'autre dans ses différentes dimensions, promouvoir et développer une recherche diachronique sur la longue durée (les découvertes faites sur l'un des deux états de langue permettant souvent d'éclairer l'autre[74]), mieux apprécier les effets de l'interaction constante du français et du latin tout au long de la période médiévale.

La création d'un corpus bilingue français/latin offre un terrain d'échange et de discussion pour les deux communautés rassemblées dans le projet. Une partie des textes latins nous ayant été fournie par le Lasla (Université de Liège, http://www.cipl.ulg.ac.be/Lasla/), nous avons dû adapter et harmoniser nos méthodes et habitudes de travail pour construire un objet d'étude commun. Ce travail de comparaison et d'harmonisation relative concerne tous les plans abordés jusqu'ici : comparaison et gestion des spécificités des normes éditoriales propres à chaque langue (par tradition, les éditions latines sont bien plus interventionnistes et reconstructivistes que les éditions françaises, les éditions du Lasla ne comportent pas de signe de ponctuation, ne désambiguïsent pas *i*/*j* et *u*/*v*, etc.), harmonisation des métadonnées utilisées pour décrire les unités textuelles (les textes latins offrent, par exemple, des difficultés de datation encore supérieures à celles des textes français), diversité des jeux d'étiquettes adaptés à chaque langue (jeu du Lasla pour le latin, jeu Cattex2009 pour le français[75]), etc. Ce que nous montre un tel projet, c'est que la création de ressources communes peut être un prétexte ou une occasion de mettre à plat nos méthodes de recherche comme nos catégories d'analyse.

---

[74] Plusieurs recherches en cours ont déjà montré la fécondité de l'étude de ces filiations et rapprochements (voir les recherches d'A. Carlier sur *mult*, notamment Carlier 2012, et, sur les démonstratifs, les travaux de C. Marchello-Nizia, M. Fruyt et [doc. 15] Guillot & Carlier à par. c).

[75] Ces deux jeux se distinguent par le nombre, le type et la hiérarchisation interne des étiquettes. Le jeu Cattex2009 comporte 10 grandes catégories (NOM, VERBE, ADJECTIF, etc.), subdivisées en types (NOM commun, NOM propre, VERBE conjugué, etc.). Les traits morphologiques s'ajoutent à la suite (genre, nombre, personne, etc.). Le jeu du Lasla comporte 21 catégories que l'on combine aux informations morphologiques (cas, type de déclinaison, temps, personne, etc.). Les catégories sont plus nombreuses dan le jeu latin mais non hiérarchisées, et elles recouvrent parfois deux étiquettes distinctes dans le jeu français (pronom/adjectif possessif, pronom/adjectif indéfini, pronom/adjectif démonstratif etc.). La typologie des deux langues explique en partie ces disparités.

Il me semble que les différents points abordés dans ce chapitre, même s'ils correspondent pour la plupart à des activités apparemment très concrètes et pratiques, sont en réalité partie intégrante du développement de la recherche scientifique en Sciences du langage. Ils nourrissent à leur façon la réflexion méthodologique sur la nature des ressources linguistiques utilisées par le linguiste, la façon dont il les analyse, la portée et la valeur des résultats qu'il en obtient.

Ces réflexions me paraissent d'autant plus utiles et nécessaires qu'elles conditionnent également en partie la constitution d'un savoir cumulatif, en permettant de partager et de reproduire des méthodes et des catégories d'analyse communes. L'avenir des sciences du langage se trouve peut-être dans ces collaborations, dans la réflexion méthodologique et l'évolution des pratiques, dans la mise en place d'infrastructures de recherche qui permettent les échanges à grande échelle, la fixation de normes et de standards, et, finalement, la capitalisation, la diffusion et l'incrémentation des savoirs linguistiques.

Conclusion

Ce mémoire de synthèse m'a permis de revenir sur les grandes lignes de mon parcours personnel, sur ses partis pris théoriques et méthodologiques, et sur l'évolution progressive de mes pratiques de recherche. Il donnera ainsi une image que j'espère fidèle de ce qui est devenu au fil du temps le quotidien d'un nombre croissant de chercheurs en Sciences du langage (immersion dans les corpus, pratique des outils numériques, de l'annotation, implication croissante dans des réseaux et infrastructures de recherche collectives, etc.).

Les mutations survenues depuis quelques décennies dans notre discipline ont peu à peu renouvelé la définition de notre objet d'étude. Le corpus, les ressources linguistiques qui sont étudiées, la façon dont elles sont constituées et analysées, toutes ces questions ont peu à peu acquis une importance primordiale dans le processus de la recherche. Si des tendances contradictoires continuent de s'opposer, il me semble que ces réflexions ne peuvent que favoriser le développement de pratiques unifiées et offrir une issue à la dichotomie dépassée entre linguiste de bureau et linguiste de terrain.

J'espère avoir montré dans cette synthèse les principaux apports de la linguistique diachronique dans ces débats méthodologiques. L'altérité des ressources médiévales explique peut-être chez le linguiste médiéviste une certaine prédisposition à une attitude réflexive sur son objet d'étude. Elle l'oblige à se demander sans cesse quelle est cette langue qu'il étudie. Elle le rend peut-être d'autant mieux conscient aussi des limites de son travail.

J'ai insisté à plusieurs reprises sur le rôle de la concertation et de la collaboration avec les informaticiens dans les évolutions en cours. On peut regretter, d'un autre côté, que les relations entre les Sciences du langage et le monde du Traitement automatique des langues soient encore timides. Il reste certainement beaucoup de chemin à parcourir pour qu'émerge une méthodologie linguistique véritablement expérimentale et l'on peut se demander dans quelle mesure l'ouverture à d'autres disciplines conditionne une telle évolution.

Il me semble que l'optique très générale que j'ai essayé de défendre dans cette synthèse pourrait permettre aussi de poser en des termes un peu différents la question de la vitalité de notre domaine de recherche et de son attrait auprès des jeunes générations. On peut se demander, en effet, dans quelle mesure le développement de méthodes empiriques et, surtout, leur transmission au sein des formations académiques existantes offriraient des solutions à la crise que traversent actuellement la plupart des cursus en Sciences du langage en France. Mes initiatives personnelles dans cette direction (en association avec mes collègues de l'ENS et

d'ICAR) n'ont pourtant pas été très fructueuses et il me semble que les résistances de l'Université française sont encore assez fortes. Ces résistances tiennent sans doute en partie à ce que certains collègues considèrent les questions méthodologiques abordées dans ce mémoire comme de simples techniques subsidiaires. Mais de même qu'on enseigne conjointement les grands cadres théoriques et les notions et outils de la description linguistique, de même, je crois qu'une immersion précoce et systématique dans les méthodologies de constitution, de formatage et d'exploitation raisonnée des ressources linguistiques serait très utile à la compréhension des enjeux théoriques, des objets d'étude et des modes de constitution du savoir dans notre discipline. Une évolution en ce sens permettrait aussi une meilleure articulation des formations universitaires avec le monde professionnel (*via*, notamment, la transmission de savoirs et de techniques aux applications très concrètes). Mais il s'agit là de questions vastes et complexes, qui sont certainement décisives pour l'avenir de notre discipline et auxquelles ce mémoire de synthèse n'a pu apporter que des éléments de réponse très partiels.

Bibliographie

Adam, J.-M. (1999). *Linguistique textuelle. Des genres de discours aux textes*. Paris : Nathan.

Auroux, S. (1998). *La raison, le langage et les normes*. Paris : PUF.

Banniard, M. (1980). *Le Haut Moyen Âge occidental*. Paris : PUF (coll. Que sais-je ?).

Bazin-Tacchella, S. (2005). « *Lequel* dans la traduction française de la *Chirurgia Magna* de Guy de Chauliac : un outil de cohésion. In : D. James-Raoul *et al.* (éd.), *Par les mots et les textes... : mélanges de langue, de littérature et d'histoire des sciences médiévales offerts à Claude Thomasset*. Paris : PUPS, p. 37-53.

Bazin-Tacchella, S. (2007). « L'articulation des séquences textuelles dans la traduction française de la *Chirurgia Magna* de Guy de Chauliac (XVe siècle) : l'importance de la topicalisation. In : A. Vanderheyden *et al.* (éd.), *Texte et discours en moyen français : actes du XIe colloque international sur le moyen français (Anvers, mai 2005)*. Turhout : Brepols, p. 61-72.

Bédier, J. (1928). « La tradition manuscrite du *Lai de l'ombre*. Réflexions sur l'art d'éditer les anciens textes ». *Romania* 54, p. 161-196 et 321-356.

Bériou, N. (2010). « Latin et langues vernaculaires dans les traces écrites de la parole vive des prédicateurs (XIIIe-XIVe siècles) ». In : S. Le Briz et G. Veysseyre (éd.), *Approches du bilinguisme latin-français au Moyen Âge. Linguistique, codicologie, esthétique*. Turnhout : Brepols, p. 191-206.

Biber, D. (1988). *Variation across Speech and Writting*. Cambridge : Cambridge University Press.

Biber, D. (1989). « A typology of English Texts ». *Linguistics* 27, p. 3-43.

Biber, D. (1995). *Dimensions of register variation. A cross-linguistic comparison*. Cambridge : Cambridge university Press.

Biber, D. (1998). *Corpus Linguistics. Investigating Language Structure and Use*. Cambridge : Cambridge University Press.

Biber, D. (2010). « What can a corpus tell us about registers and genres? » In : A. O'Keeffe & M. McCarthy (éd.), *The Routledge Handbook of Corpus Linguistics*. London / New York : Routledge, p. 241-254.

Blanche-Benveniste, C. (1997). *Approches de la langue parlée en français*. Paris : Ophrys.

Blanche-Benveniste, C. & Jeanjean, C. (1987). *Le français parlé. Edition et transcription*. Paris : Didier-Erudition.

Bossuat, R. (1954-). *Manuel bibliographique de la littérature française du Moyen âge*, Paris : Librairie d'Argences.

Botley, S. (2001). « Demonstratives in English. A Corpus-Based Study ». *Journal of English Linguistics* 29/1, p. 7-33.

Botley, S. (2006). « Indirect Anaphora. Testing the limits of corpus-based linguistics ». *International Journal of Corpus Linguistics* 11/1, p. 73-112.

Botley, S. & Mac Enery, A. (2000). *Corpus-based and Computational Approaches to Discourse Anaphora*. Amsterdam/Philadelphia : John Benjamins Publishing Company.

Bozzolo, C. & Ornato, E. (1980). *Pour une histoire du livre manuscrit au Moyen âge : trois essais de codicologie quantitative*. Paris : Éditions du Centre national de la recherche scientifique.

Buridant, C. (2000). *Grammaire nouvelle de l'ancien français*. Paris : Sedes.

Buridant, C. (2011). « Modèles et remodelages ». In : C. Galderisi (éd.), *Translations médiévales. Cinq siècles de traductions en français au Moyen Âge (XIe-XVe siècles).*

Etude et répertoire, vol. 1 De la translatio studii à l'étude de la translatio. Turnhout : Brepols, p. 93-126.
Buridant, C. (2012). « Contribution à l'histoire de la prose française médiévale : la traduction du latin dans les textes historiographiques français et espagnols. In : C. Guillot *et al.* (éd.), *Le changement en français. Etudes de linguistique diachronique*. Bern : Peter Lang, p. 15-35.
Buridant, C. (à par.). « Les premières traductions hagiographiques en français : premiers jalons d'une étude prospective », à paraître dans les actes du colloque international DIACHRO-VI (Louvain, 17-19 octobre 2012).
Carlier, A. (2004). « Sur les premiers stades de développement de l'article partitif ». *Scolia* 18, p. 117-147.
Carlier, A. (2012). « Le très ancien français comme objet d'analyse : valeur heuristique et aspects méthodologiques ». In : C. Guillot *et al.* (éd.), *Le changement en français. Etudes de linguistique diachronique*. Bern : Peter Lang, p. 57-86.
Carlier, A. & De Mulder, W. (2010). « The emergence of the definite article : *ille* in competition with *ipse* in Late Latin ». In : K. Davidse *et al.* (éd.), *Subjectification, intersubjectification and grammaticalization*. Berlin/New York : De Gruyter, p. 241-275.
Chevalier, J.-C. & Encrevé, P. (2006). *Combats pour la linguistique, de Martinet à Kristeva. Essai de dramaturgie épistémologique*. Lyon : ENS Editions.
Colombo-Timelli (1996). *Traductions françaises de l'Ars minor de Donat au moyen âge (XIIIe-XVe siècles)* Firenze : La Nuova Italia.
Combettes, B. (2001). « Un cas de grammaticalisation en français. *En ce qui regarde* / *pour ce qui regarde*. In : C. Buridant et al. (éd.), *Par monts et par vaux : itinéraires linguistiques et grammaticaux : mélanges de linguistique générale et française offerts au professeur Martin Riegel pour son soixantième anniversaire par ses collègues et amis*. Louvain : Peeters, p. 111-126.
Combettes, B. (2007). « Evolution des structures thématiques en moyen français ». In : A. Vanderheyden, et al. (éd.), *Texte et discours en moyen français : actes du XIe colloque international sur le moyen français (Anvers, mai 2005)*. Turnhout : Brepols, p. 35-46.
Combettes, B. & Prévost, S. (2003). « Texte argumentatif et topicalisation d'une proposition : une approche diachronique ». *Scolia* 16, p. 63-75.
De Mulder, W. (1997). « Les démonstratifs : des indices de changement de contexte ». In : N. Flaux *et al.* (éd.), *Entre général et particulier : les déterminants.* Arras : Artois Presses Université, p. 137-200.
Diessel, H. (1999). *Demonstratives, form, function, and grammaticalization*. Amsterdam/Philadelphia : John Benjamins Publishing Company.
Duval, F. (2006). La philologie française, pragmatique avant tout ? In : F. Duval (éd.), *Pratiques philologiques en Europe*. Paris : Ecole des Chartes, p. 115-150.
Duval, F. (éd.) (2006). *Pratiques philologiques en Europe*. Paris : Ecole des Chartes.
Duval, F. (2010). « Le lexique de la civilisation romaine au Moyen Âge : de la diglossie à l'interlinguisme. In : S. Le Briz et G. Veysseyre (éd.), *Approches du bilinguisme latin-français au Moyen Âge. Linguistique, codicologie, esthétique*. Turnhout : Brepols, p. 63-79.
Foulet, L. (1967, 1ère éd. 1919). *Petite syntaxe de l'ancien français*. Paris : Champion.
Frank, B. (1997) (éd.). *Inventaire systématique des premiers documents des langues romanes***,** Tübingen : Gunter Nar.
Gadet, F. (2003). *La variation sociale en français*. Paris : Ophrys.
Génicot, L. (1972-) (éd.). *Typologie des Typologie des sources du Moyen âge occidental*, Turnhout : Brepols.

Groupe de recherches La civilisation de l'écrit au Moyen âge (éd.) (2001-2002). *Conseils pour l'édition des textes médiévaux*, 3 vol., Paris : Comité des travaux historiques et scientifiques : Ecole nationale des chartes.

[doc. 2] 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, p. 29-36.

[doc. 6] Guillot, C. (2009) « Ecrit médiéval et traces d'oralité : l'exemple de l'adverbe *or(e)* ». In : E. Havu *et al*. (éd.), *La langue en contexte. Actes du colloque Représentation du sens linguistique IV (Helsinki, 28-30 mai 2008)*, Helsinki : Société Néophilologique, p. 267-281.

[doc. 8] Guillot, C. (2010a) « Le démonstratif de notoriété de l'ancien français : approche textuelle », In : B. Combettes et al. (éd.). *Le changement en français. Etudes de linguistique diachronique*, Bern/Berlin/Bruxelles : Peter Lang, p. 217-233.

[doc. 9] Guillot, C. (2010b) « Les démonstratifs de l'ancien français : un système encore personnel ? ». Actes du 2e Congrès Mondial de Linguistique Française, EDP Sciences (www.linguistiquefrancaise.org), [http://dx.doi.org/10.1051/cmlf/2010085].

[doc. 11] Guillot, C. (2012a). « Système des démonstratifs médiévaux et exemples de stratégies communicatives », *Journal of French Language Studies*, disponible en ligne [CJO 2012 doi:10.1017/S0959269512000245].

[doc. 12] Guillot, C. (2012b) « Le pronom anaphorique *cil* de l'ancien français : continuité ou discontinuité topicale ? ». In : C. Denizot & E. Dupraz (éd.). *Anaphore et anaphoriques : variété des langues, variété des emplois*, Mont-Saint-Aignan : Publications des universités de Rouen et du Havre (Cahiers de l'ERIAC), p. 97-115.

[doc. 3] Guillot, C., Heiden, S. & Lavrentiev Alexei (2007). « Typologie des textes et des phénomènes linguistiques pour l'analyse du changement linguistique avec la Base de Français Médiéval », *Linx*, numéro spécial, p. 125-139.

[doc. 5] Guillot, C., Heiden, S., Lavrentiev, A. & Marchello-Nizia, C. (2008). « Constitution et exploitation des corpus d'ancien français et de moyen français », *Corpus* 7, p. 5-23.

[doc. 7] Guillot, C. & Lavrentiev, A. (dir.) (version 4.2. mars 2009). *Présentation des descripteurs du projet CORPTEF* ([doc. http://corptef.ens-lyon.fr/spip.php?rubrique60

[doc. 13] Guillot, C., Lavrentiev, A., Pincemin, B. & Heiden, S. (à par. a). « 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)*.

[doc. 14] Guillot, C., Heiden, S., Lavrentiev, A. & Pincemin, B. (à par. b). « L'oral représenté dans un corpus de français médiéval (9e-15e) : paramètres de variation diamésique (discours direct vs récit), générique et diachronique », Actes du colloque international *Les variations diasystémiques et leurs interdépendances* (19-21 novembre 2012, Copenhague)

[doc. 15] Guillot, C. & Carlier, A. (à par. c). « Evolution des démonstratifs du latin au français : le passage d'un système ternaire à un système binaire », Actes du colloque international *DIACHRO-VI Le français en diachronie* (17-19 octobre 2012, Louvain)

Guyotjeannin, O. (1998). *Les sources de l'histoire médiévale*. Paris : Librairie générale française (Le Livre de poche).

Habert, B. (2000). « Des corpus représentatifs : de quoi, pourquoi, comment ? » In : M. Bilger (éd.), *Linguistique sur corpus. Études et réflexions*. Perpignan : Presses Universitaires de Perpignan, p. 11-58.

Habert, B. (2004). « Outiller la linguistique : de l'emprunt aux rencontres de savoirs ». *Revue française de linguistique appliquée* IX/1, p. 5-24.

Habert, B., Nazarenko, A. & Salem, A. (1997). *Les linguistiques de corpus*. Paris : Armand Colin.

Habert, B. & Zweigenbaum, P. (2002). « Régler les règles ». *Traitement automatique des Langues* 43/3, p. 83-105.

Habert, B. & Fuchs, C. (2004). « Bilan et perspectives méthodologiques ». *Le français moderne* LXXII/1, p. 88-97.

Hasenhor, G. (1990). « Traductions et littérature en langue vulgaire (chapitre 8) ». In : J. Vezin & H.-J. Martin (éd.), *Mise en page et mise en texte du livre manuscrit*. Paris : Éditions du Cercle de la librairie/Promodis, p. 231-354.

Hasenhor, G. (2002). « Ecrire en latin, écrire en roman : réflexions sur la pratique des abréviations dans les manuscrits français des XIIe et XIIIe siècles ». In : M. Banniard (éd.), *Langages et peuples d'Europe : cristallisation des identités romanes et germaniques, VIIe-XIe siècle*. Toulouse : CNRS, Université de Toulouse-Le Mirail, p. 79-110.

[doc. 1] Heiden, S. & Guillot, C. (2003). « Capitalisation des savoirs par le Web : une application de la TEI pour l'encodage et l'exploitation des textes de la Base de français médiéval ». In : P. Kunstmann, F. Martineau, D. Forget (éd.), *Ancien et moyen français sur le Web. Enjeux méthodologiques et analyse de discours,* Ottawa : Les éditions David, p. 77-92.

Jauss, H. R. *et al*. (éd.) (1972-). *Grundriß der romanischen Literaturen des Mittelalters*, Heidelberg :C. Winter.

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

Kleiber, G. (1987). « L'opposition *cist*/*cil* en ancien français ou comment analyser les démonstratifs ? ». *Revue de linguistique romane* 51, p. 5-35.

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

Koch, P. (1997). « Diskurstraditionen : zu ihrem sprachtheoretischen Status und ihrer Dynamik ». In : B. Frank *et al.* (éd.), *Gattungen mittelalterlicher Schriftlichkeit*. Tübingen : Narr, p. 43-79.

Koch, P. (2008). « Le latin - une langue pas tout à fait comme les autres ? Le problème de la diglossie en Gaule septentrionale ». In : M. Van Acker, *et al.* (éd.), *Latin écrit - Roman oral? De la dichotomisation à la continuité*. Turnhout : Brepols, p. 43-67.

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 *et al*. (éd.), *Lexikon der romanistischen Linguistik*. Tübingen : Niemeyer, p. 584-627.

Kroch, A. (1989). « Reflexes of Grammar in Patterns of Language Change ». *Language Variation and Change* 1,p. 199-244.

Lavrentiev, A. (2009). *Tendances de la ponctuation dans les manuscrits et incunables français en prose, du XIIIe au XVe siècle*. Thèse de doctorat, Ecole normale supérieure de Lyon.

Le Briz, S & Veysseyre, G. (éd.) (2010). *Approches du bilinguisme latin-français au Moyen Âge. Linguistique, codicologie, esthétique*. Turnhout : Brepols.

Lee, D. Y. (2001). « Genres, registers, text types, domains, and styles : clarifying the concepts and navigating a path through the BNC jungle ». *Language Learning & Technology* 5, p. 37-72.

Leech, G. (1991). « The state of the art in corpus linguistics ». In : K. Aijmer & B. Altenberg (éd.), *English Corpus Linguistics*. London/New York : Longman, p. 8-29.

Leech, G. (1992). « Corpora and theories of linguistic performance ». In : J. Svartvik (éd.), *Directions in Corpus Linguistics. Proceedings of Nobel Symposium 82, Stockholm, 4-8 August 1991*. Berlin/New York : Mouton de Gruyter, p. 105-122.

Leech, G. (1997). « What is a corpus and What is Corpus Annotation ? » In : R. Garside *et al.* (éd.), *Corpus Annotation : Linguistics Information from Computer Text Corpora*. London / New York : Longman.

Lepage, Y. (2001). *Guide de l'édition de textes en ancien français*. Paris : Champion.

Llamas-Pombo, E. (2010). « Marques graphiques du discours rapporté. Manuscrits du *Roman de la Rose*, XVe siècle ». In : B. Combettes *et al.* (éd.), *Le changement en français. Etudes de linguistique diachronique*. Bern/Berlin/Bruxelles : Peter Lang, p. 249-269.

Lusignan, S. (1987, 1ère éd. 1986). *Parler vulgairement. Les intellectuels et la langue française aux XIIIe et XIVe siècles*. Paris / Montréal : Vrin / Les Presses de l'Université de Montréal.

Lusignan, S. (2012). *Essai d'histoire sociolinguistique - Le français picard au Moyen Âge*. Paris : Garnier.

Mac Enery, T. & Hardie, A. (2012). *Corpus Linguistics. Method, Theory and Practice*. Cambridge : Cambridge University Press.

Mac Enery, T. & Wilson, A. (2001, 1ère édition 1996). *Corpus Linguistics.* Edinburgh : Edinburgh University Press.

Marchello-Nizia, C. (1995). *L'évolution du français. Ordre des mots, démonstratifs, accent tonique*. Paris : Armand Colin.

Marchello-Nizia, C. (1997). « Evolution de la langue et représentations sémantiques : du 'subjectif' à l''objectif' en français ». In : C. Fuchs & S. Robert (éd.), *Diversité des langues et représentations cognitives*, p. 119-135.

Marchello-Nizia, C. (2012). « L'oral représenté : un accès construit à une face cachée des langues 'mortes' ». In: C. Guillot *et al.* (éd.), *Le changement en français. Etudes de linguistique diachronique*. Bern/Berlin/Bruxelles : Peter Lang, p. 247-264.

Marchello-Nizia, C. (à par. a) « Les débuts de l'"oral représenté" en français : marquage du discours direct dans les plus anciens textes », à par. dans les *Mélanges* Soutet.

Marchello-Nizia, C. (à par. b) « L'importance spécifique de l'"oral représenté" pour la linguistique diachronique », à par. dans les Actes du colloque de la Société internationale de diachronie du français (6-8 septembre 2011, Nancy).

Mazziotta, N. (2009). *Ponctuation et syntaxe dans la langue française médiévale. Étude d'un corpus de chartes originales écrites à Liège entre 1236 et 1291*. Tübingen : Max Niemeyer Verlag.

Ménard, P. (1994, 1ère éd. 1973). *Syntaxe de l'ancien français*. Bordeaux : Editions Bière.

Milroy, J. (2003). « On the role of the speaker in language change ». In : R. Hickey (éd.), *Motives for Language Change*. Cambridge : Cambridge University Press, p. 143-157.

Mindt, D. (1991). « Syntactic evidence for Semantic Distinctions in English ». In : K. Aijmer & B. Altenberg (éd.), *English Corpus Linguistics*. London/New York : Longman, p. 183-196.

Moignet, G. (1984, 1ère éd. 1973). *Grammaire de l'ancien français*. Paris : Klincksieck.

Monfrin, J. (1964). « Les traducteurs et leur public en France au Moyen âge ». *Journal des savants* 1, p. 5-20.

Perret, M. (1988). *Le signe et la mention : adverbes embrayeurs ci, ça, la, iluec en moyen français (XIVe-XVe siècles)* Genève : Droz.

[doc. 4] Pincemin, B., Guillot, C., Heiden, S. & Lavrentiev Alexei (2008). « Usages linguistiques de la textométrie. Analyse qualitative de la consultation de la Base de Français Médiéval via le logiciel Weblex », *Syntaxe & sémantique* 9, p. 87-110.

Prévost, S. (2011). *Français médiéval en diachronie : du corpus à la langue*, mémoire de synthèse de l'Habilitation à diriger les recherches.

[doc. 10] Rainsford, T., Guillot, C., Lavrentiev, A. & Prévost, S. (2012). « La zone préverbale en ancien français : apport des corpus annotés ». Actes du 3e Congrès Mondial de Linguistique Française, EDP Sciences (www.linguistiquefrancaise.org), 159-176, [http://dx.doi.org/10.1051/shsconf/20120100246].

Roques, M. (1926). « Etablissement de règles pratiques pour l'édition des anciens textes français et provençaux ». *Romania* 52, p. 243-249.

Ruby, C. (2010). « Les psautiers bilingues latin / français dans l'Angleterre du XIIe siècle. Affirmation d'une langue et d'une écriture ». In : S. Le Briz & G. Veysseyre (éd.), *Approches du bilinguisme latin-français au Moyen Âge. Linguistique, codicologie, esthétique.* Turnhout : Brepols, p. 197-190.

Schøsler, L. (2012). « Sur l'emploi du passé composé et du passé simple. "... ayant receu de voz nouvelles, ie communicquay avec luy, et la conclusion fust telle que vous ay mande..." ». In : C. Guillot *et al.* (éd.), *Le changement en français. Etudes de linguistique diachronique*. Bern : Peter Lang, p. 321-339.

Sinclair, J. (1991). *Corpus Concordance Collocation*. Oxford : Oxford University Press.

Sinclair, J. (2005). « Corpus and Text. Basic Principles ». In : M. Wynne (éd.), *Developing Linguistic Corpora. A Guide to Good Pratice*. Eynsham/Oxford : *The Information Press*, p. 1-16.

Tognini-Bonelli, E. (2001). *Corpus linguistics at work*. Amsterdam/Philadelphia : John Benjamins Publishing Company.

Traugott, E. Closs (2010). « (Inter)subjectivity and (inter)subjectification : A reassessment ». In : K. Davidse *et al.* (éd.), *Subjectification, intersubjectification and grammaticalization*. Berlin/New York : De Gruyter, p. 29-71.

Wunderli, P. (1977). « Strukturen des Possessivums im Altfranzösischen ». *Vox Romanica* 36, p. 38-66.

Zink, G. (1989). *Morphologie du français médiéval*. Paris : PUF.

Curriculum vitae

Céline GUILLOT (BARBANCE)

Adresse personnelle 13, rue Maxime Teyssier
69 120 Vaulx-en-Velin
Tel : 04 72 37 47 19
Née le 06-08-1969 à Béziers (Hérault)
Nationalité française

FONCTIONS EXERCÉES

Ayant accompli une double formation d'archiviste-paléographe et de doctorante en Sciences du Langage, j'ai d'abord exercé des fonctions au sein de la Direction des Archives de France avant de me réorienter vers l'enseignement et la recherche.

- Situation actuelle
Maître de conférences à l'ENS de Lyon (7ème section), membre de l'Institut universitaire de France (junior)

- Septembre 2005 – janvier 2006
Détachement au CNRS sur un poste de CR2 au sein de l'UMR 5191 ICAR

- Septembre 2004-août 2005
Congé parental

- 1999-août 2004
ATER à l'ENS de Lyon, UMR 5191 ICAR (détachement de mon corps d'origine)

- 1998-1999
Conseillère pour les archives à la Direction Régionale des Affaires Culturelles de Rhône-Alpes

- 1994-1998
Conservateur adjoint aux Archives Départementales de la Loire

ACTIVITÉS PÉDAGOGIQUES

1. Préparation à l'agrégation
Septembre 2005 -> Grammaire de l'ancien français (préparation à la totalité de l'épreuve de langue médiévale au concours de l'agrégation de Lettres modernes) (40h CM, d'abord comme chargée de cours en poste au CNRS puis comme MCF)
2007-2008 : Préparation à l'épreuve orale de l'agrégation de Lettres classiques (10h TD)

2. Master 1 et 2
2007 -> Phonétique historique du français (21h CM Master 1 et 2)
2007 -> 2010 Introduction à l'ancien français (42h CM Master 1 et 2)
2007-2009 : Introduction à la linguistique (21h TD Master1 et 2)
2008-2010 : Le livre et ses auteurs dans le Master Edition numérique des savoirs (4h TD Master2)
2010-2012 -> Diachronie du français et grammaticalisation (21h CM, Master 2)

3. Licence 3
2008-2010 : Master Langue et littérature françaises (5h TD)

4. Cycle de conférences
Organisation en collaboration avec mes collègues de l'ENS de Lyon des « Conf'apéros en Sciences du langage de l'ENS » (7/8 conférences chaque année ; http://cle.ens-lsh.fr/31350188/0/fiche___pagelibre).

5. Divers
Formation d'enseignants, de chercheurs, d'élèves de l'ENS et de doctorants (linguistes, littéraires, historiens, philosophes) à la constitution, à l'interrogation et à l'exploitation des corpus de textes français médiévaux, à l'encodage XML/TEI de ces textes et à l'exploitation de textes étiquetés en morphosyntaxe

Direction de mémoires et soutenances
- Participation au jury de soutenance de M2 de Vanessa Obry (3 juillet 2006)
- Participation informelle à l'encadrement de la thèse de Vanessa Obry (ENS de Lyon / Paris3)
- Codirection du mémoire de M2 de Sciences du langage de Marianne Amaré (*La chosification. De sa manifestation sa manifestation à sa destitution)* en 2008-2009
- Codirection du mémoire de M2 de Sciences du langage de Jian Wang (*Etude comparative des emprunts à l'anglais en français et en chinois*) en 2012-2013

ANIMATION SCIENTIFIQUE

Participation à des projets de recherche

- Responsabilité scientifique et administrative de la Base de Français Médiéval depuis 2006 (http://bfm.ens-lsh.fr/)
- Coordination du projet ANR *Corpus représentatif des premiers textes français* (2008-2010 ; http://corptef.ens-lyon.fr/)
- Coordination de la partie lyonnaise du projet ANR/DFG franco-allemand *Syntactic Reference Corpus of Medieval French* (2009-2011)
- Coordination du projet régional (Cluster13) *Edition numérique interactive de la Queste del saint Graal, roman en prose du XIII^{e} siècle* (2009-2010)
- Coordination du projet financé par le fonds recherche de l'ENS de Lyon *Edition numérique et diachronique de textes médiévaux* (2013-2014), direction C. Guillot.
- Coordination du projet financé par l'ILF *Evolution des démonstratifs en français* (2006-2007)
- Participation au projet ANR *Evolution linguistique et corpus* (2007-2009)
- Participation au projet ANR/DFG franco-allemand *L'évolution du système prépositionnel du français* (2013-2015)
- Participation au projet PEPS CNRS Modélisation Contrastive et Computationnelle des Chaînes de Coréférence (2011-2012)

Organisation de colloques et de journée d'études

- Journée d'études *Le Démonstratif en français* (ENS de Lyon, 30 mars 2005)
- Journée d'études *Décrire le très ancien français. Approche comparative (latin/français) et outillée* (ENS de Lyon, 30 octobre 2009)
- Colloque international DIACHRO-V *Le français en diachronie* (ENS de Lyon, 20-22 octobre 2010 ; http://diachro-v.ens-lyon.fr/)
- Journée d'étude du *Consortium pour les corpus de français médiéval* intitulée *Philologie numérique : production, usages et évaluation* (ENS de Lyon, 21 juin 2011 ; http://ccfm.ens-lyon.fr/spip.php?article58)
- Journée d'études *Latin tardif > français ancien : continuités et ruptures* (ENS de Lyon, 29 mars 2012)
- Journées d'études *Décrire le passage du latin au français à travers l'analyse d'un corpus bilingue* (ENS de Lyon, 25-26 mars 2013)

Participation à des comités scientifiques et évaluation d'articles

- Participation au comité scientifique du colloque international DIACHRO, « Evolutions en français », en 2006 (Paris), 2008 (Madrid), 2010 (Lyon) et 2012 (Louvain) ; participation à l'édition des actes du colloque en 2008 et 2010
- Evaluation d'articles notamment pour le *Congrès mondial de linguistique française*, section diachronie (2008, 2010 et 2012), pour la revue *Journal of French Language Studies*, pour les actes du colloque *New Reflections on Grammaticalization 4* et pour la *Research Foundation* (Belgique, Flandre)

PUBLICATIONS

1. Editions de texte

Barbance-Guillot, C. (1992). *Edition critique du Des cas des nobles hommes et femmes de Laurent de Premierfait (1409) et commentaire linguistique*, thèse de l'Ecole nationale des Chartes, exemplaire dactylographié.

Participation à l'édition en ligne de la *Queste del Saint Graal*, en collaboration avec A. Lavrentiev et C. Marchello-Nizia (http://portal.textometrie.org/txm/).

2. Direction d'ouvrages collectifs ou de numéros spéciaux de revues

Guillot, C. (dir.) (2006). *Le démonstratif en français, Langue française* 152, 128 p.

Guillot, C., Heiden, S. & Prévost, S. (dir.) (2006). *A la quête du sens : études littéraires, historiques et linguistiques en hommage à Christiane Marchello-Nizia*, Lyon : ENS Editions, 364 p.

Guillot, C., Heiden, S., Lavrentiev, A. & Marchello-Nizia Christiane (dir.) (2008). *Constitution et exploitation des corpus d'ancien français et de moyen français, Corpus* 7, 252 p.

Combettes, B., Guillot, C., Oppermann-Marsaux, E., Prévost, S. & Rodriguez-Somolinos, A. (dir.) (2010). *Le changement en français. Etudes de linguistique diachronique*. Bern/Berlin/Bruxelles : Peter Lang (Sciences pour la communication), 402 p.

Guillot, C., Combettes, B., Lavrentiev, A., Oppermann-Marsaux, E. & Prévost, S. (dir.) (2012). *Le changement en français. Etudes de linguistique diachronique*. Bern/Berlin/Bruxelles : Peter Lang (Sciences pour la communication), 409 p.

3. Articles
3.1. Articles rédigés en tant que seul auteur

[I-1995] Barbance-Guillot, C. (1995). « La Ponctuation médiévale : quelques remarques sur cinq manuscrits du début du XV^{ème} siècle », *Romania* 113, p. 505-527.

[II-2003] Guillot, C. (2003). « Grammaticalisation et système de la référence : *celui, icelui, cest, cestui* et *ledict* dans un texte du début du XV^{ème} siècle », *Verbum* XXV/3, p. 369-379.

[doc. 2] 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, p. 29-36.

[III-2006] Guillot, C. (2006). « Démonstratif et déixis discursive : analyse comparée d'un corpus écrit de français médiéval et d'un corpus oral de français contemporain », *Langue française* 152, p. 56-69.

[IV-2006] Guillot, C. (2006). « Introduction », *Langue française* 152, p. 3-8.

[doc. 11] Guillot, C. (2012). « Système des démonstratifs médiévaux et exemples de stratégies communicatives », *Journal of French Language Studies*, disponible en ligne [CJO 2012 doi:10.1017/S0959269512000245].

3.2. Articles rédigés en collaboration

[doc. 3] Guillot, C., Heiden, S. & Lavrentiev Alexei (2007). « Typologie des textes et des phénomènes linguistiques pour l'analyse du changement linguistique avec la Base de Français Médiéval », *Linx*, numéro spécial, p. 125-139.

[V-2007] Guillot, C., Lavrentiev, A. & Marchello-Nizia, C. (2007). « Les corpus de français médiéval : état des lieux et perspectives », *Revue française de linguistique appliquée* 121, p. 125-128.

[doc. 4] Pincemin, B., Guillot, C., Heiden, S. & Lavrentiev Alexei (2008). « Usages linguistiques de la textométrie. Analyse qualitative de la consultation de la Base de Français Médiéval via le logiciel Weblex », *Syntaxe & sémantique* 9, p. 87-110.

[doc. 5] Guillot, C., Heiden, S., Lavrentiev, A. & Marchello-Nizia, C. (2008). « Constitution et exploitation des corpus d'ancien français et de moyen français », *Corpus* 7, p. 5-23.

4. Chapitres d'ouvrages

4.1 Chapitres rédigés en tant que seul auteur

[VI-2006] Guillot, C. (2006) « Anaphores résomptives démonstratives et relations partie/tout en discours ». In : G. Kleiber, C. Schnedecker, A. Theissen (éd.), *La relation partie-tout*, Louvain-Paris : Peeters (Bibliothèque de l'Information Grammaticale), p. 89-302.

[VII-2007] Guillot, C. (2007) « Entre anaphore et deixis : l'anaphore démonstrative à fonction résomptive ». In : D. Trotter (éd.), *Actes du XXIVe Congrès international de linguistique et de philologie romanes, Aberystwyth, (1-6 août) 2004*, Tübingen : Niemeyer, vol. 3, p. 307-315.

[doc. 6] Guillot, C. (2009) « Ecrit médiéval et traces d'oralité : l'exemple de l'adverbe *or(e)* ». In : E. Havu *et al*. (éd.), *La langue en contexte. Actes du colloque Représentation du sens linguistique IV (Helsinki, 28-30 mai 2008)*, Helsinki : Société Néophilologique, p. 267-281.

[VIII-2009] Guillot, C. (2009) « Le rôle désambiguïsant du déterminant *ledit* en moyen français à l'épreuve d'un corpus de traduction ». In : D. Lagorgette et O. Bertrand (éd.), *Etudes de corpus en diachronie et en synchronique. De la traduction à la variation*, Chambéry : Presses de l'Université de Savoie, p. 29-43.

[doc. 8] Guillot, C. (2010) « Le démonstratif de notoriété de l'ancien français : approche textuelle », In : B. Combettes et al. (éd.). *Le changement en français. Etudes de linguistique diachronique*, Bern/Berlin/Bruxelles : Peter Lang, p. 217-233.

[doc. 9] Guillot, C. (2010) « Les démonstratifs de l'ancien français : un système encore personnel ? ». Actes du 2e Congrès Mondial de Linguistique Française, EDP Sciences (www.linguistiquefrancaise.org), [http://dx.doi.org/10.1051/cmlf/2010085].

[doc. 12] Guillot, C. (2012) « Le pronom anaphorique *cil* de l'ancien français : continuité ou discontinuité topicale ? ». In : C. Denizot & E. Dupraz (éd.). *Anaphore et anaphoriques : variété des langues, variété des emplois*, Mont-Saint-Aignan : Publications des universités de Rouen et du Havre (Cahiers de l'ERIAC), p. 97-115.

4.2. Chapitres écrits en collaboration

[doc. 1] Heiden, S. & Guillot, C. (2003). « Capitalisation des savoirs par le Web : une application de la TEI pour l'encodage et l'exploitation des textes de la Base de français médiéval ». In : P. Kunstmann, F. Martineau, D. Forget (éd.), *Ancien et moyen français sur le Web. Enjeux méthodologiques et analyse de discours*, Ottawa : Les éditions David, p. 77-92.

[IX-2007] Guillot, C., Lavrentiev, A. & Marchello-Nizia, C. (2007). « La Base de Français Médiéval (BFM) : états et perspectives ». In : P. Kunstmann et A. Stein (éd.), *Le Nouveau Corpus d'Amsterdam, Actes de l'atelier de Lauterbad (23-26 février 2006)*, Stuttgart : Franz Steiner Verlag, p. 143-152.

[X-2008] Guillot, C. & Mortelmans, J. (2008). « Clarté ou vérité, *ledit* dans la prose de la fin du Moyen-âge ». In : O. Bertrand *et al*. (éd.). *Discours, diachronie, stylistique du français. Etudes en hommage à Bernard Combettes*, Bern/Berlin/Bruxelles : Peter Lang, p. 307-323.

[XI] De Mulder, W., Guillot, C. & Mortelmans, J. (2010). « *Ce N-ci* et *ce N-là* en moyen français », in : L. M. Tovena (éd.), *Déterminants en diachronie et synchronie*, Paris : Projet ELICO Publications (http://elico.linguist.univ-paris-diderot.fr/book_elico.php).

[XII] De Mulder, W., Guillot, C. & Mortelmans, J. (2011). « Ce N-ci and ce N-là in Middle French », in : L. M. Tovena (ed.), *French Determiners In and Across Time*, London: College Publications, p. 29-54.

[doc. 10] Rainsford, T., Guillot, C., Lavrentiev, A. & Prévost, S. (2012). « La zone préverbale en ancien français : apport des corpus annotés ». Actes du 3e Congrès Mondial de Linguistique Française, EDP Sciences (www.linguistiquefrancaise.org), 159-176, [http://dx.doi.org/10.1051/shsconf/20120100246].

5. Publications à paraître

[doc. 13] Guillot, C., Lavrentiev, A., Pincemin, B. & Heiden, S. (sous presse). « 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)*.

[XIII] Guillot, C. & Marchello-Nizia, C. (soumis). « Le cas du démonstratif français : spécialisation morphosyntaxique et changement sémantique ». *Langages*.

[doc. 14] Guillot, C., Heiden, S., Lavrentiev, A. & Pincemin, B. (accepté). « L'oral représenté dans un corpus de français médiéval (9e-15e) : paramètres de variation diamésique (discours direct vs récit), générique et diachronique », Actes du colloque international *Les variations diasystémiques et leurs interdépendances* (19-21 novembre 2012, Copenhague)

[doc. 15] Guillot, C. & Carlier, A. (accepté). « Evolution des démonstratifs du latin au français : le passage d'un système ternaire à un système binaire », Actes du colloque international *DIACHRO-VI Le français en diachronie* (17-19 octobre 2012, Louvain)

6. Comptes-rendus d'ouvrages

Compte-rendu de S. Prévost, La postposition du sujet en français aux XVe et XVIe siècles. Analyse sémantico-pragmatique, *Bulletin de la Société de linguistique de Paris*, 2003, XCVIII, fasc. 2, p. 295-296.

Compte-rendu de F. Duval (éd.), Frédéric Godefroy. Actes du Xe Colloque International sur le moyen français, organisé à Metz du 12 au 14 juin 2002 par le centre « Michel Baude, littérature et spiritualité » et par l'ATILF, *Romanische Forschungen*, 2006, 118/2, p. 221-223.

Compte-rendu de K. Sellevold, « J´ayme ces mots. » : Expressions linguistiques de doute dans les Essais de Montaigne, *Romanische Forschungen*, 2007, 119, p. 299-301.

7. Sites et portails Web
2006 -> Site de la Base de français médiéval (http://bfm.ens-lsh.fr/)
2009 -> Site du projet Corpus représentatif des premiers textes français (http://corptef.ens-lyon.fr/)
2012 -> Portail de la Base de français médiéval (http://txm.bfm-corpus.org/)

DIPLÔMES, QUALIFICATIONS ET DISTINCTIONS

- 2010
Membre (junior) de l'Institut universitaire de France

- 2004
Inscription sur la liste de qualification aux fonctions de Maître de conférences (section 7)
Inscription sur la liste de qualification aux fonctions de Maître de conférences (section 9)

- 2003
Thèse de doctorat en Sciences du Langage
Titre : « Le rôle du démonstratif dans la cohésion textuelle au XVème siècle. Eléments de grammaire textuelle »
Date de soutenance : 15 décembre 2003
Lieu de soutenance : ENS-LSH

Directeur de thèse : Ch. Marchello-Nizia (ENS- LSH Lyon)
Jury : B. Combettes (Nancy II), Rapporteur et Président ; C. Schnedecker (Strasbourg II), Rapporteur ; Ch. Plantin (CNRS), S. Prévost (CNRS)
Mention très honorable avec les félicitations du jury à l'unanimité

- 1994
Diplôme de Conservateur du Patrimoine (Ecole nationale du Patrimoine)

- 1992
Diplôme d'Archiviste-paléographe (Ecole nationale des Chartes)

Thèse de l'Ecole nationale des Chartes
Titre : « Edition critique du *Des cas des nobles hommes et femmes* de Laurent de Premierfait (1409) et commentaire linguistique »
Directeur de thèse : F. Vielliard
Thèse signalée au Ministre

DEA de Sciences du Langage à l'EHESS
Titre : « Analyse des temps verbaux dans le *De casibus virorum illustrium* ; traduction française : *Des cas des nobles hommes et femmes* »
Direction : O. Ducrot, J.-Cl. Anscombre

Articles cités

(par ordre chronologique de publication)

[doc. 1] Heiden, S. & Guillot, C. (2003). « Capitalisation des savoirs par le Web : une application de la TEI pour l'encodage et l'exploitation des textes de la Base de français médiéval ». In : P. Kunstmann, F. Martineau, D. Forget (éd.), *Ancien et moyen français sur le Web. Enjeux méthodologiques et analyse de discours*, Ottawa : Les éditions David, p. 77-92.

[doc. 2] 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, p. 29-36.

[doc. 3] Guillot, C., Heiden, S. & Lavrentiev Alexei (2007). « Typologie des textes et des phénomènes linguistiques pour l'analyse du changement linguistique avec la Base de Français Médiéval », *Linx*, numéro spécial, p. 125-139.

[doc. 4] Pincemin, B., Guillot, C., Heiden, S. & Lavrentiev Alexei (2008). « Usages linguistiques de la textométrie. Analyse qualitative de la consultation de la Base de Français Médiéval via le logiciel Weblex », *Syntaxe & sémantique* 9, p. 87-110.

[doc. 5] Guillot, C., Heiden, S., Lavrentiev, A. & Marchello-Nizia, C. (2008). « Constitution et exploitation des corpus d'ancien français et de moyen français », *Corpus* 7, p. 5-23.

[doc. 6] Guillot, C. (2009) « Ecrit médiéval et traces d'oralité : l'exemple de l'adverbe *or(e)* ». In : E. Havu *et al*. (éd.), *La langue en contexte. Actes du colloque Représentation du sens linguistique IV (Helsinki, 28-30 mai 2008)*, Helsinki : Société Néophilologique, p. 267-281.

[doc. 7] Guillot, C. & Lavrentiev, A. (dir.) (version 4.2. mars 2009). *Présentation des descripteurs du projet CORPTEF* ([doc. http://corptef.ens-lyon.fr/spip.php?rubrique60

[doc. 8] Guillot, C. (2010a) « Le démonstratif de notoriété de l'ancien français : approche textuelle », In : B. Combettes et al. (éd.). *Le changement en français. Etudes de linguistique diachronique*, Bern/Berlin/Bruxelles : Peter Lang, p. 217-233.

[doc. 9] Guillot, C. (2010b) « Les démonstratifs de l'ancien français : un système encore personnel ? ». Actes du 2e Congrès Mondial de Linguistique Française, EDP Sciences (www.linguistiquefrancaise.org), [http://dx.doi.org/10.1051/cmlf/2010085].

[doc. 10] Rainsford, T., Guillot, C., Lavrentiev, A. & Prévost, S. (2012). « La zone préverbale en ancien français : apport des corpus annotés ». Actes du 3e Congrès Mondial de Linguistique Française, EDP Sciences (www.linguistiquefrancaise.org), 159-176, [http://dx.doi.org/10.1051/shsconf/20120100246].

[doc. 11] Guillot, C. (2012a). « Système des démonstratifs médiévaux et exemples de stratégies communicatives », *Journal of French Language Studies*, disponible en ligne [CJO 2012 doi:10.1017/S0959269512000245].

[doc. 12] Guillot, C. (2012b) « Le pronom anaphorique *cil* de l'ancien français : continuité ou discontinuité topicale ? ». In : C. Denizot & E. Dupraz (éd.). *Anaphore et anaphoriques : variété des langues, variété des emplois*, Mont-Saint-Aignan : Publications des universités de Rouen et du Havre (Cahiers de l'ERIAC), p. 97-115.

[doc. 13] Guillot, C., Lavrentiev, A., Pincemin, B. & Heiden, S. (à par. a). « 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)*.

[doc. 14] Guillot, C., Heiden, S., Lavrentiev, A. & Pincemin, B. (à par. b). « L'oral représenté dans un corpus de français médiéval (9e-15^{e}) : paramètres de variation diamésique (discours direct vs récit), générique et diachronique », Actes du colloque international *Les variations diasystémiques et leurs interdépendances* (19-21 novembre 2012, Copenhague)

[doc. 15] Guillot, C. & Carlier, A. (à par. c). « Evolution des démonstratifs du latin au français : le passage d'un système ternaire à un système binaire », Actes du colloque international *DIACHRO-VI Le français en diachronie* (17-19 octobre 2012, Louvain)