Diachronie du français et linguistique de corpus : une approche quantitative renouvelée

Sophie Prévost

Diachronie du français et linguistique de corpus : une approche quantitative renouvelée

Sophie Prévost

Lattice UMR 8094 - CNRS/ENS/Paris3-Sorbonne Nouvelle

Résumé

Le présent article aborde la question de la fréquence textuelle du point de vue des études menées sur des états de langue ancien et dans une perspective diachronique. L'appui sur des données attestées, et sur leur quantification, a toujours été indispensable pour les linguistes travaillant sur une langue sans locuteurs et visant à rendre compte des changements. La numérisation des textes et le développement conjoint d'outils de traitement a cependant provoqué un important bouleversement méthodologique, du fait de la massification des données désormais traitées. Les gains sont nombreux, en particulier pour l'établissement de chronologies affinées et le repérage des lieux du changement. La présentation de quelques aspects de l'évolution de la syntaxe du sujet pronominal permet d'illustrer les apports d'une quantification à grande échelle, et de pointer certaines limites de cette approche.

Abstract

The present study deals with textual frequencies, considered from the point of view of former states of the language, and in a diachronic perspective. Linguists working on a language without speakers and aiming to account for linguistic changes have always relied on attested data and on their quantification. But electronic databases, and the text mining software developed to process them, have brought on major methodological changes to the field, due to the massification of the data henceforth dealt with. This results in many benefits, especially when the task is to refine the chronology and pinpoint the *locus* of specific language changes. In order to illustrate the benefits of a massive quantification as well as the limits of such an approach, I present some aspects of the evolution of the pronominal subject in French.

Mots-clés : français médiéval ; diachronie ; mesures et quantification ; lecture paradigmatique ; syntaxe du sujet pronominal.

Key-words : Medieval French ; diachrony ; measures and quantification ; paradigmatic reading ; subject pronoun syntax.

Introduction

Travailler sur une langue sans locuteurs rend les textes indispensables : les données attestées constituent en effet le seul objet possible d'investigation, et l'établissement de fréquences permet d'affiner des descriptions sinon approximatives. Rendre compte des changements – qu'il s'agisse de l'émergence, de la disparition ou de la transformation d'une construction[1], d'un paradigme – suppose par ailleurs de quantifier les phénomènes observés. Les diachroniciens médiévistes ont donc toujours accordé une place privilégiée

[1] La notion de 'construction' est entendue ici dans un sens large (et non dans le cadre des grammaires de construction) : une construction peut être simple ou complexe, de nature morpho-syntaxique, syntaxique ou sémantique.

à la quantification des données attestées. La numérisation des textes et la linguistique de corpus ont cependant largement modifié leur méthodologie, en raison, principalement, de la massification des données désormais prises en compte. Nous envisagerons dans un premier temps les spécificités du travail sur une langue sans locuteurs, puis le renouveau méthodologique produit par la numérisation des textes et la linguistique de corpus. Nous montrerons ensuite combien la quantification des données est, à différents égards, constitutive des études diachroniques, ce que nous illustrerons par quelques aspects de l'évolution de la syntaxe du sujet pronominal, ce qui nous permettra de pointer certaines limites de cette approche.

1. Travailler sur une langue sans locuteurs

1.1. Quand le possible de langue rejoint le possible attesté

Dans *Introduction à une science du langage* (1989), Milner considère l'établissement du « jugement différentiel » comme constitutif de l'activité grammaticale :

> Le principe de ce jugement différentiel est que tout ne peut pas se dire. [...] Il suppose donc qu'il y a un impossible en langue. Toutefois cet impossible de langue n'est pas un impossible matériel. Autrement dit, une donnée de langue peut être possible matériellement, c'est-à-dire attestée, et impossible en langue, ou inversement (1989 : 57).

Ces deux questions – établissement du jugement différentiel et possibilité d'une grammaire – se posent différemment pour les états anciens de la langue. Ceux-ci ont en effet la caractéristique majeure d'être dépourvus de locuteurs vivants, et notre absence de compétence exclut tout recours à l'introspection. Par conséquent, possible en langue et possible matériel tendent à être coextensifs, dans les faits et par principe. En effet, ce qui nous permet d'établir le possible en langue d'une donnée langagière est précisément le fait qu'elle est attestée, observable, les textes existants constituant notre seul outil d'appréhension de la langue (sur ce point, voir ci-dessous 1.2.).

On peut néanmoins interpréter une donnée attestée comme impossible en langue, mais cette appréciation ne résulte pas d'une introspection fondée sur notre compétence : elle repose sur ce que Marchello-Nizia (1995) a appelé une « intuition de reconnaissance »[2]. Il s'agit d'un jugement permis par la construction progressive de notre connaissance de la langue ancienne, qui s'est faite par l'étude des grammaires et par la fréquentation des textes.

Notre aptitude à émettre un jugement sur le caractère possible ou impossible d'une donnée langagière provient donc de ce que nous avons *appris* des règles explicitement formulées (ou bien parfois en avons-nous inféré certaines par une fréquentation assidue et une analyse des textes). Les modalités d'acquisition et de mise en œuvre de cette *grammaire* sont évidemment fort différentes de la grammaire intérieure que nous possédons pour notre propre langue. Contrairement à ce qui se produit pour celle-ci où « …nul sujet parlant n'est censé connaître les règles, parce que les ignorer n'empêche pas nécessairement de les appliquer » (Milner 1989 : 94), pour la langue ancienne, il faut connaître les règles pour juger de leur application.

Outre le caractère discutable en soi – car normatif – de la notion d'impossible en langue quand elle s'applique à des données attestées, la situation est plus complexe pour

[2] Marchello-Nizia (1995 : 22) suggère que le linguiste médiéviste peut parfois développer, à défaut d'une compétence de « production », une compétence de « reconnaissance », distinction qui rejoint celle entre compétence active et compétence passive.

la langue ancienne : comment interpréter une donnée attestée qui nous semble être, eu égard à notre « connaissance » de la langue, un impossible grammatical ?

Si la donnée se révèle finalement non isolée, et qu'elle se rencontre dans d'autres textes, plus tardifs, on peut faire l'hypothèse qu'il s'agit de l'amorce d'un changement, et que la « règle » doit être révisée. Un impossible en langue à un moment donné peut par la suite devenir possible en langue : c'est précisément la caractéristique du changement. S'il ne s'agit pas des prémices d'un changement, que l'hapax en est bien un, plusieurs interprétations sont possibles, entre lesquelles il est difficile de trancher : soit il s'agit d'une faute de l'auteur ou du copiste, en tout cas d'un impossible en langue ; soit il s'agit d'un possible en langue, que sa rareté n'a pas permis de mettre au jour jusqu'ici. Il peut aussi s'agir d'un possible en langue propre à celui qui l'a produit : ce ne serait alors pas une faute en tant que telle, mais l'indice que l'auteur de la donnée aurait eu une grammaire différente de celle observée le plus fréquemment. On touche ici à la question des usages et de la variation, et à celle des basses fréquences.

Le jugement d'acceptabilité peut donc s'appliquer, dans une certaine mesure, aux données attestées, même si les « données attestées impossibles en langue » sont rares, et difficiles à établir avec certitude.

L'intuition de reconnaissance peut aussi parfois permettre de se prononcer sur la fréquence plus ou moins élevée d'une construction[3]. Mais son caractère est bien peu fiable, bien moins encore que celui de l'introspection, à propos de laquelle nous rejoignons tout à fait Corbin (1980) :

> …si l'introspection peut repérer certaines variations dans les pratiques langagières, elle est impuissante à décrire leur distribution dans la population : le social lui échappe par définition (1980 : 121).

1.2. Les textes : accès unique et accès restreint à la langue

Constructions attestées ou recours éventuel à une intuition de reconnaissance fondée sur notre connaissance des textes : c'est toujours, *in fine*, à travers les textes qui nous sont parvenus que nous percevons la langue ancienne. Les textes constituant par ailleurs l'objet d'étude, il convient, afin d'échapper à une possible circularité, d'envisager toujours des données nouvelles, de prendre en compte des textes encore peu sollicités : cela permet de confirmer, d'infirmer, de corriger ou de compléter des propriétés précédemment mises au jour, ou d'en dégager de nouvelles.

Toutefois, aussi nombreux et variés que soient les textes considérés, le corpus se donne d'emblée comme « restreint ». En effet, travailler sur des états anciens de la langue exclut par principe l'accès aux registres de l'oral « oralisé » (par opposition aux mises en écrit de l'oral), et, dans la pratique, à certains registres de l'écrit. Nous savons que certains genres ne nous sont parvenus qu'à travers bien peu de textes, d'où la difficulté à caractériser leur « langue ». Nous ignorons par ailleurs fort probablement l'existence de certains. Nous n'accédons par conséquent qu'à un sous-ensemble de la langue des locuteurs de l'époque et n'avons qu'une vision parcellaire de l'état de langue considéré. L'objet étudié est donc partiel dans l'absolu, mais il constitue néanmoins une totalité finie pour nous, puisque les textes susceptibles d'être analysés forment un ensemble délimité. La situation est de ce point de vue opposée à celle dans laquelle on se trouve pour la langue moderne : cette dernière est potentiellement accessible dans sa totalité, mais cette totalité est infinie, puisque de nouvelles données sont sans cesse produites.

[3] Voir sur ce point les réflexions de S. Loiseau, dans l'introduction de ce numéro, sur les notions de fréquence mesurée et de fréquence intuitive (3.1.), ainsi que celles de G. Désagulier dans son article sur le statut de la fréquence dans les grammaires cognitives (2.1.).

Reste cependant, dans les deux cas, la question de la maîtrise de la variation langagière, bien imparfaite encore, presque vingt ans après la remarque de M.-P. Péry-Woodley à propos de ce serait un « corpus équilibré » :

> 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. (Péry-Woodley, 1995 : 218).

Si l'établissement d'un modèle de la variation reste assurément un enjeu de taille, y compris pour la langue ancienne, la perspective d'explorer un jour l'ensemble des textes qui sont parvenus jusqu'à nous constitue un premier pas vers cet objectif.

Dans le cadre d'études ponctuelles, il est rarement envisageable, pour des raisons matérielles (disponibilité des textes d'une part et coût de leur traitement d'autre part) de traiter l'ensemble des données existantes. Devrait néanmoins toujours se poser la question de bien représenter l'état langagier que l'on souhaite observer, c'est-à-dire de bien « panacher » les textes. La question de la constitution des corpus, et donc de la représentativité de la langue étudiée est particulièrement cruciale lorsque l'on travaille sur un état de langue pour lequel notre compétence de locuteur ne peut servir de garde-fou aux résultats obtenus et à leur interprétation (voir par ailleurs en 3.3.2. la question du « maillage » du corpus).

2. Une tradition de travail sur les textes renouvelée par la numérisation des textes et la linguistique de corpus

2.1. De nouveaux modes d'exploration des textes

Les linguistes médiévistes ont toujours pris en compte les fréquences des phénomènes observés, en particulier lorsqu'il s'agit d'étudier des changements. Mais la numérisation des textes et le développement des outils de traitement de ces derniers ont largement modifié la démarche méthodologique. Jusqu'encore récemment, on travaillait sur les données attestées dans les livres, individuellement, et sur un nombre de textes restreint. C'est toujours sur des données attestées qu'on le fait aujourd'hui, mais celles-ci sont désormais souvent disponibles aussi sous forme numérisée, ce qui permet leur massification et un traitement facilité. A certains égards, la méthode n'a pas changé : ainsi le choix des textes composant un corpus était et reste conditionné par la construction ou le phénomène à étudier (on sait parfois que certains textes seront plus pertinents, plus révélateurs que d'autres, ce qui présuppose toutefois une relative connaissance préalable des textes) ainsi que par la qualité de l'édition des textes. On continue à faire des comptages, à rechercher les premières occurrences, à élaborer des raisonnements, avant de retourner aux textes pour vérifier les hypothèses. Mais l'on est bien moins limité par les possibilités humaines individuelles de comptage : les textes numérisés et les outils pour les explorer et les analyser ont permis un gain en temps et en souplesse.

Le gain de temps permet en outre un enchaînement beaucoup plus rapide des raisonnements, et donc, potentiellement, un accroissement de la « productivité scientifique ». On peut aussi, grâce à l'augmentation du nombre de textes traités et grâce aux outils qui les exploitent, opérer des comparaisons assez rapides entre textes : dès lors devient possible la prise en compte plus systématique de la variation dialectale, dans ses différentes instanciations, de même que l'accroissement du nombre de textes étudiés permet l'établissement de diachronies à la fois plus larges et plus fines, les jalons textuels pouvant être multipliés.

Les avantages, considérables, ont néanmoins une contrepartie négative : on lit moins les textes linéairement, voire on ne les lit plus, à la fois parce que cela n'est plus

indispensable et que ce n'est plus possible matériellement. Du coup, si l'activité est plus exploratrice au sens où elle permet de repérer plus facilement une première occurrence, un contexte inédit, elle est en revanche moins « curieuse », du fait qu'elle est davantage « pré-formatée » : en ne lisant plus au fil du texte, on ne repère plus, ou on repère moins, de phénomènes nouveaux, de constructions inédites. On ne lit plus (ou moins) les textes linéairement, mais on les lit autrement : d'une lecture continue et linéaire, on est passé à une « lecture » par classes d'occurrences (quelles qu'elles soient) extraites plus ou moins automatiquement, dans un contexte plus ou moins large ; autrement dit, à une appréhension « syntagmatique » des textes se substitue désormais, souvent, une approche « paradigmatique ». Cette lecture à la fois verticale et transversale résulte du mode d'exploration et de filtrage qui consiste à générer les concordances d'une construction. La linéarité physique du texte est en partie brisée, mais c'est à un autre mode de représentation du texte que l'on accède, par classes. On peut mesurer l'homogénéité de celles-ci, leur variation interne… bien plus facilement que si les différences occurrences constitutives de ces classes étaient appréhendées au fil du texte.
On découvre ainsi des caractéristiques qu'une lecture linéaire n'aurait pas permis de mettre au jour (ou bien difficilement) ; c'est une structuration inédite du texte que l'on construit ainsi, non perceptible à l'œil nu, sans outils.

Ces derniers rendent donc possible, non seulement une massification des données traitées, mais aussi (et c'est en partie lié) une perception différente des constructions langagières et des textes.

2.2. La constance de l'outil informatique *versus* l'inconstance de l'œil humain

Les outils informatiques ont un autre avantage : il s'agit de la régularité, de l'homogénéité de l'analyse effectuée, quelle que soit sa nature (collecte de données, dénombrements, etc.). En effet, lorsque l'on cherche un objet linguistique à l'œil nu, on peut omettre des occurrences[4]. Il se peut aussi que l'on fasse « bouger » la requête, involontairement, en ne tenant plus compte systématiquement, sans même s'en rendre compte, des critères de sélection : on génère alors du bruit et/ou du silence.

L'outil permet par ailleurs de traiter – et donc de dénombrer – des structures très abstraites ou complexes que l'œil humain ne peut percevoir qu'avec beaucoup de difficulté (et de lenteur), à condition d'avoir projeté des catégorisations adéquates.

Le recours aux textes numérisés et à leur traitement informatique peut certes occulter des constructions inédites ou des phénomènes intéressants (voir 2.1.), mais il permet en revanche de percevoir des objets invisibles à l'œil nu, et d'accéder à de nouvelles couches de description.

3. Etude des changements : une nécessaire quantification des données

3.1. Données numérisées et outillées : des quantifications multiples

L'enjeu premier de la linguistique diachronique est la mise au jour (et l'analyse) de l'émergence, de la disparition, et de la transformation de constructions. Rendre compte de ces différents types de changement suppose l'établissement de fréquences, et ce quel que soit l'état de langue considéré. Décrire une langue sans locuteurs implique par ailleurs de s'appuyer sur un *dénombrement* des constructions. Diachronie et langue ancienne : la quantification des données est doublement nécessaire.

[4] A l'inverse, l'œil humain saura reconnaître une forme sous ses différentes variantes (et la variation morphologique est importante en français médiéval), alors que l'outil ne tiendra compte que des variantes qui ont été explicitement spécifiées.

Déterminer la fréquence exacte d'une construction et des contextes dans lesquelles elle apparaît est déterminant dans une perspective diachronique, et présente au moins deux intérêts majeurs. Cela permet d'une part de repérer les basses, voire très basses fréquences, et de ne pas considérer comme non attestées des constructions rares. Les outils de quantification permettent aussi d'apprécier, rapidement, la répartition d'une construction dans les différents textes d'un corpus. C'est un point important, trop souvent négligé, en particulier lorsque l'on travaille sur l'émergence d'une construction (voir ci-dessous en 3.3.2. les remarques sur l'émergence de *quant à* et *à propos de*).

Cela permet d'autre part de mesurer l'évolution de la productivité d'une construction, ce qui est important dans tous les phénomènes de changement, en particulier pour ceux qui relèvent de la grammaticalisation.

3.2. Nature et vitesse des changements

On observe deux positions majeures en ce qui concerne la nature des changements. L'une d'elle consiste à considérer que les changements surviennent brusquement : c'est la conception qui avait été développée dans un premier temps dans le cadre de la grammaire générative : selon Lightfoot (1979), un changement linguistique ne peut être que « catastrophique », dans la mesure où il suppose une *réanalyse*, et que seuls les enfants sont capables d'accomplir une telle opération, en analysant différemment les énoncés entendus[5].

La seconde position, développée dans le cadre de la socio-linguistique et de la grammaticalisation, envisage les changements comme inscrits dans un continuum, avec la coexistence temporaire d'une forme nouvelle et d'une forme ancienne. Les textes nous montrent que, au moins en syntaxe et en sémantique, c'est plutôt selon le second mode, continu, qu'évoluent les constructions langagières.

Les changements étant progressifs, se pose la question de leur vitesse : d'une part la durée globale de chaque changement (une génération, un siècle, …), d'autre part la vitesse plus ou moins grande de la diffusion d'un changement : son rythme. A. Kroch (1989) a ainsi mis au jour, pour certains changements, un schéma d'évolution qu'il a appelé 'courbe en S' ('S-curve') : dans un premier temps les emplois augmentent lentement, gagnant progressivement de nouveaux contextes, puis dans un second temps leur fréquence s'accroît rapidement et pareillement en tous contextes, avant de ralentir, formant ainsi une sorte de palier.

Etablir les temporalités exactes – durée et rythme – n'est pas simple, et suppose de travailler sur un corpus suffisamment représentatif du ou des état(s) de langue dans le(s)quel(s) s'inscrit l'évolution du phénomène.

3.3. Contextes linguistiques et contextes textuels : repérer les lieux du changement

Concrètement, l'une des tâches majeures du diachronicien consiste à repérer les étapes du changement, à travers les textes et les contextes linguistiques.

3.3.1. Contextes linguistiques

Pour ces derniers, il s'agit d'identifier ceux où se produisent les premiers changements, et d'essayer de déterminer pourquoi, en dégageant les traits distinctifs, le cas échéant, desdits contextes. A cet égard, le modèle du changement sémantico-syntaxique proposé par Heine (2002) s'avère fort convaincant : A > Ab > aB > B. Dans ce schéma, A et B peuvent correspondre à des constructions simples ou complexes, mais aussi aux

[5] Lightfoot a par la suite adopté une position moins radicale, tandis que les travaux de Romaine (1989) et de la socio-linguistique ont de leur côté conduit à réviser cette position.

différentes valeurs d'une construction. Durant la première étape, A apparaît seul, puis, seconde étape, B apparaît mais il est minoritaire ; il devient majoritaire dans une troisième étape. Phase ultime : B supplante A, qui peut néanmoins se maintenir. L'évolution du système de la négation est un exemple de ce dernier cas de figure : d'abord utilisé seul, *ne* s'est vu progressivement renforcé par des morphèmes dénotant initialement une petite quantité (*pas*, *point*, *mie*…, le premier s'imposant peu à peu), avant que *pas* ne soit finalement utilisé seul, au moins à l'oral, et ce dès le 17ème siècle[6]. Pour autant, *ne* n'a pas disparu, la locution *ne… pas* restant dominante dans les registres soutenus.

Le modèle de Heine a l'intérêt d'affiner l'étape de transition (entre le stade originel et le stade final) : il distingue plus précisément un contexte de 'transition' (*bridging context*), dans lequel la construction d'une inférence donne lieu à une nouvelle signification qui passe au premier plan, et un contexte de 'bascule' (*switching context*), incompatible avec la signification d'origine, qui passe au second plan.

La mise en œuvre concrète de ce modèle est néanmoins difficile, en particulier pour ce qui concerne les étapes intermédiaires, décisives pour la compréhension du changement. On y observe la coexistence de deux constructions ou de deux interprétations syntaxiques et/ou sémantiques, situation qui est à la source du changement. Mais c'est précisément cette ambiguïté interprétative qui fait difficulté, en particulier lorsque l'on travaille sur une langue dont on n'a pas la compétence. L'exemple de l'émergence de la locution « *aller* + infinitif » comme périphrase temporelle exprimant le futur illustre cette difficulté : s'il est relativement aisé d'interpréter les constructions dans lesquelles la locution ne dénote encore qu'un mouvement, ainsi que celles dans lesquelles elle exprime clairement une idée de futur (ainsi dans *il va pleuvoir* la présence d'un sujet inanimé exclut de comprendre *aller* comme un verbe de mouvement), il est bien plus complexe de donner une interprétation des constructions « intermédiaires » dans lesquelles l'idée de mouvement et celle d'intention (ex. 1), ou bien celle d'intention et celle de futur, sont parfois toute deux envisageables :

(1) Nous **alomes** la messe **oïr**
Tui alomes vers le mostier (*Roman de Renart*, déb. 13ème)[7].

'Nous **allons écouter** la messe, nous allons tous à l'église'

L'étude de l'émergence de certains marqueurs discursifs soulève une difficulté analogue. Par exemple, on sait que *à (ce) propos* a développé son emploi discursif à partir d'un emploi de complément régi, lequel a pu s'émanciper de la rection verbale pour fonctionner comme introducteur de remarque incidente (Prévost 2011c). Les occurrences dans lesquelles *à (ce) propos* est clairement régi ne font pas difficulté, de même que celles dans lesquelles, au contraire, il fonctionne comme marqueur discursif. En revanche, en moyen français, époque à laquelle émerge la valeur discursive, certaines occurrences sont difficiles à interpréter. Ainsi, dans l'exemple (2), une double lecture est possible :

(2) Car entre blanc et noir qui sont couleurs contraires sont pluseurs couleurs moiennes. Et aussi, **a propos**, entre le lieu ou est le feu et le lieu ou est la terre est lieu moien qui ne peust estre vieu. (N. Oresme, 1370)

[6] On en trouve des occurrences chez le jeune Louis XIII, rapportées dans le *Journal* d'Heroard.
[7] Voir De Mulder et Vanderheyden (2008) pour l'analyse détaillée de cet exemple, et de l'évolution sémantique du verbe *aller.*

On peut effectivement y voir, déjà, la valeur du marqueur discursif, ou bien celle de complément régi d'un verbe de parole implicite : « [je dis] à (ce) propos que entre le lieu où est le feu et le lieu est la terre, il se trouve un lieu moyen qui ne peut être vu ».

Cruciale pour l'établissement de la chronologie du changement, la détermination (et l'analyse) de cette étape intermédiaire est délicate, du fait de la complexité à analyser sémantiquement (et ce faisant à quantifier) les constructions qui en relèvent.

3.3.2. Contextes textuels : le bon « maillage » du corpus

Les changements s'initient dans des contextes *linguistiques* mais ils ont aussi parfois leur origine dans certains contextes *textuels* (liés aux genres ou aux registres), et une mauvaise constitution du corpus peut conduire à une interprétation totalement erronée des résultats. Ainsi dans le cadre d'une étude du marqueur de topicalisation *quant à*, j'avais interrogé la *Base de Français Médiéval* (*BFM*)[8] pour l'ancien français, et la base du *Dictionnaire de Moyen Français* (*DMF*)[9] pour le moyen français. Ne trouvant aucune occurrence de l'expression dans la *BFM*, j'en avais conclu à son émergence plus tardive. Mais il s'est avéré que les textes de moyen français dans lesquels apparaissait l'expression étaient très majoritairement de textes argumentatifs, en tout cas non littéraires ; or la *BFM* n'était à cette époque[10] composée que de textes littéraires : il était donc peu probable d'y rencontrer des occurrences de l'expression. Il reste vrai que l'expression est encore rare à cette époque, et qu'elle se développe véritablement à partir du 14ème siècle, ce qui s'explique par le fait qu'elle est effectivement liée à certains genres textuels non littéraires, au moins à ses débuts, et que les textes appartenant à ces genres ont pendant longtemps été rédigés en latin (voir Prévost 2010a).

Outre le bon ciblage des genres, se pose, pour la constitution du corpus, la question des « ellipses » temporelles, susceptibles d'occulter une phase importante de l'émergence, de la disparition ou de l'évolution d'une construction. On a intérêt à disposer du corpus le plus « dense » possible, sous peine de tirer des conclusions dont la validité ne tiendra qu'à l'insuffisance du corpus.

La question des fréquences se décline donc à différents niveaux : il s'agit de bien déterminer ce que l'on compte, et il s'agit par ailleurs de savoir dans quoi on le compte. Je ferai deux remarques à ce propos.

Quand on considère les sources des données extraites d'un corpus (ce qui n'est d'ailleurs pas toujours le cas, le travail « en aveugle » n'étant pas exceptionnel, voir mes remarques à ce propos en 2.1.), on ne considère que rarement les textes dans lesquels la construction n'apparaît pas, ne serait-ce que du point de vue de leur nombre. Or il est intéressant de connaître la proportion de textes, sur l'ensemble du corpus constitué, qui contiennent des occurrences de la construction recherchée. Par ailleurs, et c'est en partie lié, il convient de distinguer, d'un point de vue méthodologique, les constructions qui apparaissent dans tous les textes, selon des fréquences possiblement variables (voir en 4 l'étude des différentes réalisations syntaxiques du pronom personnel sujet), et celles qui ne se rencontrent que dans certains. Rappelons d'ailleurs que la rareté d'une construction peut être temporaire, et correspondre à une phase d'émergence ou au contraire de disparition.

[8] <http://bfm.ens-lyon.fr/>
[9] <http://www.atilf.fr/dmf/textes2010.htm>
[10] Les recherches avaient été effectuées en 2000.

Ainsi, pour une étude des expressions formées sur « propos » (*à propos de*, *à ce propos*, et *à propos*)[11], j'ai utilisé comme corpus de travail la base du *DMF*, qui comprenait 218 textes (et près de 7 millions d'occurrences). Il n'est ressorti de la collecte que 295 occurrences pertinentes de *propos.* Celles-ci n'apparaissaient en outre que dans 29 des 218 textes, et elles étaient largement concentrées chez trois auteurs. Plus étonnant encore, les rares occurrences de *à propos de X* en position initiale (29 en tout) étaient toutes présentes dans quatre textes du même auteur, Christine de Pizan. Il aura donc fallu un corpus de 7 millions de mots pour extraire 29 occurrences d'une construction utilisée exclusivement par un auteur. Il aurait suffi que les textes de C. de Pizan ne figurent pas dans le corpus pour en conclure que l'expression n'existe pas à l'époque, au moins dans les textes du *DMF* (base dont on peut considérer la taille comme respectable). Ce qu'aucune intuition de locuteur ne serait venue corriger. L'expression est certes encore bien rare en moyen français, mais elle existe.

3.4. Mesures et quantifications : une vision (faussement ?) objective

Dès lors qu'on procède à une quantification des données, quelle qu'elle soit, on semble objectiver le phénomène décrit. Outre le fait que les résultats mis au jour ne valent, en toute rigueur, que pour les textes qui composent le corpus, ils ne constituent, surtout, qu'une certaine vision dudit phénomène, en réalité largement subjective. En effet, dès lors que l'on dépasse le simple registre des fréquences, bon nombre de calculs requièrent une interprétation assez fine, susceptible de varier selon l'interprétant. Par ailleurs, même pour des calculs assez simples, se pose, en amont des résultats obtenus, la question de ce qui a été compté. Cette question a été évoquée en 3.3.1. à propos de l'interprétation sémantique de certaines occurrences. Elle émerge également au niveau (morpho-) syntaxique. Ainsi, selon le modèle théorique adopté, la notion de « langue V2 » ne revêt pas la même signification : il peut s'agir, pour le verbe, d'une position de surface, ou bien d'une position dans la structure profonde. On ne proposera pas le même codage des données dans les deux cas, et on ne dénombrera donc pas les mêmes structures. Tout codage résulte d'une grille de lecture, d'un parti-pris de classification, et toute classification est une construction conventionnelle, elle n'est pas « naturelle ». En tant que telle, elle informe nécessairement l'objet auquel elle est appliquée[12]. La marge de variation est particulièrement forte en syntaxe, champ dans lequel les approches théoriques, parfois radicalement différentes, peuvent donner lieu à des dénombrements très divergents des constructions et des phénomènes étudiés.

Le caractère non consensuel de l'objet mesuré et la variabilité des quantifications n'est certes pas spécifique à la langue ancienne. Mais, comme cela a déjà été souligné, nous ne pouvons mettre en œuvre une quelconque compétence de locuteur, susceptible de faire contrepoids à nos interprétations. En conséquence, les « mesures » établies à partir des textes n'en acquièrent que plus de poids, elles sont parfois perçues comme une forme de réalité, alors qu'elles ne sont pour une large part qu'une manière de donner à voir cette réalité.

Dans la mesure où c'est à partir de chiffres que nous formulons de nouvelles hypothèses, que nous révisons des règles précédemment établies, la prudence est de mise, tant dans la constitution du corpus que dans la détermination de ce qui doit être mesuré et dans la manière dont on établit les quantifications.

---

[11] Prévost (2008).

[12] Sur cette question, voir S. Loiseau dans l'introduction de ce numéro (3.1.), de même que les travaux de A. Desrosières, en particulier (Desrosières, 2001). J'ai par ailleurs développé cette discussion dans (Prévost 2011b).

Pour finir, je présenterai ci-dessous quelques aspects d'une étude menée sur l'évolution de l'expression et de la position du sujet pronominal en français médiéval, ce qui illustrera l'apport des quantifications dans un corpus relativement massif, mais aussi certaines limites de cette démarche.

4. Evolution de la syntaxe du sujet pronominal en français médiéval : comment concilier tendances générales et spécificités ?

4.1. Bref rappel historique

L'ancien français a hérité du latin une relative souplesse de l'ordre des mots (au sens où celui-ci n'est pas principalement conditionné par les fonctions syntaxiques, comme en français moderne, mais aussi, et surtout, par des considérations pragmatiques) et la tendance à ne pas exprimer le sujet quand celui-ci est cognitivement accessible (c'est une langue dite à sujet nul, comme la plupart des langues romanes). On rend généralement compte de la possibilité de ne pas exprimer le sujet par l'existence d'une morphologie verbale riche, qui permet l'identification du référent. La relative liberté dans l'ordre des mots est quant à elle expliquée par la présence d'une déclinaison nominale (réduite à deux cas en ancien français), qui permet d'identifier les fonctions des différents syntagmes, ainsi que par la contrainte du verbe en seconde position (l'ancien français est considéré comme une langue V2) , qui conduirait, lorsque un élément autre que le sujet occupe la première position, à postposer ce dernier au verbe. Perte de la morphologie verbale (Foulet 1930, Vance 1997), disparition de la déclinaison nominale, recul de la contrainte du verbe en seconde position : ces différents facteurs auraient convergé, dès l'ancien français, pour aboutir à une fixation de l'ordre des mots et plus spécifiquement, pour ce qui nous intéresse ici, de l'expression du sujet en position préverbale (la syntaxe moderne est quasiment acquise au 17ème siècle). Il s'agit d'un résumé assez schématique de l'évolution de la syntaxe du sujet, mais l'objectif n'est pas ici de développer cette question (voir Prévost 2010b, 2011a et sous presse). Je rappellerai simplement quelques points de discussion.

Tout d'abord, s'il y a bien eu, sur la période médiévale, une convergence chronologique entre la réduction de la richesse morphologique (verbale et nominale) et la fixation de l'expression du sujet en position préverbale, il n'est pas pour autant certain qu'un rapport de stricte causalité a joué entre les deux. On observe en effet très tôt en français des irrégularités dans la déclinaison (Schøsler 1984), qui n'est donc pas le seul facteur distinctif entre fonctions majeures. On sait par ailleurs que la morphologie verbale ne permettait plus de discriminer systématiquement les personnes, pour certains verbes et certains temps. Buridant considère ainsi que l'érosion phonétique n'a joué qu'un rôle de « catalyseur » (2000 : 438). Nul ne conteste le rôle du facteur phonétique, mais l'évolution de la syntaxe du sujet s'inscrit dans un mouvement plus général de fixation de l'ordre des mots, qui résulte du passage d'un principe d'organisation informationnel à un principe d'organisation grammatical, selon les fonctions (Vennemann 1976, Combettes 1988). Concernant le développement de l'expression du sujet, une autre explication a d'ailleurs été proposée, non exclusive de la précédente : c'est à partir des effets de mise en relief liés à l'expression du pronom en ancien français que celle-ci se serait progressivement systématisée. Je reviendrai sur ce point plus loin.

Par ailleurs, on considère le plus souvent que le sujet non exprimé correspond potentiellement à un pronom personnel, pour des raisons référentielles: la non-expression suppose un degré d'activation cognitive maximal du référent (même si l'on rencontre de rares exceptions), caractéristique qui rapproche le sujet non exprimé du pronom personnel. Plusieurs approches ont en outre établi un lien très fort en ancien français entre

la non-expression du sujet et la position postverbale du pronom personnel (entre autres Foulet, 1930 Skårup 1975, Adams 1987, Vance 1997, Buridant 2000), les tenants de la grammaire générative considérant le pronom sujet postverbal et le sujet non-exprimé comme deux contextes différents pour une même grammaire V2 (voir entre autres Kroch 1989). L'assimilation du sujet non-exprimé à un pronom postverbal omis s'appuie sur deux arguments principaux. Le premier, syntaxique, est lié à la contrainte du verbe en seconde position : en raison de la fréquence des séquences CV(X), on en conclut que, s'il avait été exprimé, le sujet pronominal aurait suivi le verbe Or, les séquences CSpV sont possibles en ancien français, même si elles sont rares :

(3) Par cele foi que je vos doi / Se cel anel de vostre doi / Ne m'envoiez, si que jel voie,/
***Rien qu'il deïst* ge ne croiroie** (*Beroul,* fin 12ème)
On Sp V
'je ne croirais rien de ce qu'il dirait'

Le second argument est de nature pragmatique : on rencontrerait les sujets non exprimés et les sujets postverbaux dans des contextes discursivement analogues. Ce n'est cependant pas le cas, le pronom personnel postposé apparaissant dans des contextes très spécifiques (voir Prévost 2010b et sous presse), contrairement au sujet non exprimé. Pour ces raisons, il semble préférable de ne pas assimiler pronoms postverbaux et sujets non exprimés, et d'aborder séparément la question de l'expression et celle de la position du sujet en général, sans pour autant nier les liens entre les deux.

C'est pour cette raison (mais aussi pour des raisons de place) que je me limiterai ici à l'étude du pronom personnel et n'envisagerai pas celle des autres sujets, en particulier nominaux. De plus si le sujet nominal et le pronom personnel sujet ont connu une évolution similaire, en termes de fixation en position préverbale, celle-ci n'en relève pas moins de motivations et de mécanismes en partie différents. En outre, alors que la postposition du sujet nominal en ancien français, et encore en moyen français (Prévost 2001), pouvait être extrêmement fréquente (supérieure à 50% de l'ensemble des sujets nominaux exprimés dans certains textes), celle du pronom personnel sujet a toujours été assez rare.

4.2. Une langue caractérisée par la variation idiolectale

L'étude[13] menée sur le recul de la non-expression et de la postposition du sujet pronominal (pronoms personnels de 1ère, 3ème et 6ème personnes) dans les propositions déclaratives s'appuie sur un corpus de 14 textes (458 000 mots), étiquetés morpho-syntaxiquement, et qui s'étalent de 1100 à 1400. Bon nombre proviennent de la *Base de Français Médiéval*, et ont été exploités avec le logiciel TXM[14]. Nous avons élaboré des requêtes pour collecter les séquences avec sujet préverbal, postverbal et sans sujet exprimé. Pour les séquences avec sujet – SpV et VSp – nous avons combiné étiquettes morpho-syntaxiques et chaînes de caractères, en recensant à cet effet l'ensemble des formes possibles pour le pronom personnels, 31 en tout pour les personnes 1, 3 et 6 (la casse étant neutralisée): la longueur de la liste est due à la présence d'une forte variation morphologique ainsi qu'à l'existence de nombreuses formes contractées pour la première personne (par exemple : pronom + pronom : *jel* (*je + le*)). La requête a tenu compte des éléments susceptibles de s'intercaler entre le pronom et le verbe (pronom complément, négation, ….), et nous avons essayé d'éliminer les mots interrogatifs et subordonnants en tête de phrase, afin d'exclure les propositions interrogatives et subordonnées. Le bruit s'élevait entre 2% et 12% selon les textes. Pour les sujets non exprimés, la requête s'est

[13] Voir Prévost 2012 et sous presse.
[14] <textometrie.ens-lyon.fr>

appuyée sur la catégorie morpho-syntaxique du verbe. Elle a été conçue dans le but de refuser la présence d'un mot subordonnant et /ou d'un pronom sujet devant le verbe, d'exclure au mieux les verbes de personnes 2, 4 et 5 (une liste de désinences a été conçue pour cela), et d'exclure les sujets pronominaux postverbaux. Le bruit, assez élevé (entre 40% et 50% des occurrences collectées, selon les textes), concerne majoritairement des séquences avec un sujet nominal, qu'il n'était pas possible d'éliminer sans courir le risque, en l'absence d'annotation syntaxique, d'éliminer un complément nominal.

L'ensemble des données collectées pertinentes se répartit ainsi :

	Sp préverbal	Sp postverbal	Sp non exprimé	Total
P1 (*je*)	1 709	252	1 488	3 449
P3 (*il, elle, ils, elles*)	2 300	401	9 691	12 392
Total	4 009	653	11 179	**15 841**

Tableau 1 : Fréquences absolues des sujets pronominaux préverbaux, postverbaux et non exprimés.

Voici un exemple de chacune des constructions :

SpV :

(4) Li reis Marsilie m'ad tramis ses messages.
De sun aveir me voelt duner grant masse,
[…].
Mais **il me mandet** que en France m'en alge… (*Chanson de Roland*, v.181-187)

'mais **il me demande** que je m'en aille en France'

VSp :

(5) Aussi dist on qu'il appareille / Une feste trop honnourable / Qui sera assez plus notable/ Que nulle qu'il fëist pieça ./ Et pour ce **croy je** mieux qu'il a/Haulte dame a femme rouvee,…(*Griseldis*, 1395)

'Et pour cela **je crois** qu'il a trouvé pour femme une noble dame'

S0 (non exprimé) :

(6) Amis le voit, moult en **est** esperduz./Or **se demente** et **dist** : « Las ! tant mar **fuz**/Que tu venis en terre » (*Amile*, vers 1200)

'Ami le voit, **il** en **est** tout éperdu; **Il se désole** et **dit** : hélas, c'est pour ton malheur que tu vins en cette terre'

Les deux tableaux suivants présentent, par texte, les fréquences absolues et relatives de la non-expression du sujet et de sa postposition.

	Date	Dialecte	forme	Domaine	Non expression de P1[15]	Non expression de P3	Non expr. P1+P3
Roland	1100	angl-normand	V[16]	litt.	*62.2* (166[17])	*94.4* (1341)	*89.3*
Eneas	1155	normand	V	litt.	*71.7* (241)	*85* (1446)	*82.8*

[15] Les fréquences relatives (en %) de non-expression de P1, de P3, et de P1 + P3, en italiques, sont calculées sur les ensembles respectifs P1 exprimés ou non, P3 exprimés ou non et P1 + P3 exprimés ou non.

[16] V = vers ; P = prose ; M= mixte.

[17] Entre parenthèses, en normal, sont indiquées les fréquences absolues de P1 et de P3 non exprimés.

Tristan, Béroul	1165-1200	traits norm.	V	litt.	*61.7* (240)	*89.7* (1037)	*82.7*
Ami et Amile	1200	non marqué	V	litt.	*64.1* (200)	*87.3* (816)	*81.5*
Consqtantinople, Clari[18]	après 1205	picard	P	histo.	*25* (12)	*83.3* (900)	*80.8*
Aucassin et Nicolete	1ère ½ 13e	traits picards	M	litt.	*37.5* (39)	*77* (357)	*69.7*
Miracles, G. de Coinci	1218-1227	non marqué	V	relig.	*70.5* (110)	*87.3* (563)	*84*
Queste del Saint Graal	1230	non marqué	P	Litt.	*21.5* (53)	*76.5* (1125)	*68.6*
Coustumes, Beaumanoir	1283	traits picards	P	jurid.	*26.7* (8)	*27* (55)	*26.9*
Mémoires, Joinville	1305- 1309	non marqué	P	histo.	*21.1* (71)	*54.1* (417)	*44.1*
Chroniques, Froissart	1369- 1400	franco-picard	P	histo.	*24.6* (30)	*73.2* (859)	*68.6*
Estoire de Griseldis	1395	traits picards	V	litt.	*65.1* (244)	*82.8* (106)	*69.6*
Manieres de Langage	1396, 1399	non marqué	M	didact.	*4.5* (16)	*36.6* (79)	*16.7*
Quinze Joyes de Mariage	1400	non marqué	P	litt	*15.5* (58)	*57.3* (591)	*46.1*

Tableau 2: Fréquences de non-expression de P1, P3, et P1+P3.

	Inversion P1[19]	Inversion P3	Inversion P1+P3
Roland (1100)	*17.8* (18[20])	*25.3* (20)	*21.1*
Eneas (1155)	*12.6* (12)	*8.7* (22)	*9.7*
Tristan, Béroul (1165-1200)	*16.1* (24)	*19.3* (23)	*17.5*
Ami et Amile (1200)	*25.9* (29)	*21.8* (26)	*23.8*
Conqueste, Clari (après 1205)	*0*	*51.1* (92)	*42.6*
Aucassin et Nicolete (1ère moitié 13ème)	*16.9* (11)	*5.6* (6)	*9.9*
Miracles, G. de Coinci (1218-1227)	*19.6* (9)	*23.2* (19)	*21.9*
Queste del Saint Graal (1230)	*33.1* (64)	*15.9* (55)	*22.1*
Coustumes de Beauvaisis, Beaumanoir(1283)	*18.2* (4)	*15.4* (23)	*15.8*
Mémoires, Joinville (1305-1309)	*15* (40)	*10.2* (36)	*12.3*
Chroniques, *Froissart* (1369-1400)	*5.4* (5)	*9.9* (31)	*8.8*
Estoire de Griseldis (1395)	*11.4* (15)	*27.3* (6)	*13.7*
Manieres de Langage (1396,1399)	*3* (10)	*7.3* (10)	*4.2*
Quinze Joyes de Mariage (1400)	*3.5* (11)	*7.3* (32)	*5.7*

Tableau 3: Fréquences de postposition de P1, P3, et P1+P3

Outre le critère chronologique, les textes avaient été sélectionnés en fonction de différents critères (*domaine* : littéraire, religieux, historique…, *dialecte* et *forme* : vers, prose, et mixte), lesquels ne se sont révélés discriminants que de manière très marginale. Le but de cette étude était triple : il s'agissait d'affiner, en termes de fréquences, le recul de l'inversion et de la non-expression du sujet pronominal, et de déterminer si les critères autres que chronologique jouaient un rôle dans cette évolution. Il s'agissait par ailleurs d'affiner la description des deux phénomènes en distinguant les personnes 1 (P1 : *je*) et 3 (P3 : *il, ils, elle, elles*) afin de voir si elles avaient connu une évolution différenciée[21]. Il s'agissait enfin de déterminer si, statistiquement, les deux évolutions étaient corrélées.

Je commencerai par ce dernier point, que je traiterai rapidement. La question de la possible corrélation entre les deux évolutions a plusieurs origines. D'une part, si l'on

[18] *La Conqueste de Constantinople,* de Robert de Clari

[19] Les fréquences relatives (en %) d'inversion de P1, de P3, et de P1 + P3, en italiques, sont calculées sur les ensembles respectifs P1 exprimés, P3 exprimés et P1 + P3 exprimés.

[20] Entre parenthèses, en normal, sont indiquées les fréquences absolues de P1 et de P3 inversés.

[21] Nous avons laissé de côté, pour cette étude, les personnes 4 et 5, dont les effectifs sont bien inférieurs à ceux de P1 et P3.

considère une chronologie large (ancien et moyen français) on constate que l'inversion et la non-expression ont reculé simultanément ; d'autre part, il n'est pas aberrant, linguistiquement, de considérer qu'un lien ait pu exister entre les deux évolutions : même si je défends l'idée que postposition et non-expression sont deux réalisations différentes du sujet pronominal, il n'en demeure pas moins qu'elles sont en partie liées. Enfin, il me semblait intéressant d'apporter une autre vision que celle proposée par Kroch.

Dans son étude de 1989, Kroch applique à l'évolution du sujet (pronominal) l'hypothèse de l'effet du taux constant (Constant Rate Effect). Selon lui, inversion et omission auraient connu un déclin simultané, l'hypothèse se fondant sur le postulat que sujet postverbal et sujet non exprimé constituent deux contextes différents pour une même grammaire V2. Dépourvue d'un tel postulat théorique et munie du seul indice de corrélation, j'en suis arrivée à un résultat fort différent. Le calcul a été établi pour évaluer la liaison statistique entre l'inversion et la non-expression, pour P1 d'une part, pour P3 d'autre part. Pour P1, l'indice de corrélation est de 0.37 tandis qu'il est de 0.43 pour P3. Les deux valeurs correspondent à une probabilité supérieure à 10% d'atteindre ou de dépasser ces coefficients : on ne peut donc rejeter l'hypothèse nulle, et il faut conclure à l'absence de dépendance statistique entre les deux évolutions.

Cet exemple illustre le propos développé en 3.4. à propos de l'incidence des postulats théoriques sur ce que l'on recense, et sur la manière dont on le recense.

Le choix de distinguer les personnes 1 et 3 trouve en partie son origine dans l'hypothèse émise par de Detges (2003), selon laquelle le développement du pronom se serait produit à partir d'effets de mise en relief liés à son expression. Il aurait d'abord été utilisé à des fins de stratégie discursive, dans des contextes de prise de parole (impliquant P1). Son usage répété aurait entraîné une dévaluation rhétorique, et donc une généralisation de l'emploi, elle-même cause de son affaiblissement. Cette hypothèse a été confirmée par les textes du corpus : dès les plus anciens textes, on observe une fréquence systématiquement accrue de l'expression de P1 comparée à celle de P3 (sauf dans un texte). Sans surprise, le recours au khi2 fait apparaître des valeurs très élevées : cela signifie que la probabilité que la distribution résulte du hasard est faible. Plus précisément, on observe une attraction entre la non-expression et P3 et/ou une attraction entre l'expression et P1. C'est particulièrement net pour les 3 textes historiques, qui présentent un même khi2 et une configuration analogue (attraction entre P1 et l'expression). Le seul texte qui déroge à cette tendance massive est un texte juridique (*Coustumes de Beauvaisis* de P. de Beaumanoir, 1283). Cela est dû à une expression élevée de P3 (elle est identique à celle de P1), ce qui s'explique probablement par la nécessité, dans ce type de texte, d'expliciter plus systématiquement les référents.

Si la fréquence nettement plus élevée de l'expression de P1 trouve une justification assez convaincante dans l'hypothèse émise par Detges, on peut aussi suggérer qu'elle est liée à la situation de discours direct, dans lequel le pronom de première personne apparaît très préférentiellement. De fait, de premiers sondages dans quelques textes montrent que, pour P3, l'expression est environ 1.5 fois plus élevée en discours direct qu'en récit (une telle comparaison n'est guère pertinente pour les occurrences de P1, qui n'apparait que de manière marginale en récit). Cette piste est à poursuivre (Foulet l'avait déjà suggérée en 1930), et est d'ailleurs pleinement compatible avec l'hypothèse de Detges.

Par ailleurs, une étude, en cours, sur la syntaxe du pronom personnel dans les subordonnées fait apparaître de résultats bien différents pour ce qui concerne l'expression du pronom personnel sujet : d'une part, dès les plus anciens textes elle y est bien plus élevée que dans les déclaratives, et elle passe la barre des 90% au début du 13ème siècle ; d'autre part les écarts entre les fréquences d'expression selon la personne sont minimes.

Nous différons l'interprétation approfondie de ces données (déjà signalées et étudiées, dans un cadre générativiste, par Dupuis 1988 et Hirschbühler 1989 et 1990), mais elles soulignent la nécessité d'un traitement différencié des différents types de propositions.

Si l'on considère maintenant la position du sujet, dans les déclaratives[22], on observe une situation différente de celle constatée pour l'expression : les écarts de fréquence entre P1 et P3 ne sont pas négligeables, mais la fréquence la plus élevée n'est pas systématiquement au profit de la même personne, la situation varie selon les textes.

Toutefois, qu'il s'agisse de l'inversion ou de la non-expression, on observe dans certains textes des écarts importants entre P1 et P3, masqués lorsque l'on recourt à une fréquence moyenne P1 + P3. Ainsi, dans la moitié des textes du corpus, le rapport entre la fréquence de non-expression de P1 et celle de P3 oscille entre 2 et 4. Les *Chroniques* de Froissart et *l'Estoire de Griseldis* (deux textes de la fin du 14ème siècle) présentent une même fréquence globale de non-expression (69%), mais ce chiffre recouvre une fréquence de non-expression de P1 de 24.6% dans *Chroniques*, contre une fréquence de 65.1% dans *Griseldis*.

L'inversion est globalement peu fréquente (hormis P3 dans la *Conqueste de Constantinople* de R. de Clari, 1205 : 51%, chiffre tout à fait atypique pour les textes médiévaux). Si l'amplitude de variation entre textes est assez faible (le texte de Clari mis à part), qu'il s'agisse de la fréquence de P1 ou de P3, le rapport entre les fréquences d'inversion de chacune des personnes n'en présente pas moins certains écarts importants. Ainsi, dans 6 textes du corpus (*Conqueste de Constantinople, Aucassin et Nicolete,* début 13ème, *Queste del Saint Graal*, 1230, *Estoire de Griseldis, Manières de langage* et *Quinze Joyes de Mariage*, 1400), le rapport est supérieur à 2, au profit de P1 ou de P3**.**

Cela signifie, comme pour l'expression, qu'il faut relativiser la moyenne d'inversion, qui dissimule parfois de fortes disparités entre P1 et P3. Ainsi dans la *Queste del Saint Graal*, la fréquence globale d'inversion est de 22%, mais cette moyenne recouvre une fréquence de 33% pour P1 et de seulement 16% pour P3.

On observe par ailleurs une forte variabilité des fréquences entre textes, y compris pour une même période. Ainsi, si l'on considère les cinq textes du début du 13ème siècle (*Ami et Amile, Miracles* de Coinci, *Conqueste de Constantinople, Aucassin et Nicolete* et *Queste del Saint Graal*), la fréquence de non-expression moyenne de P1 est de 43%, mais ce chiffre recouvre des fréquences qui vont de 21% (*Queste*) à 64 % (*Ami Amile*).

Si l'on considère par ailleurs la fréquence d'inversion de P3 dans ces 5 mêmes textes, on obtient une moyenne de 23.5 %, chiffre très voisin de la fréquence d'inversion dans deux des textes (*Ami Amile* et *Miracles*), mais qui recouvre par ailleurs une grande disparité des chiffres : de 5.6% dans *Aucassin* à 51% dans le texte de Clari.

Certes, c'est le propre d'une moyenne de pouvoir donner à voir une image assez artificielle de la réalité, mais la variation idiolectale que connaît le français médiéval, au moins en ce qui concerne la syntaxe du sujet pronominal, rend particulièrement problématique le recours à ce calcul, à moins d'expliciter la réalité qu'il recouvre, ou bien de le compléter par d'autres calculs (écart type, médiane, …), qui permettent d'affiner le caractère un peu brut de la fréquence relative. Encore rare il y a quelques années, le recours à des calculs statistiques plus complexes que la simple fréquence relative se développe parmi les médiévistes.

Il n'en demeure pas moins que le comportement très spécifique de certains textes fait difficulté. *La Conqueste de Constantinople* présente ainsi des fréquences d'inversion tout à fait atypiques, exceptionnellement élevée pour P3, mais nulle pour P1. Dans *Les*

[22] En subordonnée, l'inversion du pronom sujet est extrêmement rare.

Coustumes de Beauvaisis, c'est au contraire la non-expression qui se singularise, puisque les fréquences sont identiques pour P1 et P3, et très basses. Dans *Griseldis*, les fréquences de non-expression de P1 et P3 sont relativement élevées (de même que l'inversion de P3). Enfin, dans la *Queste del Saint Graal*, la fréquence d'inversion de P1 est très élevée, et supérieure à celle de non-expression de P1. On peut s'interroger sur le traitement à adopter pour ces textes : qu'en faire ? Du point de vue synchronique, ils « cassent » les moyennes, et du point de vue diachronique, ils provoquent des pics et des creux très marqués.

4.2. Représentations graphiques : des « courbes » trompeuses.

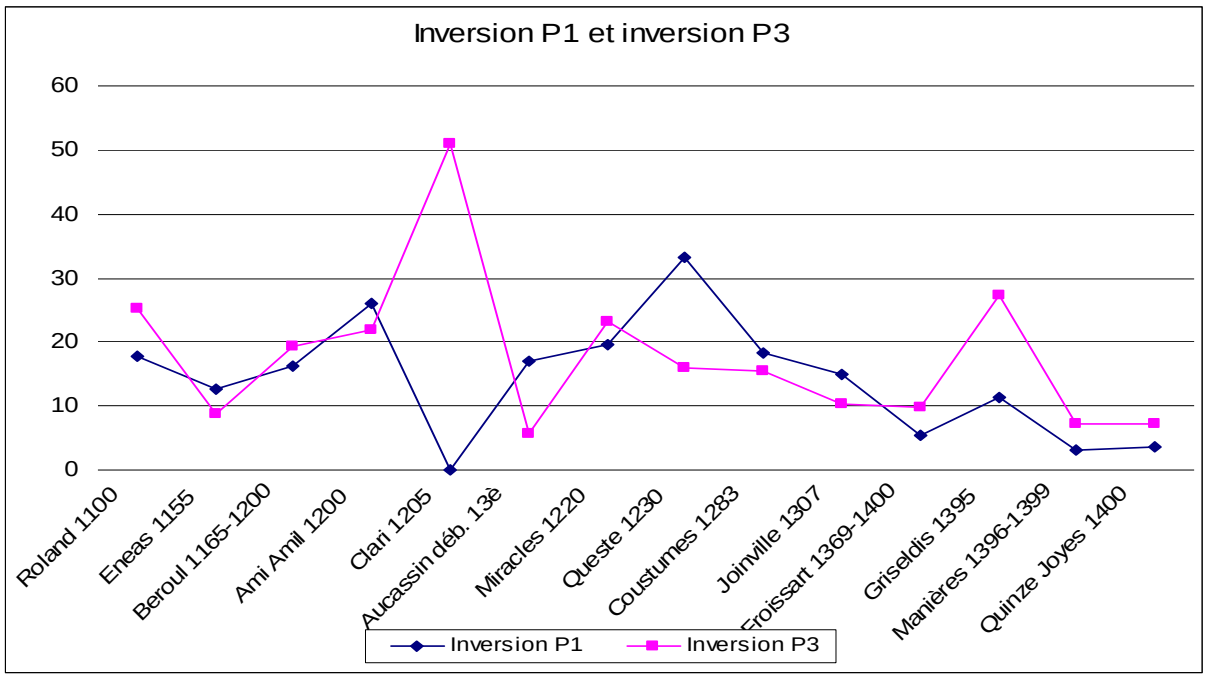
Les représentations dites en 'courbes' permettent une appréhension globale, synthétique et immédiate des différentes fréquences d'un phénomène, et facilitent la comparaison des fréquences de différents phénomènes. Elles dessinent aussi une évolution, dont la nature est définie par l'axe des abscisses, qui, dans l'étude en question, comprend les textes et leurs dates, ordonnés selon un ordre chronologique. Mais la représentation dite en courbe est en partie trompeuse. En effet, la ligne qui relie deux points oblitère ce qui se passe possiblement « entre » ces deux points, donnant à voir une ligne droite ascendante ou descendante, sans accident de parcours[23]. Or l'exemple du texte de Clari nous montre combien un seul texte peut changer l'allure d'une courbe : supprimons *Clari* du corpus (ce qui ne provoquerait même pas de trou dans la chronologie, la période étant représentée par d'autres textes), et la courbe du graphique 1 ci-dessous n'aura plus du tout la même allure[24].

Se pose donc à nouveau la question de la représentativité du corpus, qui se décline ici en termes de granularité du découpage de l'espace temporel et du nombre de textes (diversifiés) à prendre en compte. Peut-on, concernant ces deux aspects, définir des seuils à partir desquels on puisse affirmer que les courbes (ou les lignes) des représentations graphiques dessinent bien des évolutions, et non le simple parcours d'un texte à l'autre ? On ne sait cependant sur quoi s'appuyer objectivement pour déterminer ces seuils.

Par ailleurs, faut-il fusionner les textes d'une même date, ou appartenant à une même période, et fournir des chiffres globaux ? Au vu de la diversité des pourcentages pour les textes du début du 13ème siècle, on peut en douter. Une éventuelle fusion des fréquences d'inversion de différents textes contemporains à une même date ne doit en tout cas pas occulter les fréquences propres à chaque texte.

[23] Raison pour laquelle l'appellation 'graphique en ligne' est plus juste que celle de 'graphique en courbe'.

[24] Je laisse ici de côté le problème du biais introduit par le caractère chronologiquement non proportionnel des espaces entre les textes, qui génère une vision 'déformée', la période du début du 13ème siècle étant étirée de façon disproportionnée.



Graphique 1: Inversion P1 et inversion P3 (axe des ordonnées : pourcentages)

Conclusion

La quantification accrue des données permet assurément une meilleure appréhension des phénomènes de changement, plus complète et précise. Mais elle conduit aussi à s'interroger, peut-être plus qu'avant, ou en tout cas différemment, sur la pertinence des fréquences établies. Si autrefois l'on pouvait parfois douter de la légitimité à dégager des régularités, voire des règles, à partir d'un nombre trop restreint de textes, on peut mettre en cause, aujourd'hui, la validité de résultats qui dissimulent parfois une forte variabilité entre textes.

La quantification des données tend à objectiver le phénomène décrit ; les chiffres, qui ne sont en toute rigueur que l'interprétation d'une réalité langagière, en viennent parfois à se substituer à cette dernière, avec laquelle les nouveaux modes de lecture peuvent par ailleurs nous faire perdre contact. Dès lors, si l'objectif reste bien un accroissement des données, et donc des textes, celui-ci doit s'accompagner de leur traitement raisonné et ordonné, sous peine de voir le traitement quantitatif perdre sa pertinence pour l'interprétation qualitative.

Bibliographie

ADAMS, M. (1987), *Old French, Null Subjects and Verb Second Phenomena*, Ph.D. Dissertation, University of California, Los Angeles.

BYBEE J.(2003), « Mechanisms of change in grammaticization : The role of frequency », *in* B. Joseph & J. Janda (eds.) *The Handbook of Historical Linguistics.* Oxford: Blackwell. 602-623.

COMBETTES, B. (1988): *Recherches sur l'ordre des éléments de la phrase en moyen français* (Thèse pour le Doctorat d'Etat, Université de Nancy ; exemplaire dactylographié).

CORBIN P. (1980), « De la production des données en linguistique introspective », in A.-M. Dessaux-Bertheneau (éd.) *Théories linguistiques et traditions grammaticales*. Villeneuve d'Asq : PU de Lille.

DE MULDER W. & VANDERHEYDEN A. (2008), « Grammaticalisation et évolution sémantique du verbe *aller*. Inférence, métonymie ou métaphore? », *in* B. Fagard & al. (éds) *Evolutions en français – Etudes de linguistique diachronique*, Bern : Peter Lang, 21-44 .

DESROSIERES A. (2001), « Entre réalisme et conventions d'équivalence : les ambiguïtés de la sociologie quantitative », *Genèses* 43, 112-127.

DETGES U. (2003), « Du sujet parlant au sujet grammatical. L'obligatorisation des pronoms sujets en ancien français dans une perspective pragmatique et comparative », *Verbum* XXV, 3, 307-333.

FOULET L. (1930,1ère ed. 1919), *Petite syntaxe de l'ancien français*. Champion, Paris.

HEINE B. (2002), « On the role of context in grammaticalization », *in* I. Wischer & G. Diewald (éds), *New Reflections on Grammaticalization*, Amsterdam : John Benjamins, 83-101.

HIRSCHBÜHLER, P. (1989). «On the existence of null subjects in embedded clauses in Old and Middle French», dans C. Kirschner and J. de Cesaris, *Studies in Romance Linguistics*, Benjamins, Amsterdam: 155-176.

HIRSCHBÜHLER, P. (1990). «La légitimation de la construction V1 en subordonnée dans la prose et le vers en ancien français», *Revue Québécoise de Linguistique* 19.1: 33-55.

KROCH T. (1989), « Reflexes of grammar in patterns of language change », *Language Variation and Change* 1, 199-244.

LIGHTFOOT D. (1979), *Principles of diachronic Syntax*. Cambridge : Cambridge University Press.

MARCHELLO-NIZIA C. (1995), *L'évolution du français : ordre des mots, démonstratifs, accent tonique*. Paris : Armand Colin.

MILNER J.-C. (1989), *Introduction à une science du langage*, édition abrégée, Paris : éditions du Seuil, collection Points Essais.

PERY-WOODLEY M.-P. (1995), « Quels corpus pour quels traitements automatiques ? », *TAL*, 36, (1-2), 213-232.

PREVOST S. (2010a), « *Quant à X* : du complément à l'introducteur de topique en passant par l'introducteur de cadre » *in* B. Combettes & al. (éds), Le *changement en français. Etudes de linguistique diachronique.* Bern : Peter Lang, 325-343.

PREVOST S. (2010b), « Evolution de la position du sujet pronominal en français médiéval: une approche sémantico-pragmatique », in Neveu F., Muni Toke V., Durand J., Klingler T., Mondada L., Prévost S. (éds.), *Congrès Mondial de Linguistique Française - CMLF 2010*, Paris: Institut de Linguistique Française, 305-320 [http://dx.doi.org/10.1051/cmlf/2010106]

PREVOST, S. (2011a), « Expression et position du sujet pronominal en français », *Mémoires de la Société de Linguistique de Paris*, Tome XIX, 13-33.

PREVOST S. (2011b), « Français médiéval en diachronie : du corpus à la langue » (Mémoire de synthèse HDR, http://tel.archives-ouvertes.fr/docs/00/66/71/07/PDF/prevost_HDRsynthese080711.pdf

PRÉVOST S. (2011c), « *A propos :* from verbal complement to 'utterance marker' of discourse shift », *Linguistics*, 49 : 2, p 391-413.

PREVOST S. (2012), « Expression et position du sujet pronominal du 12ème au 14ème siècle : une approche quantitative et contrastive », *in* A . Dister, D. Longrée, et G.

Purnelle (éds) *Actes des 11èmes Journées Internationales d'Analyse des Données textuelles (JADT 2012)*, édition en ligne

PREVOST, S. (sous presse) « Recul de la non-expression et de l'inversion du sujet pronominal du 12ème au 14ème siècle: une approche quantitative et qualitative », in *Actes du colloque Diachro VI*, Carlier A., Goyens, M. et Lamiroy B. eds, Bern : Peter Lang.

ROMAINE, S. (1989), « The role of children in linguistic change », *in* Breivik, L.E. & Jahr, E.H. eds, 199-226.

SKÅRUP, P. (1975), *Les premières zones de la proposition en ancien français. Essai de syntaxe de position.* Etudes romanes de l'Université de Copenhague, *Revue Romane*, numéro spécial 6, Akademisk Forlag.

VANCE, B. (1997), *Syntactic Change in Medieval French: Verb-Second and Null Subjects*, Kluwer Academic Publishers, Dordrecht-Boston-Londres.

VENNEMANN, T. (1976). «Topics, subjects and word-order : from SXV to SVX via TVX », in J.M Anderson & C. Jones eds. *Proceedings of the first international congress of Historical Linguistics*. Amsterdam: 339-376.