Constitution et exploitation d'un corpus de français médiéval : enjeux, spécificités et apports

Sophie Prévost

Chapitre 4

Constitution et exploitation d'un corpus de français médiéval : enjeux, spécificités et apports[1]

4.1. Introduction

Les linguistes qui s'intéressent à des états de langue contemporains peuvent choisir de recourir ou non au « corpus »[2], comme source principale de leurs données ou comme complément de leur intuition de locuteur ou de celle de leurs informateurs. Les linguistes qui travaillent sur des états de langues disparus, en l'occurrence le français médiéval, n'ont pas ce choix : le corpus est indispensable, puisqu'il conditionne l'existence même de l'objet à décrire. Cette situation n'est pas nouvelle, mais elle n'a pas pour autant suscité, de manière précoce, un questionnement véritable de l'utilisation du corpus : la réflexion liée à la langue ancienne s'est esquissée dans la mouvance de la réflexion générale, née de l'intérêt croissant que la linguistique descriptive a porté, depuis une bonne vingtaine d'années, au travail sur les corpus (désormais numérisés). Elle a cependant suivi en

---

Chapitre rédigé par Sophie PRÉVOST.

1. Merci à Benoît Habert pour la relecture minutieuse qu'il a faite de ce chapitre, ainsi qu'à Christiane Marchello-Nizia et Bernard Victorri pour leur lecture des premières pages. Ce chapitre doit en outre beaucoup aux travaux de C. Marchello-Nizia, dont nous espérons, lorsque nous l'avons citée, ne pas avoir trahi la pensée.

2. Le terme est pris pour l'instant dans le sens assez vague de textes, écrits ou oraux, sans précision quantitative et/ou qualitative.

partie sa propre voie, car travailler sur le français médiéval[3] soulève des difficultés proprement liées à cet objet d'étude.

Le présent chapitre se penche donc sur les précautions, les contraintes et les spécificités liées à la constitution et à l'exploitation d'un corpus de français médiéval dans le cadre d'une étude de linguistique historique, que ce soit dans une perspective synchronique ou diachronique, tout en envisageant, à l'aide de quelques exemples, en quoi et comment le travail sur corpus permet d'enrichir – voire de modifier – la connaissance d'un état de langue dont nous n'avons pas la compétence.

4.2. Les spécificités de l'objet étudié

4.2.1. *Une langue sans locuteurs*

Bien que le français que nous parlons aujourd'hui soit l'héritier de celui du 15e siècle, lui-même issu de celui des siècles précédents, la plupart des locuteurs contemporains restent bien perplexes face à un texte d'ancien ou de moyen français. En quelques siècles, la langue a profondément changé, et si un texte du 15e siècle nous « parle » encore un peu, ce n'est pas le cas pour un texte d'ancien français, *a fortiori* du 11e siècle.

Le locuteur d'aujourd'hui n'a pas la compétence de la langue ancienne, et il ne peut en outre faire appel à des informateurs possédant cette compétence. Il ne peut que consulter des textes datant de cette époque, dont la caractéristique majeure est de nous avoir été transmis sous la seule forme écrite, même si certains d'entre eux avaient été composés pour être dits. Pour reprendre une distinction proposée par C. Marchello-Nizia (Marchello-Nizia, 1995 : 22), le linguiste médiéviste, à défaut d'une compétence de « production », possède au mieux, une compétence de « reconnaissance »[4].

---

3. « Le » français médiéval : il faut d'emblée souligner le caractère trompeur de cette appellation, qui semble conférer un semblant d'unité à son objet, lequel n'est qu'une construction *a posteriori*, et qui présente en outre une variation importante : du 9e au 15e siècle, la langue a grandement évolué, de même qu'elle a revêtu, sur une même période, des formes assez différentes selon les régions. Ce label recouvre en fait deux périodes, celles de l'ancien et du moyen français (9e-13e et 14e-15e siècles), mais celles-ci ne correspondent pas non plus à des états de langue stables et homogènes. Nous reviendrons sur ce point.

4. Toutefois, pour éviter l'ambiguïté attachée à ce terme, selon qu'il est pris dans un sens plus ou moins étroit (voir la double notion de « compétence » / « performance » proposée par Chomsky), nous lui préfèrerons celui d' « intuition », moins lourd de connotations.

Contrairement au français moderne pour lequel il y a une évidence vécue de la réalité de l'objet, notre connaissance du français médiéval a donc ceci de particulier qu'elle provient nécessairement et exclusivement des textes, c'est-à-dire de l'objet même d'étude. La fréquentation assidue de ces derniers conduit à en abstraire les règles de fonctionnement, ou bien, plus souvent, elle complète et exemplifie une grammaire acquise par ailleurs, cette dernière s'étant bâtie sur l'exploitation d'un corpus, et, dans une mesure variable, sur la consultation des premières grammaires du français au 16e siècle (H. Estienne, T. de Bèze, J. Dubois, Meigret, etc. et avant eux celle de Palsgrave en Angleterre) ou des témoignages de l'époque[5]. A la dépendance à l'égard d'un « corpus » écrit s'ajoute le fait que les conditions de production de ce dernier restent souvent assez obscures, de même que ses modalités d'utilisation dans les grammaires ne sont pas toujours bien claires[6].

Le linguiste médiéviste semble soumis à une situation circulaire : étudier des textes avec une connaissance de la langue exclusivement fondée sur ces mêmes textes ! Une démarche critique permet néanmoins de sortir de ce cercle vicieux : l'analyse de textes déjà étudiés le conduit à enrichir la connaissance de la langue en se penchant sur des faits non encore mis au jour ou approfondis, et la prise en

5. Le rôle de ces grammaires et témoignages est d'ailleurs complexe. Ils fournissent un éclairage sur certains usages de leur époque, ou de l'époque précédente, mais, du fait de leur caractère souvent prescriptif, il n'est pas facile de faire la part entre les recommandations et les pratiques réelles, écrites ou orales. Il est en outre souvent difficile de déterminer la place qu'ont occupée ces témoignages dans l'élaboration des grammaires contemporaines du français médiéval : ont-ils été consultés directement et exhaustivement, ou simplement utilisés indirectement à travers le savoir qu'ils ont transmis au fil des siècles ?

6. Les textes utilisés sont à notre connaissance toujours cités, et il est d'ailleurs intéressant de constater des écarts assez importants (liés à la fois aux possibilités nouvelles offertes par la numérisation et à des choix personnels). A titre d'exemple, *La Petite Syntaxe de l'ancien français* de L. Foulet (1963, 1re éd. 1919) s'appuie sur 18 textes, la *Grammaire de l'ancien français* de Moignet (1984) sur 102 textes, et dans sa *Grammaire nouvelle de l'ancien français,* C. Buridant (2000) mentionne plus de 350 textes auxquels s'ajoutent des index et concordances électroniques. Pour la période suivante, la *Syntaxe du moyen français* (1980) de R. Martin et M. Wilmet repose sur 24 textes (dont 3 numérisés et analysés exhaustivement), tandis que *La Langue française aux 14e et 15e siècles* de C. Marchello-Nizia (1997, 1re éd. 1979) mentionne 69 textes. Les écarts sont certes importants (même si l'on regrette de ne pouvoir les préciser en nombre d'occurrences), mais l'essentiel réside davantage dans l'utilisation potentiellement variable des textes : ont-ils subi un dépouillement exhaustif ou été l'objet de sondages réguliers ? Ou bien ne sont-ils que des réservoirs à exemple ? Ont-ils par ailleurs tous été soumis au même traitement ? Ces questions ne sont guère élucidées (même si C. Marchello-Nizia distingue les « textes dépouillés ou consultés » et ceux « auxquels il est fait occasionnellement référence »). Ces choix ne sont pourtant pas sans conséquence sur les descriptions proposées.

compte de textes encore peu sollicités lui permet de confirmer ou d'infirmer l'existence de phénomènes ou de règles précédemment avancés, et peut-être de compléter ces dernières. C'est grâce à cet accroissement du corpus que peut progresser la connaissance de la langue (et grâce aussi, nous le verrons, aux réflexions théoriques et méthodologiques sur la nature du corpus et son exploitation différenciée).

Le corpus est indispensable au linguiste médiéviste, mais ce dernier a la possibilité de relativiser ce que révèlent les textes, par une maîtrise de leurs caractéristiques externes et de leurs conditions de production. Cette démarche, nécessaire, rencontre cependant certaines difficultés, comme on le verra.

4.2.2. *La difficile délimitation de l'objet langagier*

Nul n'entendra jamais le « son » de l'ancien ou du moyen français, les prononciations reconstruites gardant leur part d'inconnu et d'incertitude et ne nous disant en outre que bien peu sur la prosodie de cette époque[7]. Outre la musique de la langue, ce sont aussi ses spécificités, en particulier syntaxiques, qui nous échappent. Or on peut supposer que, comme aujourd'hui, l'écart entre écrit et oral n'était pas négligeable. Nous possédons certes quelques indications, certaines provenant des grammaires, manières de dire et autres témoignages de l'époque, d'autres d'une mise en écrit de l'oral. Mais qu'il s'agisse de dialogues dans un roman ou une pièce de théâtre, on sait que l'écart est souvent important avec l'oral réel, celui-ci étant passé au crible de la norme, ou au contraire caricaturé.

C'est donc bien la seule langue écrite que peut étudier le linguiste, mais parler de « la » langue, même écrite, n'a pas grand sens, *a fortiori* lorsqu'il s'agit d'une période de sept siècles ! Français médiéval, langue de la Renaissance, français classique, etc. : ces différents labels, utiles, ont néanmoins l'inconvénient de favoriser une conception discrétisante de la langue, et en partie artificielle, dans la mesure où ils résultent d'un découpage *a posteriori*, et en fonction de critères assez modernes, même si les locuteurs de l'époque (certains au moins) ont parfois pu avoir le sentiment de parler une langue suffisamment modifiée pour qu'il faille lui donner un nouveau nom. C'est une telle situation qu'a entérinée le Concile de Tours en 813 en demandant aux ecclésiastiques de désormais prononcer leurs prêches en *lingua*

7. Même si l'on peut supposer, comme le suggère C. Marchello-Nizia (1995 : chap. 6), que le français médiéval a connu un accent de microsyntagme, entre l'accent de mot du latin et l'accent de macrosyntagme ou de phrase du français moderne.

romana rustica, le latin s'étant suffisamment érodé au fil des siècles pour que sa forme classique ne puisse plus être comprise des fidèles, ni que la langue parlée ne puisse porter ce nom.

Le terme de « français médiéval » recouvre certes l'ancien et le moyen français, le premier s'étendant du 9e au 13e siècle (de 1150 à 1300 pour l'ancien français « classique »), le second couvrant les 14e et 15e siècles[8]. Mais au sein de ces périodes des évolutions ont eu lieu, que l'on ne perçoit d'ailleurs qu'avec le recul du temps. Par exemple, comme l'a montré C. Marchello-Nizia (Marchello-Nizia, 1995), la position de l'objet nominal est devenue majoritairement postverbale au 13e siècle, ce qui signifie que, en deux siècles, sa syntaxe a largement évolué. Mais en ce qui concerne d'autres faits, en particulier la syntaxe du sujet, nominal ou pronominal, la situation reste globalement stationnaire durant cette période et ne commencera à évoluer qu'au siècle suivant, et ce pendant trois siècles.

Il apparaît donc qu'ancien français et moyen français, du fait de l'évolution de certains faits langagiers, ne correspondent pas à des périodes globalement homogènes et stables : au sein de ces deux vastes synchronies s'opèrent des changements qui introduisent une nécessaire dimension diachronique. Comme le rappelle C. Marchello-Nizia (Marchello-Nizia, 1997 : 7) : « En fait, tout état de langue est à la fois, nécessairement, "transition" et "stabilité", dans la mesure où toute langue naturelle change, continûment, tout en maintenant un équilibre indispensable à l'intercompréhesnion. »

Il n'est donc pas simple de déterminer la langue que l'on étudie, la période pertinente variant selon la perspective adoptée : soit rendre compte de l'ensemble des faits langagiers, du « système », à une période donnée (et si celle-ci est longue, deux siècles par exemple, il faut introduire une perspective diachronique ou bien figer certains faits), soit s'attacher à un phénomène particulier, saisi dans sa stabilité ou au contraire dans son évolution (et c'est alors l'objet lui-même qui détermine la période pertinente). Ainsi, pour décrire un système bicasuel nominal encore bien vivant et stable, mieux vaut considérer l'ancien français jusqu'au 12e siècle. En revanche, pour rendre compte de son effritement mieux vaut s'intéresser au 13e

8. L'exacte délimitation entre les deux périodes n'est pas totalement consensuelle, et il faut d'ailleurs préciser qu'elle s'appuie non seulement sur des critères linguistiques (dont la réduction des diphtongues et des hiatus, et la disparition de la déclinaison bicasuelle à la jonction des 13e et 14e siècles), mais aussi politico-socioculturels. Pour une présentation détaillée de cette question, voir Marchello-Nizia (1997 : 3-9).

siècle, et, plus encore à la période suivante, le moyen français, qui voit voler le système en éclats.

4.2.3. *Synchronie et diachronie*[9]

Il est essentiel de distinguer approche synchronique et approche diachronique (la première n'excluant nullement la prise en compte d'une dimension diachronique), même si c'est pour évoquer les difficultés que l'une et l'autre soulèvent.

La première tient à ce qu'une démarche véritablement diachronique ne se réduit pas à la prise en compte d'états synchroniques successifs. Elle doit expliquer comment l'on passe de l'un à l'autre. Il ne suffit pas d'enregistrer les changements, il faut rendre compte de l'évolution : « Le changement, c'est le résultat, ce que l'on constate. L'évolution, c'est le processus invisible, et largement inconscient, que le linguiste a pour tâche d'expliquer, c'est-à-dire de reconstruire » (Marchello-Nizia, 1995 : 28-29).

Qu'elle soit synchronique ou diachronique, la démarche se heurte à la difficile délimitation « chronologique ». Pour une approche synchronique, se pose la question de la taille de la fenêtre temporelle adéquate pour rendre compte d'une large période (ancien ou moyen français dans leur totalité par exemple). Faut-il considérer les données sur l'ensemble de l'intervalle, ou peut-on se contenter de coupes plus étroites mais supposées représentatives de ladite période ? Dans ce cas, quelle est la taille de la bonne fenêtre ? Dix ans[10] ? Vingt ans ? La réponse dépend probablement du niveau d'analyse visé, les faits ne bougeant pas au même rythme en sémantique lexicale, en morphologie ou en syntaxe[11].

Pour une approche diachronique, il ne suffit certes pas de décrire des synchronies successives, mais c'est néanmoins un préalable nécessaire, et il faut donc déterminer à la fois la taille de celles-ci et leur fréquence : tous les dix ans ? tous les vingt ans ? Ou bien faut-il au contraire envisager une couverture exhaustive (par exemple :

9. Rappelons (voir Marchello-Nizia 1995 : 28) que la linguistique diachronique s'oppose à la linguistique historique en ce sens que, contrairement à la seconde, elle ne prend en charge que l'histoire interne de la langue, rejoignant en cela la linguistique synchronique, dont elle se distingue en traitant de la dimension évolutive de la langue.
10. C'est le choix fait par Martin et Wilmet pour la *Syntaxe du moyen français* (1980) : ils retiennent la décennie 1455-1465.
11. Merci à Benoît Habert de m'avoir signalé ce point, qui exigerait une étude approfondie.

1220-1230, 1230-1240, 1240-1250, etc.) ? Laisser dans l'ombre une période risque d'occulter des faits notables, en particulier s'il s'agit de dater l'apparition ou la disparition d'une forme, ou bien d'attester une forme rare.

Au-delà, c'est la nature même de l'évolution qui est en cause, selon que l'on considère que les changements surviennent brusquement, ou qu'ils s'inscrivent dans un continuum, avec par exemple coexistence temporaire d'une forme nouvelle et d'une forme ancienne, la seconde, initialement dominante, devenant minoritaire avant de disparaître ou de se refaire une jeunesse dans d'autres emplois. C'est par exemple le cas de l'adverbe « moult », progressivement supplanté en moyen français par « très » et « beaucoup ». Les textes nous montrent que c'est plutôt selon un mode continu qu'évoluent les faits langagiers, au moins en ce qui concerne la morphologie et la syntaxe[12]. On peut dès lors s'interroger sur le bien-fondé d'une approche qui discrétise (et avec des ellipses) l'espace temporel.

Les problèmes de délimitation temporelle de la langue étudiée ne sont pas exclusifs des langues anciennes : le français décrit il y a vingt ans est-il encore celui d'aujourd'hui[13] ? L'abondance des documents et l'existence de locuteurs compétents diminue cependant le risque de « passer » à côté d'une d'une construction[14-15].

4.2.4. *La variation dialectale*

A la variation temporelle s'ajoute la variation dialectale, dont nous rappellerons quelques aspects essentiels[16].

A ses débuts, le « français » ne correspondait pas à une langue commune à tous les locuteurs, mais à des dialectes (francien, normand, champenois, picard, orléanais, etc.), situation que favorisait en particulier l'absence de langue officielle vernaculaire. Contrairement à ce que l'on a longtemps pensé, ce qui deviendra le français ne semble pas résulter de l'accession d'un dialecte, le francien, au rang de

12. On observe dans le domaine lexical des phénomènes plus abrupts, dûs à la création ou à l'importation de l'étranger d'un nouveau terme.
13. Encore faudrait-il préciser ce que l'on entend par « aujourd'hui »...
14. *Forme* et *construction* ont dans notre usage un même sens général : morphème, syntagme, construction verbale, tournure phrastique etc.
15. Il est vrai qu'à l'inverse la difficulté à couvrir de façon exhaustive toutes les productions, même sur une période brève (1995-2000 par exemple), accroît ce risque.
16. Largement inspirés de Perret (1998 : 52-62).

langue officielle, puis majoritaire. En effet, on rencontre dès les premiers textes une langue qui se veut commune, *scripta transdialectale* pour reprendre les termes de B. Cerquiglini (Cerquiglini, 1991), dans laquelle, il est vrai, les traits de la région Ile-de-France élargie dominent. Cette langue s'étend à partir du 13e siècle, gagnant du terrain dans les documents juridiques et reléguant progressivement les dialectes au rang de simples patois.

Si la variation dialectale reste longtemps admise dans les textes littéraires (au moins jusqu'au 14e siècle), elle n'en demeure pas moins – *scripta* oblige – assez limitée, ne compromettant pas la compréhension du texte par les locuteurs d'autres dialectes. Concernant en premier lieu le phonétisme (et donc en partie la morphologie), les graphies et le lexique, elle affecte en revanche peu la syntaxe.

L'incidence de la variation dialectale est donc plus limitée que l'on pourrait le penser, au moins à l'écrit, seule trace tangible qui nous reste de cette époque.

A ces deux formes de variation, temporelle et dialectale, s'ajoute celle liée aux genres, aux registres. Elle n'est assurément pas spécifique à l'objet « français médiéval », mais celui-ci induit cependant certaines particularités, qui seront évoquées en même temps que la question de la représentativité du corpus.

4.3. Corpus et bases textuelles de français médiéval

4.3.1. *Corpus vs base textuelle*

Depuis plusieurs années, les travaux sur « corpus » se multiplient, signe d'une prise de conscience quant à la nécessité d'affronter les textes mêmes. Bien que ce mouvement se soit accompagné d'une réflexion sur la notion même de corpus, il a aussi vu éclore un certain nombre de travaux qui se disent « sur corpus », simplement parce qu'ils s'appuient sur quelques textes. Le caractère parfois intempestif de ce label n'en trahit pas moins la difficulté à appréhender la notion.

S'il ne suffit pas de rassembler quelques textes pour qu'ils constituent un corpus, déterminer des caractéristiques de l'ensemble pour qu'il puisse prétendre à ce statut demeure complexe et sujet à variation. Il existe cependant un relatif consensus sur deux points essentiels et corrélés : le corpus est constitué en vue d'un certain but, et, eu égard à ce but, il doit offrir une certaine cohérence et une relative représentativité (quantitative et qualitative). Il s'oppose en cela aux bases textuelles (ou bases de

données), dont la constitution ne répond pas à un objectif précis, et obéit donc à des critères moins rigoureux en matière de représentativité[17]. Bien qu'une base textuelle soit souvent une « réserve » au sein de laquelle on puise pour forger un corpus, la limite entre les deux n'est cependant pas toujours bien nette. En effet, comme le souligne C. Marchello-Nizia (Marchello-Nizia, 1999 : 32) : « il n'est pas rare qu'un corpus finalisé puisse se transformer par la suite en base de données généraliste offerte à d'autres chercheurs et ouverte à de nouvelles recherches ; et inversement il n'est pas rare qu'une base de données soit adoptée telle quelle comme corpus pour telle ou telle recherche. Dans ces deux cas on conviendra que la frontière est difficile à maintenir ». Nous verrons que le second volet de cette remarque trouve une résonnance particulière en ce qui concerne la langue ancienne.

En relation avec les notions de base textuelle et de corpus, il convient d'évoquer celle de « corpus de référence »[18], généralement associé à la notion de langue générale. Tout en relevant encore, selon nous, de la gageure, y compris pour les états de langue moderne, la constitution d'un tel corpus n'en demeure pas moins un idéal vers lequel il est nécessaire de tendre[19]. Les travaux de D. Biber[20] ont en effet souligné le caractère artificiel d'une « langue générale » conçue comme une entité : la variation sous-tend la langue, ensemble composite de différents « registres » dont il convient de dégager les caractéristiques. Mais, comme le fait remarquer M.-P. Péry-Woodley (Péry-Woodley, 1995 : 218) en envisageant la constitution d'un corpus équilibré représentatif de la langue générale, la non-maîtrise de la variation langagière reste encore un obstacle majeur : « le corpus équilibré est sans doute celui qui a "de tout un peu", mais encore faudrait-il savoir ce qu'est "tout", c'est-à-dire quelles sont les classes à représenter – ce qui nécessite un modèle complet de la variation -, et avoir accès à des textes les représentant ». C'est un point de vue analogue qui conduit B. Habert (Habert, 2000) à proposer une définition de la notion

17. La réflexion, abondante en ce qui concerne les corpus, est en revanche assez pauvre pour ce qui est des bases textuelles. Or, que celles-ci ne soient pas soumises aux mêmes exigences que les « corpus » ne signifie pas pour autant qu'elles ne doivent obéir à aucune contrainte, ne serait-ce que quantitative : parlera-t-on de « base textuelle » pour le groupement de deux textes de 3 000 caractères chacun ? Y a t-il un seuil minimal – et si oui de quel type ? – qui autorise à parler de base textuelle ?

18. Souvent mise en relation avec celle de « corpus de spécialité », que nous n'aborderons pas ici. Pour une discussion à ce sujet, voir, entre autres, les différents travaux de D. Biber, de B. Habert, ainsi que Péry-Woodley (1995) et Condamines (2000).

19. De ce point de vue, le BNC (British National Corpus) est assez exemplaire : corpus de 100 millions de mots étiquetés, il définit précisément les conditions extralinguistiques de ses données textuelles, représentatives d'une grande variété de situations de communication.

20. Biber (1988), (1990), (1993) et (1995) en particulier.

de corpus qui, tout en s'appuyant sur celle de J. Sinclair (Sinclair, 1996 : 4), la restreint à deux égards : « un corpus est une collection de données langagières qui sont sélectionnées et organisées selon des critères linguistiques et *extralinguistiques* explicites pour servir d'échantillon d'*emplois déterminés d'une langue* ». Outre l'ajout de critères extralinguistiques, B. Habert limite en effet le concept de langage à celui d' « emplois déterminés d'une langue », justifiant cette position par « notre ignorance de la population d'événements que constitue un langage dans son ensemble »[21]. Une telle réserve est plus fondée encore pour la langue ancienne, notre « ignorance de la population d'événements » étant bien plus grande. La notion de corpus de référence est donc d'emblée exclue en ce qui concerne toute période du français médiéval : nous ne pouvons en maîtriser la variation, ne serait-ce que parce que nous n'aurons jamais accès à certains registres (en particulier ceux de l'oral), et que nous ignorons même probablement l'existence de beaucoup.

4.3.2. *La représentativité du corpus*

Ecarter la perspective d'un corpus de référence pour « la » langue n'exclut pas de tendre vers un corpus offrant la meilleure représentativité possible. Variant en fonction du but à atteindre (rendre compte d'une langue dans sa variété ou au contraire de certains usages), celle-ci se décline néanmoins toujours selon deux dimensions, quantitative et qualitative, dont l'antagonisme de fait tend désormais à se résoudre dans une conception complémentaire des deux[22] : l'accumulation des données, permise par les progrès spectaculaires de l'informatique, est en effet peu utile si elle ne s'accompagne pas de leur diversification raisonnée.

Pour l'aspect quantitatif, il convient de choisir le nombre de catégories textuelles retenues, et, parmi elles, le nombre de textes ou d'échantillons[23], ainsi que leur taille en mots (question qui ne se pose pas pour les textes intégraux, mais leur possible variation de taille exige néanmoins de pondérer les résultats des relevés opérés). Les propositions avancées, en particulier par Biber[24], demeurent indicatives, destinées à

21. Habert (2000). La question est en outre discutée dans Habert et Zweigenbaum (2002).
22. Pour un historique de cette question, voir Péry-Woodley (1995).
23. Précisons que le recours à l'échantillonnage est souvent motivé (aujourd'hui encore) par le coût du traitement de documents entiers.
24. S'appuyant sur les caractéristiques du LOB corpus, Biber (1990) suggérait de retenir 10 textes par catégorie, et, s'appuyant sur la représentativité de nombreux traits grammaticaux, il recommandait une moyenne de 1 000 mots par échantillon, tout en admettant le besoin d'un échantillon plus important pour certains traits plus rares.

varier selon le type d'étude à mener : faute de critères totalement objectifs pour la fonder, il est difficile d'envisager une « norme » absolue en matière de taille. La démarche consiste pour beaucoup à s'appuyer sur ce qui a déjà été fait, en France ou à l'étranger, tout en veillant à la cohérence et à l'homogénéisation quantitatives internes du corpus.

Le choix d'utiliser des textes intégraux ou des échantillons revêt une dimension particulière lorsqu'il s'agit de la langue ancienne. Précisons que la question ne se pose pas pour les textes très brefs, qui, même dans leur intégralité, peuvent sembler bien courts pour être représentatifs. Toutefois, ce sont parfois les uniques témoignages que nous possédons d'un certain état de langue. C'est le cas, pour le 9e siècle, des *Serments de Strasbourg* (842, une petite centaine de mots) et de la *Séquence de sainte Eulalie* (vers 880, moins de 200 mots).

D'une manière générale, le recours à des textes intégraux est souvent nécessaire, dans la mesure où bon nombre d'études impliquent la quête de constructions rares, ou l'attestation de formes à l'existence incertaine. Alors que le locuteur moderne peut faire appel à sa compétence pour décider si la forme existe ou non[25], le linguiste médiéviste doit s'en remettre aux seuls textes, et ne peut se contenter d'échantillons.

La prudence dans l'échantillonnage est nécessaire aussi du fait qu'un texte est susceptible d'être hétérogène. Il peut l'être d'un point de vue narratif[26], mais aussi « chronologique », la rédaction de l'œuvre ayant pu être longue, et la langue de l'auteur se modifier quelque peu en cours de route. C'est la cas des *Mémoires* de Commynes, dont les livres successifs ont été rédigés de 1489 à 1498[27]. Il se peut aussi que l'écriture se soit faite à plusieurs mains, avec un éventuel décalage dans le temps : c'est le cas du *Roman de la rose*, dont la première partie a été rédigée par G. de Lorris en 1225-1230, et la seconde par J. de Meun en 1269-1278.

Pour ce qui est de la dimension qualitative, la diversité revêt une importance cruciale lorsqu'il s'agit d'élaborer un corpus à visée généraliste. La difficulté préalable réside dans la détermination des genres, des conditions de production et des buts visés, ainsi que de certaines marques linguistiques jugées pertinentes et

25. Même s'il est vrai qu'un locuteur peut ignorer l'existence de certaines constructions.
26. Voir les remarques de Guillot (2003) à propos *Des cas des nobles hommes* (Boccace, traduit par Laurent de Premierfait), qui alterne plusieurs types narratifs.
27. L'étude approfondie du livre 1 et des sondages dans le livre 7 (l'avant-dernier) nous ont montré que la syntaxe du sujet a évolué de manière non négligeable de l'un à l'autre.

caractéristiques[28]. La classification « idéale » conjugue ces différents critères... mais se heurte à leur évolution dans le temps.

La représentativité quantitative et qualitative du corpus est pareillement nécessaire lorsqu'il s'agit du français médiéval, mais l'on se heurte à des difficultés supplémentaires. En préalable, il faut évoquer le manque de disponibilité ou la rareté de certaines données : il n'est pas toujours facile de réunir pour telle époque tel nombre de textes de telle catégorie, *a fortiori* numérisés.

Se pose en outre la question de la classification des textes. Complexe pour les états de langue moderne, elle l'est davantage encore pour ceux de langue ancienne. Tout d'abord, les « genres »[29] de l'époque diffèrent en partie de ceux d'aujourd'hui : les « romans » des 12e et 13e siècles n'ont ainsi que bien peu à voir avec le roman moderne[30]. Il est de plus difficile de classer certains écrits : c'est le cas des textes « historiques » (chroniques, mémoires, etc.), ne serait-ce que parce qu'ils comprennent bien souvent une part de fiction (caractéristique des genres littéraires) et dénotent une implication assez forte de l'auteur. Ils sont généralement apparentés au genre littéraire.

Par ailleurs, certains genres ne nous sont parvenus qu'à travers bien peu de textes, d'où la difficulté à les caractériser. Peut-être même ignorons-nous l'existence de certains ? En outre, si les visées de certains textes nous sont connues, nous ne maîtrisons que bien imparfaitement leurs conditions de production et de réception. A ces différents handicaps s'ajoute le fait que des genres apparaissent, d'autres disparaissent, et surtout certains évoluent, ce qui est complexe à gérer dans la perspective d'une approche diachronique : faut-il considérer qu'il s'agit toujours du même genre, ou au contraire d'un autre ? Sur quels critères objectifs s'appuyer pour mesurer le changement, et quel est le « seuil » éventuel au-delà duquel on ne peut plus considérer que l'on a affaire au même genre ?

---

28. Aux classifications *a priori* des textes s'opposent des typologies inductives, qui dégagent des types de textes à partir d'un traitement multidimensionnel de traits linguistiques. Pour une présentation des deux démarches, voir Habert (2000), ainsi que les travaux de Biber pour une présentation détaillée de la seconde (1998 entre autres pour un bilan).

29. Nous adoptons ici une terminologie simplifiée, distinguant les « genres », fondés sur des critères extralinguistiques (en partie intuitifs et donc subjectifs) et les « types », fondés sur des critères linguistiques. Pour les genres, nous opérons une macrodistinction entre genres littéraires et non-littéraires. Parmi les premiers on trouve : roman, nouvelle, chanson de geste, poésie, théâtre (qui peuvent donner lieu à des sous-genres : comédie, tragédie, etc.). Parmi les seconds on trouve : texte juridique, charte, coutumier, etc.

30. Si tant est que l'on puisse définir celui-ci.

Le bon ciblage des genres considérés est essentiel, et ne pas prendre en compte ce facteur peut conduire à des conclusions hâtives et erronnées. Dans le cadre d'une étude consacrée à l'évolution du marqueur de topicalisation « quant à » (Prévost, 2003), la grande rareté des occurrences dans les textes d'ancien français de la BFM nous a d'abord fait penser qu'il s'agissait d'une forme peu utilisée à cette époque. Toutefois, au vu des occurrences récoltées pour le moyen français dans la base du DMF, cette interprétation s'est avérée en partie fausse.

C'est en effet dans des textes non littéraires (didactiques, argumentatifs, etc.) que nous avons rencontré la majorité des occurrences, et les textes d'ancien français consultés étant au contraire tous littéraires et narratifs, il était de ce fait peu probable de faire une collecte bien fructueuse (d'autant qu'il est vrai que la forme, tous genres confondus, reste plus rare en ancien qu'en moyen français).

Certes, plus un corpus allie importance quantitative et diversité qualitative, plus il est à même de représenter un nombre important d'états langagiers et donc de nous donner une vision plus complète de la langue.

C'est cependant toujours avec prudence qu'il faut interpréter et *a fortiori* généraliser les résultats obtenus, d'autant que, face à un corpus de langue ancienne, notre intuition de locuteur ne peut servir de garde-fou en mesurant l'écart plus ou moins grand entre l'objet construit et ce qu'il prétend représenter.

4.3.3. *Editions critiques vs éditions diplomatiques*

Le linguiste qui travaille sur une langue ancienne est potentiellement conduit à choisir entre l'utilisation d'éditions diplomatiques ou critiques[31]. Bien que le présent chapitre soit davantage consacré aux secondes, il est nécessaire d'évoquer les différents enjeux qui leur sont associés.

31. Pour une présentation détaillée de cette question voir Heiden et Lavrentiev (2004), dont nous nous inspirons ici en partie. Les éditions diplomatiques reproduisent le texte d'un manuscrit, sans y apporter de modifications, alors que les éditions critiques résultent de la comparaison de plusieurs manuscrits d'un même texte, avec corrections de certaines erreurs, le but étant d'approcher au plus près de l'original, généralement inaccessible. L'attitude des éditeurs a varié au fil des époques, aujourd'hui moins « interventionniste » qu'elle ne l'a été à une époque.

Le recours aux éditions diplomatiques garantit la fiabilité des données[32], mais suppose de consulter plusieurs manuscrits pour une même œuvre, ce qui, pour des raisons pratiques, limite souvent le nombre de textes du corpus. S'appuyer sur des éditions critiques permet à l'inverse le dépouillement d'un nombre plus important de textes, mais l'édition utilisée est toujours susceptible de porter trace de corrections intempestives.

Le choix demeure en partie conditionné par des considérations matérielles : il n'existe que rarement d'édition diplomatique numérisée pour l'ensemble des manuscrits d'une œuvre, ce qui oblige à se contenter de celles existantes ou à consulter directement les autres manuscrits. Le choix est en outre motivé par le type d'étude à mener. Pour certaines, le recours aux éditions critiques n'est pas envisageable, en particulier si la recherche porte sur la ponctuation ou la segmentation des mots, largement modernisées dans les éditions critiques, ainsi que sur certaines variantes morphologiques, parfois neutralisées par les éditeurs modernes. En revanche, bon nombre d'études syntaxiques, sémantiques ou pragmatiques, voire morphologiques, se satisfont de l'utilisation d'éditions critiques.

En ce qui concerne les éditions diplomatiques, il faut signaler que les manuscrits originaux des œuvres médiévales ne nous sont généralement pas parvenus, la plupart s'étant perdus. Ce n'est donc pas sur le manuscrit original que l'on travaille, mais sur celui d'un copiste, qui peut-être l'a rédigé bien plus tard, et parlait un autre dialecte. Il peut donc, involontairement, avoir introduit des traits langagiers propres à son usage, d'autant qu'il n'a pas eu nécessairement l'original entre les mains. C'est la raison principale des variations qui se rencontrent d'un manuscrit à l'autre. La langue d'un manuscrit n'est donc pas exactement celle de l'auteur, et si les éditions critiques modernes visent précisément à approcher au plus près cette langue, il ne s'agit jamais que d'une reconstruction.

4.3.4. *Les données numérisées pour le français médiéval : partiel état des lieux*

Alors que les linguistes médiévistes ont toujours travaillé « sur corpus »[33], celui-ci a longtemps revêtu la forme papier et, comparativement à ce qui existe pour les

32. D'autant plus qu'il est désormais possible de numériser les manuscrits et donc d'éviter bien davantage la part d'interprétation inhérente aux éditions papier. Voir à ce sujet le n°40 de la revue *Le Médiéviste et l'Ordinateur* (« La numérisation des manuscrits médiévaux »).

33. Au sens large du terme, et sans forcément s'interroger sur la représentativité des textes utilisés ni sur l'exploitation qui en est faite.

états de langue moderne[34], le retard est important en matière de données numériques disponibles. Plusieurs raisons expliquent cette situation.

Tout d'abord, malgré un intérêt croissant de nombreux linguistes non médiévistes pour la diachronie, le français médiéval demeure encore pour beaucoup un objet marginal. Cela signifie aussi que les débouchés commerciaux des corpus restent plus limités, et qu'ils ont bénéficié, tant de la part du secteur privé que public, d'une aide financière relativement restreinte. Or la constitution d'un corpus de langue ancienne est onéreuse en temps humain. L'absence de compétence ralentit en effet le rythme de la numérisation[35], d'autant que les textes médiévaux connaissent, en particulier en moyen français, d'importantes variations graphiques et morphologiques, y compris au sein d'un même texte. Ainsi, dans *Jehan de Paris* (roman de la fin du 15e siècle), on trouve pour la forme « lesquels » les graphies *lesquelz, lesquieulx* et, pour « lesdits » : *lesdicts, lesdictz*.

Il faut par ailleurs signaler que, en amont de cette coûteuse numérisation, on se heurte à des obstacles juridiques (droits d'éditeurs) pour pouvoir rendre publics, et donc accessibles à tous, les textes numérisés. Cette situation, regrettable, semble heureusement évoluer dans un sens favorable.

En dépit de ces difficultés, différentes bases textuelles ont désormais vu le jour. Nous n'en mentionnerons que certaines, qui toutes contiennent des textes intégraux[36].

Nous préférons parler de « base textuelle » plutôt que de corpus : ces ressources documentaires n'ont pas été élaborées selon des critères toujours bien arrêtés[37],

34. Voir pour l'anglais et l'américain l'inventaire dressé dans Kennedy (1998, 88-203), et pour le français, Habert *et al.* (1997), Habert (2000). Ces inventaires exigent probablement une mise à jour.

35. Le recours au scanner est moins économique que l'on pourrait le penser : non fiable à 100 %, en particulier à cause de certains caractères plus difficiles à reconnaître, il implique donc pareillement plusieurs relectures.

36. Pour la mention d'autres bases ou corpus que celles et ceux ici évoqués, voir Marchello-Nizia (1999) et Heiden & Lavrentiev (2004). Nous recommandons en outre la consultation du site de l'Institut de Recherche et d'Histoire des Textes (<http : //www.irht.cnrs.fr>), ainsi que de la revue *Le Médiéviste et l'Ordinateur* (accessible en ligne), en particulier les numéros 25 : « L'informatique et les textes en français médiéval », 38 : « Le français médiéval sur Internet », et 40 (voir note 31).

37. En raison de considérations matérielles qui obligent à des compromis dans le choix des textes.

hormis, pour beaucoup, le critère chronologique. Elles n'ont en outre pas été conçues en vue d'une utilisation unique, mais au contraire pour permettre un grand nombre de recherches, destinées à jouer le rôle de « réservoirs à corpus », quitte à parfois devenir elles-mêmes des « corpus » d'étude. Elles ont donc une visée généraliste, et s'opposent de ce point de vue à un autre type de ressources documentaires, les corpus rassemblant les différents manuscrits d'une même œuvre. C'est le cas du *Projet Charrette*[38], réalisé à l'université de Princeton à l'inititative de Karl Uitti, et visant à réunir tous les manuscrits du *Chevalier a la Charrete* de Chrétien de Troyes, tant sous forme de photos numériques que de trancriptions, à comparer avec l'édition de référence.

Pour ce qui est des bases textuelles généralistes, on mentionnera tout d'abord la base des Textes de Français Ancien (TFA)[39], établie sous la direction de P. Kunstman, au Laboratoire de Français Ancien de l'université d'Ottawa. Rassemblant initialement des textes des 12e et 13e siècles, elle s'est désormais ouverte sur le moyen français, et comporte en tout une bonne centaine de textes (dont une quarantaine de *Miracles*), qui représentent plus de 3 millions de mots (balises incluses).

En France, on citera la Base du Dictionnaire du Moyen Français (DMF), qui rassemble 218 textes et près de 7 millions de mots. Initialement conçue, à l'instar de la base Frantext pour les siècles suivants, comme devant servir de base à l'élaboration d'un dictionnaire pour le moyen français (sous l'égide de Robert Martin), elle couvre la période 1330-1500. Comme le souligne C. Marchello-Nizia (Marchello-Nizia, 1999 : 33), cette base souligne bien l'ambiguïté évoquée plus haut entre corpus et base textuelle : s'agit-il d'une base à visée généraliste, d'un corpus visant à la constitution d'un dictionnaire de langue, ou bien encore d'une tentative de corpus de référence pour la langue des 14e et 15e siècles ? Nous la concevons comme une base à visée généraliste, qui représente une source fort précieuse pour cette époque. Non ouverte au public encore récemment (en raison de son objectif initial), elle est désormais destinée à être intégrée à la base de Frantext, et est même, pour l'instant, accessible à tous[40].

---

38. <http : //www.princeton.edu/~lancelot>. Pour une présentation plus détaillée de ce projet, voir Heiden & Lavrentiev (2004), section 2.
39. <http : //www.uottawa.ca/academic/arts/lfa>
40. Sous réserve que cette information soit encore pertinente au moment de la parution de cet ouvrage : <http : //zeus.inalf.fr/dmf.htm>.

Nous évoquerons pour terminer la Base de Français Médiéval (BFM) à l'élaboration de laquelle nous avons participé[41], et que nous avons souvent utilisée[42]. La BFM s'est progressivement constituée depuis 1989 sous la direction de C. Marchello-Nizia, et au sein d'équipes successives[43]. Elle regroupe actuellement une soixantaine de textes intégraux saisis à partir d'éditions critiques, ce qui correspond à un total de 2,7 millions d'occurrences. Dans la mesure où la BFM s'est voulue complémentaire de la base du DMF, il s'agit principalement de textes d'ancien français, les deux tiers étant antérieurs à 1320, avec, il est vrai peu de textes des 9e et 10e siècles en raison de la rareté des documents existants pour cette époque. Les textes sont plutôt littéraires, les autres genres restant encore minoritaires dans les éditions modernes (un effort de diversification a néanmoins permis l'introduction de chartes et d'un livre de coutumes), et parmi les textes littéraires la plupart sont narratifs, poésie et théâtre étant peu représentés.

Le critère de la distinction vers/prose s'avère peu pertinent pour la BFM. En effet, jusqu'au 12e siècle, la plupart des textes, y compris narratifs, sont en vers. A l'inverse, en moyen français, les textes narratifs sont en prose, mais les textes poétiques étant absents de la base pour cette période, la distinction n'est pas plus significative que pour la période précédente. Pour ce qui est enfin de la diversité dialectale, celle-ci n'a pas été un critère de sélection des textes. En effet, au démarrage de la BFM, les linguistes susceptibles de l'utiliser menaient principalement des études syntaxiques, or c'est un domaine où la variation dialectale est relativement peu importante. Les textes sont donc majoritairement en anglo-normand, mais d'autres dialectes y sont néanmoins représentés : champenois, orléanais, et picard en particulier.

Au regard de ses caractéristiques tant quantitatives que qualitatives, on peut dire que la BFM est représentative, non pas de « la » langue de telle période, mais d'un certain état de langue qui correspond aux textes qui la constituent, c'est-à-dire la langue littéraire écrite et narrative des 11e-13e siècles, et des 14e-15e siècles dans une moindre mesure[44].

---

41. De 1994 à 1998 : au sein de l'Equipe d'accueil ELI de l'ENS de Fontenay / St-Cloud dans un premier temps, puis de l'UMR 8503 « Analyses de Corpus » (désormais intégrée à l'UMR 5191 ICAR, ENS-LSH Lyon).

42. Et je tiens à en remercier les équipes successives qui ont géré cette base, et plus particulièrement C. Marchello-Nizia.

43. Voir note 40.

44. Les textes de la première période (9e-10e siècles), trop peu nombreux pour être jugés représentatifs, n'en demeurent pas moins un précieux témoignage de l'époque.

Notons enfin que la BFM, encore privée pour des raisons juridiques, mais devant prochainement intégrer Frantext, a été enrichie d'un balisage normalisé au format XML selon les recommandations de la TEI (Text Encoding Initiative)[45].

4.4. Les corpus numérisés : des apports considérables pour le français médiéval

4.4.1. *Découvrir, compléter et corriger les faits*

La numérisation des corpus a permis de grandement faciliter leur traitement, et elle a autorisé le linguiste médiéviste à brasser une quantité et une diversité de données inenvisageables auparavant. C'est capital pour une langue ancienne, car cela permet d'en modifier notablement notre connaissance sur certains points, qu'il s'agisse de la découverte de certains faits, de leur enrichissement ou même de leur rectification.

Nous illustrerons ce propos par un apport assez récent de C. Marchello-Nizia à la nature de l'opposition sémantique entre les deux séries de démonstratifs (« cist » et « cil ») en ancien français[46]. Pendant longtemps ont prévalu sur cette question des interprétations localistes (proximité / éloignement, discours / récit, etc.), qui laissaient néanmoins trop d'exemples inexpliqués, et ne s'avéraient donc pas pleinement satisfaisantes. Les différents travaux de G. Kleiber (en particulier 1987) ont eu le mérite de renouveler l'approche de cette question. Tout en restant de type localiste, son approche a déplacé la question de la sémantique référentielle vers la sémantique grammaticale, en proposant de centrer la recherche du référent sur l'entourage même de l'occurrence du démonstratif, symbole indexical opaque. Pour « cist », forme à « appariement référentiel contigu saturé », la recherche du référent doit se faire dans le voisinage même (linguistique ou extralinguistique) de la forme. Quant à la forme « cil », elle ne s'oppose pas directement à la précédente, mais correspond à « une forme non-marquée du point de vue de l'obligation de saturation contiguë » (Kleiber, 1987 : 20). Sans remettre totalement en cause cette analyse,

45. Pour une présentation détaillée des modalités d'encodage de la BFM, voir Heiden et Guillot (2003) et Heiden et Lavrentiev (2004).

46. Voir Marchello-Nizia (à paraître). Rappelons qu'il existe deux séries, celle dite en –ST (*cist, cest, ceste, cestui, icestui…*) et celle dite en –L (*cil, cel, cele, celui, icelui…*). En ancien français l'opposition entre les deux est d'ordre sémantico-pragmatique, et non pas morphologique comme en français moderne (déterminant *vs* pronom). L'évolution vers le système moderne, entamée au 13e siècle, se poursuit et s'achève en moyen français. Pour l'étude détaillée de cette question, voir Marchello-Nizia (1995 : chap. 4 et 5).

Marchello-Nizia a montré qu'il était nécessaire de la compléter, voire de la réviser pour la première partie de l'ancien français.

S'appuyant sur une observation de l'ensemble des textes de la Base de Français Médiéval, et sur l'analyse détaillée d'un texte, (*Ami et Amile*, chanson de geste de la fin du 12e siècle, manuscrit de la fin du 13e), C. Marchello-Nizia a constaté que dans certains contextes, et pour la période la plus ancienne (9e-12e), « cil » ne remplace jamais « cist », ce qui met donc en cause son caractère de forme non marquée. Il apparaît en revanche que les occurrences de « cist » se caractérisent par la présence d'un trait sémantico-pragmatique spécifique : l'appartenance du référent à la « sphère du locuteur ». En revanche, lorsque « cil » apparaît là où, selon la seule hypothèse de l'appariement référentiel contigu saturé, « cist » serait possible, on constate que le locuteur exclut le référent de sa sphère personnelle, parfois de manière agressive. Le corpus révèle par la suite, dans le courant du 12e siècle, la coexistence des deux systèmes, avant que celui proposé par Kleiber ne s'impose au siècle suivant.

L'étude d'un gros corpus, en massifiant les occurrences « récalcitrantes » à l'analyse (et plus rares que les autres) a permis d'en dégager les caractéristiques, et, de corriger l'interprétation initiale, principalement en affinant la chronologie des faits. Cette remarque nous conduit à envisager la quantification des faits.

4.4.2. *Quantifier les faits*

Si l'introspection permet parfois au locuteur moderne d'évaluer la fréquence plus ou moins grande d'une construction qui lui est familière, il est toujours à la merci d'une interprétation erronnée, et incapable d'en proposer une quantification exacte, ce que permettent au contraire les corpus, *a fortiori* numérisés. Cet apport majeur est d'autant plus essentiel lorsqu'il s'agit d'un état de langue pour lequel nous ne pouvons de toute façon faire appel à l'introspection.

Déterminer la fréquence exacte d'une construction permet de ne pas s'en tenir à sa simple attestation, et de mesurer la possible évolution de sa productivité. Une telle approche rend par ailleurs possible l'affinement de la notion de « construction dominante », en introduisant celle de « construction majoritaire »[47], distinction qui conduit à une description bien plus précise de nombreux phénomènes. C'est le cas

47. Sur cette question et celles développées dans ce paragraphe, voir Marchello-Nizia (1995 : 24-26).

pour ceux touchant à l'ordre des mots, mais aussi pour ceux relevant du processus de grammaticalisation, en particulier en ce qui concerne les mécanismes de réanalyse et d'extension.

A l'opposé, quantifier les faits sert aussi au repérage des basses, voire très basses fréquences. Signalons d'ailleurs que ces constructions peuvent facilement passer à la trappe – dans des corpus *de facto* trop peu représentatifs –, et être du coup considérées comme non attestées. Le dénombrement de ces basses fréquences permet en particulier de donner plus de consistance à la notion de forme ou de construction « marquée », en dépassant le simple label « minoritaire ».

Il est enfin un point qu'il convient d'évoquer en relation avec la quantification des phénomènes. Il s'agit de la variation morphosyntaxique, que l'on rencontre en particulier en moyen français. Celle-ci se manifeste entre textes répondant à des critères externes différents (en particulier la date et le dialecte), mais aussi entre textes répondant à de mêmes critères externes. On rencontre par exemple, au sein de la BFM, pour la 4e personne de l'imparfait du verbe « avoir », les formes suivantes : *aviiens, avïens, aviens, aviions, avyons*, et pour l'adverbe d'intensité « moult » : *molt, mult, mout, moult*[48]. Mais, plus déroutant encore pour nous, cette hétérogénéité se rencontre aussi au sein d'un même texte, comme en témoignent les exemples proposés plus haut (4.3.4). Il convient donc de dresser préalablement à toute recherche l'inventaire des éventuelles variantes, car l' « oubli » de l'une d'entre elles peut sérieusement déformer les résultats. L'accès à des textes lemmatisés est de ce point de vue fort précieux.

4.4.3. *Corréler les faits*

Outre l'attestation et la quantification de faits isolés, le recours aux corpus, par la quantité des données brassées, permet d'établir des corrélations entre différents faits linguistiques[49] et de mettre ainsi au jour des macroévolutions. Nous en citerons quelques exemples, tout en renvoyant, pour leur développement, aux études qui leur ont été consacrées.

---

48. Nous n'entrons pas dans le détail de cette question : sur le plan linguistique, les variations évoquées ici relèvent assurément de plans différents.

49. Il existe des outils pour mettre en évidence les corrélations entre faits langagiers, en particulier l'analyse multidimensionnelle en statistique, mais c'est, à notre connaissance, une démarche encore peu développée pour les langues anciennes, au moins le français médiéval.

Le premier concerne à nouveau les démonstratifs. Alors que les deux séries – « cist » et « cil » – se distinguaient en ancien français sur le plan sémantique, en moyen français la nature de leur opposition se transforme pour devenir morphologique. Une telle évolution est interprétable comme un mouvement du subjectif vers l'objectif[50]. En effet, que l'opposition sémantique soit conçue en termes de présence / absence d'une référence explicite au contexte immédiat ou à la situation d'énonciation, ou bien en termes d'appartenance à / exclusion de la sphère du locuteur, il y a dans les deux cas une dimension subjective qui se perdra dans le passage à l'opposition morphologique. Il faut d'ailleurs noter que, s'il est bien vrai que l'opposition entre les démonstratifs se serait d'abord construite autour du pôle-locuteur (voir analyse de Marchello-Nizia en 4.4.1), puis autour du pôle-occurrence (voir analyse de Kleiber), il y aurait eu, au sein même du système sémantique, une première évolution du subjectif vers l'objectif. Le point intéressant est que ce mouvement est à mettre en relation avec celui de deux autres formes. La première est le verbe *cuidier* (« croire »), qui, assez fréquent en ancien français, va disparaître entre le 15e et le 17e siècle. Or ce verbe dénote une implication forte du locuteur-énonciateur, qui se pose comme le détenteur de la vérité ou de la fausseté de ce qu'il asserte. Dans le second cas, il s'agit de l'apparition de l'adjectif « reel » (tout d'abord dans des textes juridiques), qui va désormais prendre en charge la « réalité », c'est-à-dire la vérité assertée comme objective, alors que celle-ci était jusque là dénotée par l'adjectif *verai/voir*, celui-là même qui exprime l'idée de vérité assumée comme subjective : la séparation est désormais nette entre objectif et subjectif. C. Marchello-Nizia (Marchello-Nizia, 1997) propose de voir dans ces trois changements une présence amoindrie du locuteur-énonciateur-sujet, et, donc, une objectivation de la langue, évolution de nature métasémantique. A cette dernière s'ajoute, pour les démonstratifs, une évolution de nature métamorphologique.

La transformation que connaît leur système s'inscrit en effet dans un mouvement plus large qui prend place entre le 12e et le 15e siècle et qui consiste à distinguer des éléments de nature et de rôle différents. Se développe ainsi pour les démonstratifs une opposition entre déterminants et pronoms, qui va toucher d'autres formes, en particulier les indéfinis : en moyen français émerge la forme « chaque » qui va prendre en charge les emplois de déterminant jusque-là assumés par « chacun », désormais dévolu aux seuls emplois pronominaux. Sans doute par analogie, « quelqu'un » connaît une évolution similaire, avec l'émergence de la forme « quelque ».

50. Voir Marchello-Nizia (1997) pour la corrélation des trois phénomènes qui vont être évoqués.

C'est un mouvement du même ordre qui est à l'origine des destins croisés de trois morphèmes : « moult », « très », et « beaucoup »[51]. Alors que le premier pouvait porter sur des catégories de niveaux différents, il va progressivement être supplanté par les deux autres (« beaucoup » résultant d'une grammaticalisation, et « très », ancien préfixe verbal, d'une réanalyse), chacun d'eux connaissant, à l'inverse de « moult », une spécialisation syntaxique : « beaucoup » va porter sur des éléments de niveau supérieur (nom et verbe), tandis que « très » va au contraire affecter ceux de niveau inférieur (adjectif et adverbe).

On peut citer, à titre d'illustration supplémentaire de ce vaste mouvement, le fait que de nombreux éléments multifonctionnels en ancien français connaissent en moyen français une spécialisation progressive de leurs emplois, soit comme préposition, soit comme adverbe, soit encore comme préfixe verbal.

Il ne faut certes pas non plus surestimer le rôle de l'exploitation des corpus dans la mise au jour de ces phénomènes : beaucoup, dans leur caractère général, étaient connus, et bien que l'observateur d'une langue soit souvent peu sensible au repérage des faits de corrélation, l'accumulation des études sur la langue ancienne avait permis d'en remarquer certaines. Et si ces dernières n'étaient pas formulées en termes de « macroévolution », ce n'est pas tant le recours aux corpus qui a changé les choses qu'une approche théorique différente. Il n'en demeure pas moins que l'accès à des données massives et diversifiées, ainsi que la facilitation de leur traitement, a permis, d'une part de quantifier et comparer les évolutions convergentes, et donc d'affiner leur description (c'est le cas pour « moult / très / beaucoup »), et d'autre part de donner plus d'assise aux faits assertés en établissant des chronologies assez précises.

4.4.4. *Utiliser des corpus enrichis*

Si la mise au jour de macroévolutions correspond typiquement à celle d'évolutions en profondeur, le recours à des corpus étiquetés est aussi, d'un autre point de vue, une manière de dépasser ce qui se donne à voir de prime abord, et d'accéder à une autre « couche » de description : on ne raisonne plus simplement sur des formes en surface, mais sur des catégories morphologiques, des fonctions syntaxiques, ou des étiquettes sémantiques. Les deux démarches – mise au jour de macroévolutions et travail sur des formes enrichies – sont d'ailleurs souvent

51. Voir Marchello-Nizia (2000) et (2003).

complémentaires, le fait de raisonner sur des catégories, quelles qu'elles soient, permettant de déceler des faits non seulement inédits, mais aussi de les mettre en relation. Un exemple très simple permettra d'illustrer le propos. La BFM renferme cinq textes étiquetés morphologiquement[52], deux d'ancien français (13e siècle[53]), et trois de moyen français (début du 15e pour l'un, fin du 15e pour les deux autres[54]). La simple observation de la fréquence des différentes catégories montre une évolution intéressante : en ancien français, la catégorie la plus représentée est le verbe conjugué, alors que c'est le nom commun à partir du 15e siècle, y compris dans les deux textes d'un genre identique (roman/nouvelle) à ceux de la période précédente. Il ne faut certes pas tirer de conclusions trop hâtives à partir de cinq textes, en outre tous littéraires. Le phénomène a pourtant été pareillement constaté dans le cadre d'un étiquetage par apprentissage. Appliqué sur un texte d'ancien français (*La Queste del Saint Graal*), l'étiqueteur a proposé comme première règle d'étiquetage des formes inconnues : « tout mot contenant le caractère « e » est à étiqueter "verbe conjugué" ». Le même étiqueteur, entraîné sur un corpus de français moderne (mêlant des genres différents), a fourni, comme même première règle : « tout mot contenant le caractère « e » est à étiqueter "nom commun" ».

Si une démarche aussi simple que le dénombrement de catégories permet d'obtenir des informations aussi intéressantes que celle-ci, il est légitime de penser qu'une exploitation plus fine permettra des découvertes d'importance. Mais alors que le recours aux corpus enrichis est désormais un acquis pour le français moderne, c'est encore loin d'être le cas pour le français médiéval. Cela s'explique principalement par l'absence d'étiqueteur « prêt à l'emploi » (c'est-à-dire incluant grammaire et dictionnaire) conçu pour les différents états de langue de cette époque. Et l'on comprend la difficulté de la chose, la langue étant sujette à évolution, et connaissant, dans ses phases de relative stabilité une variation morphologique et syntaxique (ordre des mots beaucoup plus souple qu'aujourd'hui) qui complexifie fort l'élaboration de règles. D'ailleurs, la seule conception d'un jeu d'étiquettes consensuel n'est déjà pas une mince affaire, ne serait-ce que parce que certaines catégories apparaissent tandis que d'autres disparaissent[55].

---

52. Pour une présentation du jeu de 58 étiquettes, voir <http : //weblex.ens-lsh.fr/projects/bfm/jeu_d_etiquettes>.
53. Il s'agit des romans *La Mort Artu* et *la Queste del Saint Graal*.
54. Respectivement *Les quinze joyes de mariage*, recueil de nouvelles, et *Le roman de Jehan de Paris* et le livre 1 des *Mémoires* de Commynes.
55. Pour un développement de ces différents points en corrélation avec la démarche menée au sein de la BFM, voir Prévost et Heiden (2002).

Il est certes possible d'envisager des procédures par apprentissage[56], mais cela suppose, d'une part d'être déjà en possession d'un ou plusieur(s) texte(s) étiqueté(s), et d'autre part de procéder à des vérifications systématiques, la variation étant susceptible de générer davantage d'erreurs que pour le français moderne.

4.4.5. *Il y a corpus et corpus…*

Après cette longue apologie des corpus « gros et variés », il convient de rappeler que, pour l'étude de certains phénomènes, il n'est pas possible, pour des raisons matérielles, d'accumuler les textes exploités.

Le repérage des faits constitue un premier facteur discriminant. Le coût n'est en effet pas le même selon que l'on travaille sur des formes bien circonscrites et dont le relevé peut se faire de manière automatisée, ou à l'inverse sur des constructions qui n'autorisent pas une telle démarche. Il n'est pas égal de collecter, dans des textes non étiquetés, les occurrences de sujets pronominaux ou celles de sujets nominaux. A cela s'ajoute la fréquence des occurrences : leur traitement sera plus rapide si elles sont rares que si elles sont nombreuses. Outre la taille du corpus de départ, il faut donc considérer celle du « sous-corpus » que constituent les faits pertinents.

Mais c'est surtout la nature même de l'étude qui est décisive. Etablir la typologie des emplois d'une forme n'est certes pas simple, *a fortiori* si les contextes d'occurrence sont ambigus. Toutefois, analyser une structure complexe, par exemple sur un plan sémantico-pragmatique, nous semble présenter un degré supplémentaitre de complexité. L'expérience nous a confrontée à ces différentes démarches.

4.4.5.1. *Gros corpus ...*

Nous avons ainsi étudié l'évolution de différentes formes, de l'ancien français au 16e siècle, en travaillant sur de gros corpus[57]. Il s'agit de l'adverbe « aussi »[58] en position initiale, du marqueur de topicalisation « quant à X »[59] et du trinôme « les

56. Voir Prévost et Heiden (2002) et Stein (2003).
57. La Base de Français Médiéval pour l'ancien français, celle du DMF pour le moyen français, et celle de Frantext pour le 16e.
58. Prévost (1999).
59. Prévost (2003).

aucuns » / « d'aucun(s) » / « aucun(s) »)[60]. Pour les deux premiers, une fois sélectionnées les seules formes en position initiale, le total des données restait inférieur au millier (377 occurrences provenant de 57 textes pour « quant à », 696 occurrences émanant de 49 textes pour « aussi »). Le nombre était en revanche bien plus élevé pour /aucun/ : 10 444 occurrences pour la seule période médiévale (16e siècle inclus), et plus de 30 000 pour les siècles suivants.

Ceci est une illustration typique de la différence qu'il peut y avoir, à partir d'un même corpus, entre les sous-corpus que constituent les faits pertinents. Le traitement des données n'a donc pas été le même. Les caractéristiques morphosyntaxiques ont été dégagées dans les trois cas, de manière assez précise pour /aucun/, pour lequel l'un des enjeux principaux était précisément de rendre compte de l'évolution de la distribution entre déterminants et pronoms[61]. La dimension sémantique de l'étude s'est en revanche déclinée selon des modalités différentes. Pour « aussi » et « quant à », nous avons pu décrire en détails les caractéristiques sémantico-pragmatiques de l'ensemble des occurrences, et mettre au jour des tendances assez nettes et précises ainsi que leur évolution, que seul le recours à un corpus conséquent a permis de dégager, les occurrences demeurant dans l'ensemble peu fréquentes. Pour /aucun/, il n'était pas possible d'envisager une étude sémantico-référentielle détaillée sur l'ensemble du corpus de travail, qui a donc été réduit, soit par le choix des formes mêmes (« les aucuns » et » d'aucuns » suscitant une description plus minutieuse que « aucun » pour le français médiéval), soit par celui de leurs caractéristiques morphologiques (formes plurielles et pronominales par exemple), soit encore en opérant une sélection générique parmi les textes. L'étude exige donc d'être complétée par une analyse sémantique détaillée des formes laissées de côté. Toutefois, la démarche adoptée (qui a fait émerger des évolutions que nous pensons non exclusives des occurrences considérées) nous semble être, face à une telle avalanche de données, un compromis acceptable, qui allie analyse de détail et étude plus grossière, prise en compte exhaustive des données et sélection parmi elles.

Il ne s'agit pas de décider du seuil au-delà duquel la gestion exhaustive du corpus et de ses données n'est plus possible : un tel seuil n'existe pas de manière absolue, tant il dépend de facteurs divers (nature de l'analyse, degré de finesse de

60. Ici abrégé par commodité en /aucun/. Voir Prévost et Schnedecker (à paraître). L'étude a porté aussi sur les siècles suivants (17e-20e), cette période étant plus spécifiquement prise en charge par C. Schnedecker.

61. Et nous pensons avoir mis au jour de ce point de vue des phénomènes assez inattendus.

celle-ci ?, etc.). Il n'en est pas moins utile de garder à l'esprit les contraintes qu'impose le « quantitatif »[62].

4.4.5.2. *...et petits corpus*

A l'inverse de la démarche précédemment évoquée, nous avons mené deux études portant sur des corpus bien plus restreints. Il s'agit, d'une part de l'évolution de la postposition du sujet du 15e au 16e siècle, réalisée dans un cadre sémantico-pragmatique[63], et, d'autre part, de l'étude des phénomènes d'encadrement temporel en relation avec l'expression du sujet dans un corpus du 15e siècle (éclairé par un texte du 13e)[64]. En ce qui concerne les corpus, il s'agit de six textes littéraires dans le premier cas, dont cinq narratifs et un historique (le premier livre des *Mémoires* de Commynes), et de cinq textes pareillement littéraires (dont le même livre 1 de Commynes) dans le second. Si les corpus constitués avaient comme vocation première de témoigner des seuls textes qui les composaient, cela n'exclut pas de poser la question de leur représentativité et de la possible généralisation des résultats dégagés, en particulier en ce qui concerne la première étude.

Rappelons qu'il est désormais admis que l'ancien français est une langue qui, tout en obéissant à la contrainte assez rigide du verbe en seconde position, connaît un ordre des mots organisé selon un principe, non pas grammatical comme en français moderne, mais pragmatico-informationnel, formulé en termes de « thème-rhème » ou de « topique-commentaire » selon les approches. C'est ce qui explique la souplesse de l'ordre des mots, et en particulier la position variable du sujet. La transition vers le système moderne s'opère en moyen français, et les cas d'inversion du sujet qui se maintiennent à cette époque sont généralement analysés comme un vestige de cette organisation informationnelle. Or nous avons précisément montré que ce n'est pas le cas, et qu'il convient, pour rendre compte du maintien de l'ordre verbe-sujet, de dépasser la dimension strictement informationnelle pour une approche sémantico-pragmatique plus large. Ces résultats infirment donc partiellement les interprétations précédentes. Mais cette invalidation ne vaut-elle que pour les six textes de notre corpus, ou peut-on la généraliser ?

62. Certains outils informatiques nous aideraient beaucoup, par exemple en indiquant le vocabulaire lexical ou grammatical associé à la forme étudiée : l'analyse en aval en serait grandement allégée !
63. Prévost (2001).
64. Prévost (à paraître).

Nous terminerons par une question qui touche aussi aux procédures de généralisation. Qu'il s'agisse de l'évolution sémantico-pragmatique de la postposition du sujet ou de celle de l'encadrement temporel corrélée à l'expression du sujet, l'étude a dans les deux cas mis au jour des fonctionnements parfois très variables entre textes relevant pourtant de mêmes critères externes. Deux textes de la même époque et/ou du même genre peuvent ainsi différer sur certains points (comme la structure informationnelle dominante ou la catégorie prévalente d'adverbial temporel), et s'apparenter au contraire pour ces mêmes points à des textes de date et/ou de genre différent(s). Il faut insister sur le fait que les corrélations établies ne sont pas fixes, mais associées à tel ou tel trait. Le double-constat, affinités inattendues et corrélations à géométrie variable, laisse quelque peu perplexe.

On sait que la prudence est de mise dès qu'il s'agit de généraliser les résultats, mais l'on pourrait penser que des textes présentant de mêmes critères externes s'y prêtent plus facilement. Or pour le moyen français (au moins pour les textes et pour les phénomènes que nous avons observés), cette démarche ne va pas de soi, tant les tendances stylistiques des auteurs, qui touchent parfois à l'idiolecte, peuvent « typer » les textes.

Cela prouve que le recours aux gros corpus et l'automatisation de certains traitements ne doit pas exclure, d'une part une attention constante à la spécificité éventuelle des résultats de chaque texte, et d'autre part une « descente » fréquente dans le détail des textes, cela sous peine de voir écrasées certaines oppositions, neutralisées certaines variations… qui font précisément toute la richesse de la langue : la quantité des données ne doit pas se substituer à la qualité de leur analyse.

4.5. En guise de conclusion …

Le recours aux gros corpus, *a fortiori* enrichis, représente un considérable apport quantitatif et qualitatif pour notre représentation du français médiéval, mais il a en outre permis un « déplacement qualitatif dans le mode de raisonnement lui-même » (Marchello-Nizia, 1999 : 39). Idéalement, la connaissance d'une langue devrait s'appuyer sur le recours conjugué à l'introspection et aux corpus. La première étant exclue pour cette période, on peut espérer que le développement croissant des bases textuelles va permettre de « compenser » autant que possible l'absence de locuteur natif, et d'élaborer progressivement une nouvelle grammaire, l'exploitation d'un corpus permettant de confirmer ou d'infirmer les règles existantes, pour ensuite les

confronter à de nouveaux textes et éventuellement les corriger ou les affiner. Ce mouvement de va-et-vient entre règles et corpus permettra peut-être, à long terme, de doter la grammaire ainsi élaborée d'un caractère prédictif, sinon de la « vraie » langue de l'époque, au moins des différents états qu'elle nous donne à voir à travers l'ensemble des documents qui nous sont parvenus, et dont le traitement aussi exhaustif et réfléchi que possible constitue donc un enjeu capital.

4.6. Bibliographie

Biber, D. (1988). *Variation across speech and writing*. Cambridge : Cambridge University Press.

Biber, D. (1990). Methodological issues regarding corpus-based analyses of linguistic variation. *Literary and Linguistic Computing, 5* (4), 257-270.

Biber, D. (1993). Using register-diversified corpora for general language studies. *Computationnal linguistics,* 19 (2), 219-241.

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

Biber, D., Conrad, S. & Reppen, R. (1998). *Corpus Linguistics. Investigating Language Structure and Use.* Cambridge : Cambridge University Press.

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

Cerquiglini, B. (1991). *La naissance du français.* Paris : PUF, collection « Que sais-je ? ».

Condamines, A. (2000). Les bases théoriques du groupe toulousain 'Sémantique et Corpus' : ancrages et perpsectives. *Cahiers de Grammaire, 25*, 5-28.

Foulet, L. (1963). *Petite syntaxe de l'ancien français*. Paris : H. Champion. (Première publication 1919).

Guillot, C. (2003). *Le rôle du démonstratif dans la cohésion textuelle au 15e siècle* Eléments de grammaire textuelle. Thèse de doctorat. Lyon : Ecole normale supérieure Lettres et Sciences humaines.

Habert, B & Zweigenbaum, P. (2002). Régler les règles. *TAL,* 43(3), 83-105.

Habert, B. (2000). Des corpus représentatifs : de quoi, pour quoi, comment ?. In M. Bilger (Ed) , *Cahiers de l'université de Perpignan, 31,* '*Linguistiques sur corpus. Etudes et réflexions'* (pp. 11-58). Perpignan : Presses universitaires de Perpignan.

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

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. Kunstman, F. Martineau & D. Forget (Eds.), *Ancien et moyen français sur le web. Enjeux méthodologiques et analyse du discours* (pp. 77-92). Ottawa : Les Editions David.

Heiden, S. & Lavrentiev, A. (2004). Ressources électroniques pour l'étude des textes médiévaux : approches et outils. *Revue Française de Linguistique Appliquée*, *IX,* (1), 99-118.

Kennedy G. (1998). *An introduction to corpus linguistics*. London : Longman, Studies in language and linguistics.

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

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

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

Marchello-Nizia, C. (1997). *La langue française aux 14e et 15e siècles*. Paris : Nathan. (Première publication 1979 Bordas).

Marchello-Nizia, C. (1999). Corpus diachroniques. *Revue française de linguistique appliquée*, IV(1), 31-39.

Marchello-Nizia, C. (2000). Les grammaticalisations ont-elles une cause. *L'information grammaticale, 67*, 3-9.

Marchello-Nizia, C. (2003). Changes in the structure of grammatical systems : the evolution of French. *Court of the University of St Andrews, XXXIX*, (4) , 372-385.

Marchello-Nizia, C. (à paraître). « Se vos de ceste ne voz poéz oster, Je voz ferai celle teste coper » (*Ami et Amile 753*) : la sphère du locuteur et la deixis en ancien français. In *Mélanges Venckeler*.

Martin, R. & Wilmet, M. (1980). *Syntaxe du moyen français*. Bordeaux : Sobodi.

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

Perret, M. (1998). *Introduction à l'histoire de la langue française*. Paris : Sedes.

Pery-Woodley, M.-P. (1995). Quels corpus pour quels traitements automatiques ?. *TAL,* 36(1-2), 213-232.

Prévost, S. & Heiden, S. (2002). Etiquetage d'un corpus de français médiéval : enjeux et modalités. In C.D. Pusch & W. Raible (Eds.), *Romance Corpus Linguistics - Corpora and Spoken Language* (pp.127-136). Tübingen : Gunter Narr Verlag Tübingen.

Prévost, S. &. Schnedecker C. (sous presse). *Aucun(e)(s) /d'aucun(e)(s) / les aucun(e)(s) :* évolution du français médiéval au français moderne. *Scolia,18*, 39-73.

Prévost, S. (1999). *Aussi* en position initiale : évolution sémantico-syntaxique du 12e au 16e siècle. *Verbum, XXI*, (3), 351-380.

Prévost, S. (2001). *La postpostion du sujet aux 15e et 16e siècles : une approche pragmatico-sémantique*. Paris : éditions du CNRS.

Prévost, S. (2003). *Quant a :* analyse pragmatique de l'évolution diachronique (14e-16e siècles). In B. Combettes, A. Theissen & C. Schnedecker (Eds.), *Ordre et distinction dans la langue et le discours* (pp. 443-459). Paris : H. Champion.

Prévost, S. (à paraître). Encadrement temporel en français médiéval. *Revue Québéquoise de Linguistique*.

Sinclair, J. (1996). *Preliminary recommendations on Coprus Typology*. Technical report, EAGLES, (Expert Advisory Group on Language Engineering Standards). <http : //www.ilc.cnr.it/EAGLES96/pub/eagles/corpora/corpustyp.ps.gz>

Stein, A. (2003) Etiquetage morphologique et lemmatisation de textes d'ancien français. In P. Kunstman, F. Martineau & D. Forget (Eds.), *Ancien et moyen français sur le web. Enjeux méthodologiques et analyse du discours* (pp. 273-284). Ottawa : Les Editions David.