Expression et position du sujet pronominal du 12ème au 14ème siècle : une approche quantitative (recherche inédite)

Sophie Prévost

Expression et position du sujet pronominal du 12ème au 14ème siècle : une approche quantitative

Sophie Prévost

Chargée de recherche au Lattice (UMR 8094), ENS /Paris-3

Recherche inédite
en vue de l'obtention de
l'Habilitation à diriger des recherches
sous la direction de
M. Benoît Habert, Professeur des universités

Jury :

M. Michel Charolles
M. Bernard Combettes
M. Benoît Habert
Mme Christiane Marchello-Nizia
Mme Marie-Paule Péry-Woodley
Mme Lene Schøsler

Ecole Normale Supérieure de Lyon
2011

Table des matières

Introduction*

J'avais étudié dans ma thèse, publiée sous forme remaniée dans Prévost 2001, l'évolution de la postposition du sujet nominal et pronominal dans six textes de moyen français. L'approche était de type informationnel, fondée sur le cadre théorique développé dans Lambrecht 1994, et je montrais que l'explication consistant à considérer les énoncés verbe-sujet comme une rémanence de l'organisation informationnelle de l'énoncé qui avait prévalu en ancien français[1] ne suffisait pas. En effet, les énoncés 'verbe-sujet nominal' (qui représentent entre 13% et 52% de l'ensemble des énoncés à sujet nominal dans les 6 textes étudiés, qui vont du milieu du 15ème au milieu du 16ème siècle) ne répondaient pas systématiquement à la présence d'un sujet cognitivement nouveau ou porteur d'information nouvelle, et en outre, et surtout, les énoncés 'verbe-sujet pronominal' constituaient d'emblée une énigme dans la mesure où le pronom, et en particulier *il*, le plus fréquent d'entre eux, constitue un thème ou un topique par excellence : rien ne justifie donc, dans une perspective informationnelle, sa position postverbale. Cela explique probablement la rareté de celle-ci dès les plus anciens textes (entre 8% et 23% des énoncés à sujet pronominal dans ce même corpus).

J'avais montré que la postposition du sujet, nominal ou pronominal, s'explique en revanche par la présence d'une relation de rupture avec le contexte précédent, diversement actualisée. Pour les sujets nominaux, elle se réalise selon quatre modalités principales :

a) par des transitions narratives (introduction de nouveaux référents et opérations de recentrage sur un référent) :

(1) Et alors se changea tout ordre et tout conseil, car chascun se mectoit à en dire son advis, et ja estoit commencée **une grosse et forte escarmouche** au bout du villaige de Montlehery, toute d'archiers d'un costé et d'autre (*Mémoires* de Commynes, 1489-1490).

b) par des ruptures énonciatives (passages du récit au discours ou l'inverse) :

* Je remercie Benjamin Fagard pour la relecture de ce mémoire et pour ses remarques.

[1] Organisation formulée en termes de 'thème-rhème' (progression du plus connu vers le moins connu), ou bien en termes de 'topique-commentaire' (on pose d'abord 'ce dont on parle', puis on énonce 'ce qu'on en dit').

(2) ... sy vous requier tres acertes que vous en dïez vostre advis et en rendez ycy vostre determinacion. » Atant se teult **le sage roy de Castille**,… (*Roman du conte d'Artois*, 1453-1467)

c) par des enchaînements narrativement inattendus (jeux de surenchère informative)

(3) En ce temps vint nouvelles en Espaigne comme le roy de France alla de vie a trespas, dont fut demené ung merveilleux dueil par le roy et la royne et les barons du pays, et n'y eut monastere, esglise ne convent ou le roy ne fit faire obseques, prieres et oraisons pour l'ame du bon roy et en porterent **le roy et la royne** le dueil ung an et moult bien en firent leur devoir. (*Jehan de Paris*, 1494)

d) enfin, par des ruptures syntaxiques (antéposition au verbe de compléments essentiels, dont l'objet nominal).

(4) ...car *tant* sont grandes **les doubtes que sa dame ne en perde, et preine desplaisir** que un seul deshonneste penser n'en est en lui; (A. de la Sale, *Jehan de Saintré*, 1456)

En ce qui concerne les sujets pronominaux, l'étude a montré qu'il convient de traiter séparément les pronoms indéfinis, impersonnels et personnels, les plus nombreux. Dans les énoncés incluant ces derniers, les ruptures passent le plus souvent par des relations adversatives, consécutives, comparatives, ou explicatives, liées à des jeux de retournements argumentatifs, comme dans l'exemple ci-dessous (relation adversative) :

(5) Et combien que son maistre l'envoyast souvent querir, si ne retourna **il** poinct à la court, qu'il ne fust bien guery de toutes ses playes,… (M. de Navarre, *L'Heptameron*, 1549)

On peut interpréter la position inhabituelle du sujet pronominal (au regard du principe fonctionnel et du principe grammatical en train de se mettre en place) comme le signal d'une validation problématique de la relation prédicative : l'expression du sujet y est indispensable pour bien montrer que cette relation est *malgré tout* validée, mais l'on signale en même temps la difficulté de cette opération (ou son caractère inattendu) par un positionnement inhabituel du sujet. Etant donné la raréfaction croissante des énoncés verbe-sujet à partir du moyen français, les structures à pronom postverbal qui se maintiennent à partir de ce moment-là apparaissent comme particulièrement marquées. Elles seront bientôt limitées à la présence de quelques adverbes à caractère épistémique ou argumentatif (*Paul est malade ; peut-être* ***viendra-t-il*** *malgré tout*), et correspondent désormais toutes, comme l'a montré Guimier (1997), à des contextes de validation problématique de la relation prédicative, et donc à des énoncés non

pleinement assertifs. Il reste que, pour le locuteur moderne, ces constructions sont plutôt interprétées comme des tournures figées que comme dénotant une quelconque inscription du locuteur dans son énoncé.

D'emblée cette recherche s'était donné certaines limites : la période envisagée était assez courte (1450-1550), appréhendée à travers trois coupes synchroniques (début, milieu et fin de la période) pour chacune desquelles deux textes avaient été sélectionnés. Le caractère relativement restreint du corpus était justifié à différents égards : une large part des constructions recherchées (les sujets nominaux) ne pouvaient l'être que « manuellement », les textes sur lesquels je travaillais, numérisés en mode texte, n'étant pas enrichis d'informations syntaxiques (la collecte des données était plus simple pour les sujets pronominaux, qui correspondent à une liste finie de formes). Par ailleurs, la finesse de l'analyse informationnelle menée sur les énoncés à sujet postverbal n'aurait pas été compatible, dans le temps imparti, avec le traitement d'un nombre élevé de cas. Enfin, une partie importante de la recherche avait été consacrée à l'analyse critique des cadres d'analyse fonctionnels existants, et à l'élaboration, à partir du modèle proposé dans Lambrecht (1994), d'un cadre d'analyse adapté aux spécificités de la langue envisagée[2]. Il s'agissait certes de proposer des avancées en matière d'évolution de la syntaxe du sujet en moyen français, mais aussi, et tout autant, de tester un cadre d'analyse non initialement prévu pour cet état de langue, et par conséquent remanié pour y être adapté.

J'avais délibérément laissé de côté deux extensions possibles à l'analyse proposée. Je n'avais pas pris en compte la période plus ancienne de l'ancien français, en partie pour des raisons matérielles de temps, mais surtout parce que m'intéressait prioritairement la période de transition que constitue le moyen français (passage d'une organisation informationnelle de l'ordre des mots à une organisation selon les fonctions syntaxiques). J'avais par ailleurs laissé de côté la question de l'expression du sujet, cela pour des raisons très clairement pratiques. En effet, traiter les sujets non-exprimés constituait une tâche dont la lourdeur n'était pas compatible avec le cadre restreint d'un travail de thèse : le repérage automatique des constructions n'était pas possible en l'absence de textes enrichis syntaxiquement et lesdites expressions s'avéraient en

[2] En particulier, le moyen français présente un fonctionnement des expressions référentielles assez différent de celui du français ou de l'anglais modernes, et il n'est par ailleurs pas possible de s'appuyer sur la prosodie (ni même sur une prosodie intérieure et silencieuse, comme on peut le faire pour un texte écrit moderne).

outre très fréquentes, d'où un traitement très long. Le travail avait donc été centré sur la position du sujet exprimé, nominal ou pronominal. Il ne m'en est pas moins resté la conviction que si l'évolution de la syntaxe du sujet pronominal devait être envisagée en relation avec celle du sujet nominal, elle devait aussi l'être avec celle du sujet non exprimé. Il est en effet admis de manière consensuelle que les sujets non exprimés correspondent à de potentiels sujets pronominaux (l'absence de toute expression explicite du sujet signifie en effet que son référent est hautement accessible du point de vue cognitif, et l'expression grammaticale généralement associée à une accessibilité référentielle élevée est le pronom personnel). L'accord est bien moindre en ce qui concerne la position potentielle du sujet pronominal non exprimé. Nous reviendrons sur ce point.

Si les six textes étudiés dans le cadre de ma thèse ont permis à la fois de tester un cadre d'analyse et de dégager des conclusions que je pense intéressantes, ils n'en constituent pas moins un corpus restreint, tant par le nombre de textes que par la période couverte. La réflexion menée depuis des années dans le cadre des linguistiques de corpus ainsi que mon questionnement personnel sur les corpus de langue ancienne m'ont largement convaincue de la nécessité de travailler sur un corpus aussi représentatif que possible du point de vue quantitatif et qualitatif. C'est d'autant plus indispensable lorsqu'il s'agit d'une langue pour laquelle nous n'avons pas de compétence susceptible de jouer le rôle de garde-fou vis-à-vis de nos interprétations, langue qui se révèle en outre particulièrement sujette à la variation d'un texte à l'autre, au moins pour ce qui concerne la syntaxe du sujet (voir (Prévost 2010) où est développée l'idée de micro-systèmes propres aux textes).

L'évolution de la syntaxe du sujet nominal n'a assurément pas livré tous ses secrets. C'est néanmoins celle du sujet pronominal, et plus spécifiquement des pronoms personnels, que j'ai décidé d'explorer de nouveau, cela pour plusieurs raisons.

La première tient à ce que certains aspects de la syntaxe du sujet pronominal (entendu désormais comme du seul pronom personnel, abrégé Sp) et de son évolution résistent encore à l'analyse. Ainsi en est-il de l'inversion[3] du sujet. Nous avons vu plus haut que qu'il est difficile d'appréhender la variation positionnelle du sujet pronominal

[3] Le terme d' « inversion » ne suppose ici aucun déplacement, mais dénote simplement le caractère toujours minoritaire de la position postverbale du sujet. Il est donc synonyme de « postposition ». Les deux sont ici employés indifféremment.

dans un cadre informationnel, sa position postverbale constituant d'emblée une aberration de ce point de vue (à l'inverse, même si tous les cas de postposition du sujet nominal ne correspondent pas à la présence d'un sujet porteur d'information nouvelle, la perspective informationnelle reste un cadre pertinent pour envisager bon nombre de cas). Pour autant, l'explication selon laquelle, dans le cadre d'une grammaire où le verbe occupe (très majoritairement) la seconde position (ancien français et encore moyen français), la présence d'un élément en tête d'énoncé « provoque » l'inversion du sujet, n'est pas pleinement satisfaisante, et ce d'autant moins que l'étude menée sur les textes de moyen français a prouvé que c'est loin d'être le cas et que les ressorts de l'inversion pronominale sont bien plus complexes. De même l'idée fréquemment avancée selon laquelle les sujets non exprimés correspondraient à des sujets inversés, puis omis, laisse à désirer, pour des raisons que j'exposerai plus bas.

Il m'a donc paru intéressant de poursuivre l'étude des sujets pronominaux, et il m'a semblé nécessaire de le faire en examinant conjointement l'évolution de la position des sujets pronominaux et celle de leur expression, afin de valider ou au contraire d'infirmer, au moins en termes quantitatifs, la relation forte souvent postulée entre les deux phénomènes. La prise en compte de l'expression du sujet constitue l'un des apports à l'étude précédemment menée. Le second apport réside dans l'élargissement de la période considérée, qui inclut désormais l'ancien français. En effet, dans la mesure où l'inversion pronominale ne répond pas à des motivations de type informationnel, la période du moyen français, qui voit l'organisation de l'ordre des mots passer d'un principe informationnel à un principe syntaxique, ne constitue plus, d'emblée, un moment aussi décisif. Et l'exploration, même superficielle, de quelques textes d'ancien français montre, d'une part que les séquences 'verbe-sujet pronominal' ont toujours été largement minoritaires, et d'autre part que l'interprétation en termes de rupture par rapport à ce qui précède a déjà toute sa pertinence en ancien français, même si elle ne revêt pas exactement les mêmes modalités qu'en moyen français.

Enfin, le troisième apport réside dans l'adoption d'une approche quantifiée réalisée sur un corpus qui tente de répondre autant que possible à des exigences de représentativité sur les plans quantitatif et qualitatif. Comme je l'ai rappelé plus haut, une telle démarche s'avère particulièrement nécessaire lorsque l'on travaille sur une langue dont on n'a pas la compétence. Les premiers résultats collectés jusqu'ici m'ont en outre montré combien les textes anciens mettent en œuvre des micro-systèmes qui

leur sont propres (au moins en ce qui concerne la syntaxe du pronom sujet). Un enjeu majeur consiste précisément à dépasser ces derniers, sans les nier, pour essayer de mettre au jour les caractéristiques communes, si tant est qu'il s'en trouve. Cela suppose d'étudier un nombre conséquent de textes variés, afin que le corpus ainsi constitué puisse prétendre à une relative complétude au regard de l'objectif visé.

(Prévost 2002) s'appuyait sur les acquis de (Prévost 2001) en élargissant quelque peu le corpus, et en envisageant l'évolution de la syntaxe du sujet pronominal dans la perspective de la grammaticalisation. Plus précisément, l'étude a montré que la fixation progressive du sujet pronominal devant le verbe correspond bien à un cas de grammaticalisation. En effet, le pronom lui-même connaît un processus de cliticisation ; sa position, qui était motivée par des facteurs pragmatiques, est dorénavant contrainte du point de vue syntaxique, et elle constitue un indice de la fonction sujet ; enfin, corrélat direct, la position préverbale, qui était celle du topique, devient celle du sujet. Aussi bien le contenu de la position préverbale que la position du sujet pronominal répondent désormais à des contraintes grammatico-syntaxiques, et non plus pragmatico-informationnelles. Il s'agit donc, dans une large mesure, d'une fixation des stratégies discursives en des structures morpho-syntaxiques, type de grammaticalisation initialement envisagé par Givón (1971, 1979), et largement reconnu depuis.

Après un détour, pendant quelques années, et pour quelques articles, par la périphérie gauche de la phrase (étude des marqueurs de topicalisation et des marqueurs discursifs), je suis revenue récemment au cœur de celle-ci en me concentrant à nouveau sur le sujet pronominal. (Prévost 2010) et (Prévost sous presse) ont apporté quelques éléments inédits sur la position du sujet en ancien français, ainsi que d'autres, encore assez grossiers, sur son expression.

L'étude ici proposée part de faits partiellement explorés, tout en présentant un caractère inédit : elle vise à approfondir les résultats déjà obtenus, et à envisager les faits étudiés dans une nouvelle perspective, en mettant en relation l'étude de certains d'entre eux et en adoptant une approche quantitative qui dépasse les simples calculs fréquentiels utilisés jusqu'ici. Plus précisément, il s'agit d'évaluer les modalités de réalisation de l'expression du sujet (expression/non-expression) et de sa position (antéposition/postposition au verbe) en fonction d'un critère interne, celui de sa

personne. Pour cela, il a été adopté un traitement différencié des personnes du « discours » (‘je’, ‘tu’, ‘nous’, ‘vous’[4]) et de celles du « récit » (‘il’, ‘elle’, ‘ils’, ‘elles’), afin de déterminer s’il existe ou non des préférences distinctes en matière de position et d’expression de Sp. Cela doit en particulier permettre de tester sur corpus l’hypothèse avancée par Detges (2003) selon laquelle le développement de l’expression du sujet se serait fait à partir de la première personne, pour des raisons d’expressivité[5].

Il s’agit donc d’évaluer, sur le plan quantitatif, les possibles relations, voire corrélations, entre les évolutions de l’expression et de la position du sujet pronominal, plus spécifiquement en ce qui concerne le recul de son inversion et de sa non-expression, cela en fonction de la personne, mais aussi au regard d’un certain nombre de critères externes : la date bien sûr, mais aussi le dialecte, le genre des textes, et leur forme, versifiée ou non. Il se peut que la démarche fasse émerger une nouvelle classification des textes, qui ne soit plus uniquement fondée sur leurs caractéristiques externes, mais aussi sur certains traits linguistiques[6].

Cette étude est, aussi, l’occasion de poursuivre et d’illustrer sur un cas concret la réflexion entamée sur la constitution d’un corpus et sur le traitement des données.

A l’origine de ce projet, il était prévu qu’une analyse qualitative complète l’approche quantitative. Je suis en effet convaincue – comme beaucoup – que l’exploration et l’analyse des données, même partielle, doit accompagner la dimension strictement quantitative, au moins en ce qui concerne les questions qui relèvent du domaine syntactico-sémantique, comme c’est le cas pour celle-ci. Il s’agissait d’évaluer si l’élargissement du corpus mettait en cause ou non les caractéristiques des constructions ‘verbe-sujet pronominal’ observées jusqu’ici, en ancien et en moyen français. Il s’agissait aussi de mettre au jour certaines des caractéristiques des constructions sans sujet exprimé. Pour les deux types de constructions, le but était de dégager les modalités selon lesquelles les contextes qui autorisent ces séquences ont régressé. Mais l’approche méthodologique et l’analyse quantitative ont

[4] Finalement réduites au seul pronom de première personne : voir 3.2.1.
[5] Idée d’ailleurs déjà formulée, en d’autres termes, par Foulet (1930/1965 : 345-356).
[6] Cette approche a à cet égard des affinités avec la démarche inductive de D. Biber (1988 en particulier). Partant d’un regroupement assez lâche de ‘registres’ (sortes de genres), Biber définit des ‘dimensions’ (constellations de traits linguistiques) mobilisées différemment selon les registres, dimensions permettant de regrouper en ‘types’ les textes qui les utilisent de la même façon. Dans la présente étude il s’agit, plus simplement, d’essayer de dégager des regroupements entre textes selon certains traits linguistiques (position et expression du sujet pronominal).

progressivement occupé une place croissante dans ce travail, laissant peu d'espace et de temps pour une étude des contextes de recul de l'inversion et de la non-expression du sujet. Ce ne sont donc finalement que quelques éléments d'analyse qui seront esquissés à l'issue de l'étude quantitative. Ils viendront étayer les pistes de recherche qui seront présentées pour l'exploration ultérieure de cette dimension.

Dans un premier temps, je rappellerai en quoi la syntaxe du sujet pronominal constitue, dès l'ancien français, un objet assez complexe. Complexe dans ses modalités de réalisation, ce dont ne rendent le plus souvent pas compte des descriptions simplificatrices, et complexe aussi parce que les explications proposées pour en rendre compte ne sont pas pleinement satisfaisantes. Je reviendrai ensuite sur le fait de travailler « en corpus », propos délibérément laissé en suspens dans le document de synthèse présenté conjointement à cette recherche. J'évoquerai plus précisément la question de l'extraction et du traitement des données, celle de la constitution du corpus de travail, ainsi que les apports des donnés numérisées et des outils informatiques. Cette seconde partie nous conduira, dans un troisième temps, à la présentation de la méthodologie adoptée pour la présente étude, tant en ce qui concerne le choix des textes étudiés que la collecte des données et leur traitement. Je proposerai enfin, dans une dernière partie, l'analyse quantitative des données et leur interprétation. Une dernière partie, prospective, proposera quelques pistes pour l'analyse, sur le plan qualitatif, des constructions à sujet inversé et à sujet non-exprimé.

Chapitre 1. Evolution de l'expression et de la position du sujet pronominal : une question complexe

Je ne présenterai pas une revue de l'art des travaux portant sur la syntaxe du sujet pronominal. Plus modestement, je présenterai les caractéristiques majeures du sujet pronominal en ancien français, et son évolution, pour envisager ensuite certaines des explications qui ont été proposées, et les points sur lesquels elles achoppent.

En préambule, je rappellerai en quelques mots la situation en français moderne, en soulignant les trois caractéristiques qui distinguent le fonctionnement actuel du sujet pronominal de celui qu'il avait en ancien français.

Aujourd'hui le sujet pronominal est toujours exprimé, sauf dans les cas de coordination immédiate entre prédicats de même personne et de même temps (*elle se leva, prit son sac et sortit*). Il occupe par ailleurs très majoritairement la position préverbale, les inversions ne se rencontrant qu'en présence, en position initiale de la phrase, de certains adverbes épistémiques ou argumentatifs (*peut-être, sans doute, probablement, aussi, aussi bien, en vain, à peine, au moins, du moins, au pire, à peine, tout juste, tout au plus, ainsi, de même, pas davantage, encore, toujours…*[7]), qui ont en commun, comme l'a montré Guimier (1997), de déclencher une mise en balance de la prédication, et de conférer ainsi un caractère non pleinement assertif à l'énoncé (*Paul est très fatigué en ce moment :* ***aussi a-t-il décidé*** *de renoncer à son voyage ;* ***peut-être changera-t-il*** *d'avis dans quelques jours* ; énoncé construit). Enfin, troisième caractéristique, le sujet pronominal, quelle que soit sa position, est clitique : il ne peut former un groupe syntaxique autonome. Hormis dans le cas de la tournure figée « *je, soussigné Léon Petit, déclare,…* », ce sont les formes compléments des pronoms (*moi, lui…*) qui apparaissent dans les contextes de disjonction d'avec le verbe : *lui qui pensait partir en vacances ce soir (il) est furieux contre les grèves* (énoncé construit).

Ces trois caractéristiques distinguent nettement le français moderne à la fois de la

[7] Il s'agit d'adverbes qui ne tolèrent pas l'inversion nominale : **Sans doute connaîtra Paul ses résultats demain*, mais qui supportent en revanche l'inversion complexe : *Sans doute Paul connaîtra-t-il ses résultats demain*.

majorité des autres langues romanes[8] et de l'ancien français.

1.1. Caractéristiques de l'ancien français

L'ancien français se caractérise par un ordre des mots globalement plus souple qu'en français moderne, au sens où il n'est pas régi par les fonctions syntaxiques : l'objet nominal peut se trouver en position préverbale, et le sujet en position postverbale. En ce qui concerne plus précisément le sujet, l'ancien français présente, assez naturellement, les caractéristiques d'une langue romane. D'une part la non-expression du sujet y est fréquente, c'est une langue à sujet nul[9]) :

(6) Li rei Marsilie esteit en Sarraguce.
Alez en est en un verger suz l'umbre.
Sur un perrun de marbre bloi se **culchet**; (*Chanson* de *Roland*, vers 1100),

et elle autorise d'autre part l'inversion « romane » : le sujet, qui ne peut être que nominal dans ce cas, suit l'ensemble des formes verbales (ainsi que la négation *pas*) :

(7a) Tout einsint **ont anonciee li hermite et li saint home** vostre venue plus a de vint anz (*Queste del Saint Graal*, vers 1220)

(7b) bele buce, bel vis, bele faiture,
Cum est mudede vostra bela figure ! (*Vie Saint Alexis*, 1050)

Plus étonnamment, l'ancien français présente aussi les caractéristiques d'une langue germanique. Le verbe y est en seconde position, précédé d'un élément tonique (c'est une langue dite « V2 »), et l'inversion « germanique » y est courante : dans ce cas, le sujet, nominal ou pronominal, suit immédiatement la forme conjuguée du prédicat verbal, et précède les formes non conjuguées, comme en (8a-b) :

(8a) Si **a li rois** einsi **atendu** des le tens Josephe jusqu'à ceste hore (*Queste del Saint Graal*)

(8b) … et einsi **furent il destruit** par l'anemi et par son amonestement (*Queste*)

[8] Voir (Prévost, 2011b) pour la présentation de quelques éléments contrastifs.

[9] Je n'utiliserai pas ici le label « Pro-drop », introduit dans le cadre de la grammaire générative (à ma connaissance par Perlmutter (1971)), mais ayant largement débordé de ce cadre théorique depuis. La notion de « Pro-drop » réfère, en toute rigueur, non seulement à la possibilité de ne pas exprimer le sujet, mais aussi à la possibilité d'extraire les sujets des subordonnées gouvernées par un complémenteur explicite. Kayne (1980) a par ailleurs montré que la classe de langues qui autorisent les sujets nuls et le type d'extraction sus-mentionné autorisent aussi l'inversion « libre » du sujet, contrairement aux langues qui ne supportent pas les sujets nuls, comme le montre l'opposition entre les deux exemples suivants, empruntés à l'italien et au français : *Sono arrivati molto amici/ *Sont arrivés beaucoup d'amis*. C'est la conjugaison de ces différentes propriétés qui a été érigée en paramètre sous le label 'Pro-Drop'.

L'ancien français ne peut cependant pas être considéré comme une langue strictement V2, du moins dans le cadre d'une syntaxe de surface. En effet, d'une part, l'inversion germanique n'est pas la seule possible pour le sujet nominal, et d'autre part la contrainte du verbe en seconde position n'est pas absolue : on rencontre, bien que rarement, des occurrences de verbe en première position (9a-b), en particulier au 12ème siècle dans les chansons de geste, ou en troisième position (10a-d) :

(9a) **Plurent** Franceis pur pitet de Rollant (*Chanson de Roland*)

(9b) **Ot** le Gillelmes, s'en a un ris gité (*Le Charroi de Nimes*, 12ème s.)

(10a) **Li quens Rollant *Gualter de l'Hum* apelet** (*Chanson de Roland*)

(10b) Veez m'espee, ki est e bone e lunge :
***A Durendal* jo la metrai** encuntre (*Chanson de Roland*)

(10c) 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** (*Tristan* de Beroul, fin 12ème)

(10d) …et gardez qu' il ne soit a nul home mortel conté que vos l'aiez veü en ceste voie, ne **ge *endroit moi* n'en parlerai** ja. (*Mort Artu*, 1230)

Il existe par ailleurs des différences phonétiques et syntaxiques importantes selon la position du sujet pronominal. En position préverbale, le pronom peut porter l'accent et être disjoint, comme en (11) :

(11) Et **ele *tant* le conforta**
[….] Que ele en santé le remist (*Jehan et Blonde*, vers 1230)

Il peut aussi être coordonné et déterminé :

(12a) **Jou *et mi homme*** nous voulons vengier d'aus (Clari, *la Conqueste de Constantinople*, après 1205)

(12b) Et **je, *ki ai apris autres salus de cevaliers et autres acointements,* m'en retournerai** au roiaume de la Petite Bretaigne (*Tristan en prose*, après 1240)

En revanche, en position postverbale, le sujet pronominal est conjoint, il ne peut être séparé du verbe par un élément autre qu'un pronom régime conjoint, cas qui se rencontre d'ailleurs assez peu, et principalement, semble-t-il, en proposition interrogative :

(13) **Quidés *me* vos** comme lievre escaper ? (*Aliscans*, fin 12ème, cité par Skårup (1975 :47))

Il n'est séparable du verbe par un élément autre qu'un pronom conjoint que s'il appartient à un groupe, comme en (14)[10] :

(14) or **avront *garnemenz* il *e si cumpaignon*** (*Roman de Horn*, 1394)

1.2. Du 12ème au 17ème siècle : le sujet exprimé s'impose en position préverbale

Trois changements majeurs se sont produits dans l'histoire du sujet pronominal : il s'agit du développement de son expression, de celui de sa position préverbale, ainsi que de sa cliticisation en toutes positions. On a coutume de considérer que ces trois évolutions se sont déroulées en partie conjointement. C'est vrai si l'on envisage une très large diachronie qui irait du 12ème au 17ème siècle. En effet, alors que le pronom est majoritairement omis au début de cette période, que sa position postverbale est associée à des contextes relativement variés et que les cas de disjonction d'avec le verbe sont fréquents, on peut considérer que, au début du 17ème siècle, l'expression est majoritaire, et que les cas d'inversion et d'autonomie syntaxique, non rarissimes au 16ème siècle, le sont désormais devenus. Ils sont d'ailleurs largement proscrits par les grammairiens et remarqueurs. Il faut signaler à cet égard une évolution assez nette au début du siècle : à la relative tolérance de Maupas en 1618 en matière de non expression, et surtout de postposition du sujet (liste assez longue d'adverbes supportant l'inversion) succède les restrictions et les condamnations de Vaugelas en 1647 (voir Fournier (1998 : 36). Il ne faut pas pour autant en conclure que leur action

[10] Skårup (1975 : 47-50) cite de rares cas où un élément autre qu'un pronom régime conjoint peut s'intercaler entre le verbe et le pronom postverbal ; il s'agit principalement des adverbes *or*, *donc*, *tost*, *ja*, *mie*, *point*, (exceptionnellement un infinitif), occurrences qu'il explique par des exigences de versification dans le cas de *point* : *« Ma fille, et quele devise esc**e** ? / Encore ne le cognois point j**e**,/si le saroie volontiers »*, Froissart, *Meliador*, fin 14ème), ou plus fréquemment comme des erreurs de copistes. Cela reste un phénomène extrêmement marginal. Voir à ce propos l'exemple (47), en 4.1.6.2., extrait des *Miracles* de Coinci (début 13ème).

a été décisive : non pas qu'il faille se poser ici la question de l'effectivité de la norme édictée, mais, plus simplement, les règles formulées n'ont fait qu'entériner un usage dominant. Comme le rappelle N. Fournier (1998 : 8) :

> On leur a beaucoup reproché, surtout aux remarqueurs, leur visée normative. C'est oublier que leur norme est une norme d'usage, et qu'ils ne recommandent que ce qui leur paraît le plus répandu ; leurs jugements d'acceptabilité sont ceux d'un locuteur, qui se veut, comme le dit Vaugelas, et pourquoi ne pas le croire ? « simple témoin, qui dépose ce qu'il a vu et ouï » (Préface des *Remarques*, I).

Signalons qu'il semble s'être produit, au siècle précédent, une sorte de sursaut de la non-expression du sujet. Beaucoup ont relevé ce phénomène, dans lequel Brunot (1905-1938, vol.1 : 474) voit la conséquence d'une influence latine, qui a ralenti une évolution qui aurait été sinon plus rapide. Mais aucune preuve quantitative n'est venue jusqu'ici étayer ces affirmations, qui, aussi convergentes qu'elles soient, n'en mériteraient pas moins une confirmation quantifiée.

1.3. Un phénomène complexe à différents égards

Le « fonctionnement » du sujet pronominal (et son évolution) constitue un phénomène complexe dans la mesure où il comprend différents aspects : l'expression, la position, mais aussi l'accentuation, tous trois certainement liés, mais qui n'en comportent pas moins leurs spécificités et leurs modalités de description propres.

Il est complexe aussi – et cela découle en partie de ce qui vient d'être dit – parce qu'il relève de différents domaines : du domaine syntaxique bien sûr, tant en ce qui concerne l'expression que la position, du domaine prosodique, le pronom étant différemment accentué selon sa position, mais aussi du domaine sémantico-pragmatique et textuel. Ce dernier aspect a sans doute été moins exploré, hormis en ce qui concerne l'expression du sujet, que l'on tend à corréler à des phénomènes de (dis)continuité thématique (nous y reviendrons). Complexe, la question du sujet pronominal l'est aussi parce qu'elle est indéniablement liée à celle du sujet en général, qu'il s'agisse des sujets nominaux ou des autres pronoms : le sujet pronominal n'a pas fonctionné en ancien français (puis évolué) tout seul dans son coin.

Il est certes rare qu'un phénomène linguistique soit simple : bien souvent il relève (ou du moins l'analyse que nous en faisons) de plusieurs domaines (syntaxique, phonétique, sémantique…). Il est pareillement peu fréquent qu'une construction soit

coupée du reste du système, aussi bien dans ses phases de stabilité que dans celles d'évolution : elle appartient généralement à un paradigme.

Il n'en demeure pas moins que le fonctionnement du sujet pronominal, et partant, son analyse, présente un degré de complexité particulièrement élevé, ce qui explique peut-être qu'un certain nombre de questions n'ont pas encore trouvé de réponse pleinement satisfaisante. Je laisserai ici de côté les relations qu'entretient le sujet pronominal avec les autres sujets, bien consciente néanmoins qu'une étude complète de la question du sujet pronominal se devra(it) de prendre en compte l'évolution de l'ensemble des sujets. Mais ce n'est pas le propos du présent travail, qui ne prétend pas à un traitement exhaustif de la question. Je rappellerai simplement que, alors que l'inversion pronominale a toujours été un phénomène assez rare, la postposition du sujet nominal était au contraire très fréquente en ancien français, et encore en moyen français (voir Prévost 2001), pouvant représenter dans certains textes plus de la moitié des occurrences de sujets nominaux exprimés. Par ailleurs, il a été avancé (par Foulet entre autres) que la chute de la déclinaison nominale aurait provoqué la fixation positionnelle des sujets nominaux, laquelle aurait entraîné celle des sujets pronominaux. Nous reviendrons sur la question de la chute de la déclinaison plus loin.

On s'en tiendra ici aux aspects qui concernent spécifiquement le sujet pronominal, qui le constituent en tant que phénomène, à savoir, son expression, sa position, et son accentuation. A elles seules ces trois dimensions soulèvent plusieurs questions, en particulier en ce qui concerne de possibles relations de cause à effet dans le processus d'évolution qu'a connu le pronom : la cliticisation a-t-elle provoqué l'accroissement de l'expression ? Est-ce au contraire parce qu'il était davantage exprimé que le pronom s'est cliticisé ? L'expression accrue s'est-elle faite conjointement avec celle du développement de la position préverbale ? Ou bien l'une a-t-elle précédé l'autre, et l'a-t-elle du coup provoqué, ou au moins accéléré ? Les questions qui se posent concernent à la fois l'évolution qu'a connu le pronom et l'état de relative stabilité qui a précédé, en ancien français.

Ainsi, si la non-expression du sujet en ancien français trouve une explication assez consensuelle dans le caractère saillant du référent non exprimé, les raisons qui motivent la postposition du sujet pronominal restent plus obscures. En ce qui concerne l'évolution du pronom, la question du « pourquoi » reste encore largement ouverte, de même que celle du « comment ». Déterminer les causes d'une évolution est souvent

difficile, l'interprétation risquée. Cela suppose de se fonder sur une analyse exhaustive des données, lesquelles doivent être aussi nombreuses que possible. On ne peut répondre à la question du « pourquoi » sans avoir répondu précisément à celle du « comment », au moins pour le phénomène qui nous intéresse. Je ne prétends pas apporter de réponses, dans le cadre de cette étude, à l'ensemble de ces questions, en particulier à celle des causes de l'évolution : cela supposerait non seulement la prise en compte et l'analyse fine de données plus nombreuses que celles étudiées ici, et sur une période plus large, mais aussi une mise en relation avec les autres types de sujets. Plus modestement, j'essaierai de préciser les modalités du « comment », en fournissant des données quantifiées systématiques. Cette étude est conçue comme une contribution à une connaissance accrue de l'évolution du sujet pronominal, elle constitue une étape à une meilleure compréhension du phénomène.

Je rappellerai brièvement quelques-unes des explications qui ont été proposées pour rendre compte de l'expression et de la position du sujet pronominal ainsi que de l'évolution de celles-ci. Cette présentation est partielle (et donc partiale), mais elle tente d'identifier les tendances majeures et les points de discorde.
Je ne dirai que quelques mots de l'accentuation des pronoms et de la cliticisation qu'ils ont subie en position préverbale, laissant pour l'instant de côté l'approfondissement de cette question.

1.4. La cliticisation

Alors qu'il est admis de manière consensuelle que le pronom postverbal n'était pas accentué (au moins lorsqu'il était seul : la question peut se poser dans un exemple tel que (14)), il est plus difficile d'établir la nature de son accentuation en position préverbale, et ce avant même les débuts de sa cliticisation. Franzen a souligné la complexité de l'accentuation des pronoms personnels sujet (1939 : chapitres 1-3)[11], et il suggère l'existence de variations accentuelles dues à leur sens ou aux conditions prosodiques dans lesquelles ils apparaissaient (1939 : 8).

L'évolution ultérieure est bien connue : peu à peu le pronom va se cliticiser en toutes positions. Les divergences en matière de datation n'en sont pas moins grandes. Ainsi,

[11] Franzen (1939 : 36) met en question le fait que les pronoms détachés du verbe aient été systématiquement accentués.

selon Moignet (1973) et Adams (1987), qui se fondent sur l'élision précoce de « je », la cliticisation aurait débuté dès le 13ème siècle. En revanche, Dufresne (1995) estime que ce n'est qu'au 14ème siècle que le pronom a atteint le statut de clitique, passant au préalable par celui de mot fonctionnel.
Certes, comme l'a remarqué Foulet (1930 : 152), on observe les premières occurrences de formes régimes fonctionnant comme sujet dès la fin du 12ème siècle :

(15a) S'irons tornoier **moi et vos** (*Yvain* de Ch. de Troyes, vers 1177)

(15b) **Moi et vos** fumez en une hore engendré (*Ami et Amile,* 1200)

Il n'est pas exclu d'y voir le premier signe d'une perte de prédicativité du pronom sujet. Il est vrai aussi que s'opère un tournant vers 1200. A partir de cette époque disparaît en effet la restriction connue sous le nom de 'loi Tobler-Mussafia', qui consiste en l'exclusion des pronoms clitiques en position initiale de proposition (et donc la proclise, dans les constructions à verbe initial). On peut désormais trouver des pronoms régimes atones en tête de phrase. Dès lors que la position initiale supporte des éléments non toniques on peut faire l'hypothèse que les pronoms sujets qui occupent cette position ne sont plus nécessairement accentués.

On le voit, il reste encore beaucoup d'inconnues pour dater précisément la cliticisation des pronoms sujets, la seule certitude étant que c'est une évolution qui s'est faite de manière progressive, conjointement à l'émergence des formes régimes en fonction sujet. Il est fort probable que l'affaiblissement des pronoms sujets soit lié au développement de leur expression, sans que soit pour autant avérée une relation de cause (l'affaiblissement a-t-il favorisé l'expression, ou bien l'expression accrue a-t-elle provoqué l'affaiblissement ?).

En ce qui concerne le rapport entre cliticisation et position du pronom personnel sujet, Buridant (2000 :746) estime que la cliticisation aurait contribué à soustraire le sujet pronominal à la position postverbale au profit de la position préverbale quand la position initiale était occupée[12]. Cette affirmation postule une règle V2 très stricte, qui aurait précédemment interdit au pronom personnel d'occuper la position préverbale conjointement avec un autre élément. Or nous savons que l'ancien français n'était pas une langue strictement V2, ni au 12ème ni au 13ème siècle. Nous avons déjà mentionné

[12] Hypothèse déjà avancée par Zwanenburg (1974 et 1978).

plusieurs exemples de verbe en troisième position (10a-d), en voici une autre, avec en tête un objet nominal suivi du sujet pronominal :

(16) mesire Gynglayns […]me dist k'il l'avoit veü tout forsené dedens le Marés. ***Autres nouveles*** **je ne sai** orendroit de lui (*Tristan* en prose).

Il convient donc, sur la question de la relation entre cliticisation et position du sujet pronominal, de demeurer prudent, et d'envisager, en l'état actuel de nos connaissances, de simples convergences entre les différents changements plutôt que des relations de causalité fortes.

Je laisserai cette question de côté pour me tourner plus spécifiquement vers la position et l'expression du sujet pronominal.

1.5. La position du sujet pronominal

La postposition du sujet pronominal est rare dès les plus anciens textes du français, sa fréquence toujours largement inférieure à celle du sujet nominal. C'est un fait unanimement constaté, très tôt. Brunot (1905-1938, vol.1), Foulet (1930), Franzen (1939) : tous signalent le caractère marginal du phénomène,… et tous expédient assez rapidement l'affaire. Peut-être la rareté du phénomène n'incite-t-elle pas à s'attarder, d'autant qu'une explication semble s'imposer : l'ancien français étant une langue à verbe second, dès lors qu'un élément autre que le sujet occupe la première position, celui-ci est inversé. Signalons néanmoins que Franzen, dans son étude sur les pronoms personnels sujets, évoque la possibilité que l'inversion marque une nuance de sens, même s'il est souvent difficile de déceler une intention spécifique et que « l'usage semble le plus souvent arbitraire » (1939 :132)[13]. Des années plus tard, Moignet (1973) n'évoque pas la question spécifique de la position du pronom, tandis que Ménard (1988) y consacre un court paragraphe, soulignant que la présence en tête de certains compléments « entraine 'l'inversion du sujet' » (1988 : 52). Quant à la grammaire de Buridant (2000), autant elle se montre prolixe sur la question de l'expression, autant elle est passablement expéditive sur la question de la position, « conditionnée par les paramètres syntaxiques, dont l'ordre TVS, avec adverbe,

[13] Plus loin, Franzen insiste à nouveau sur la difficulté à rendre compte de l'inversion (1939 :139). De fait son étude s'attache bien plus à la question de l'expression ou non du sujet pronominal, ainsi qu'à celle de l'accentuation.

régime objet, ciconstanciel, etc. »[14] (2000 : 432). Skårup (1975) formule les choses un peu différemment, dans le cadre de son modèle positionnel à 3 zones (zone préverbale, verbale et postverbale – auxquelles s'ajoute dans certains cas, à gauche de la zone préverbale, l'extraposition). Dans ce modèle, le sujet pronominal postverbal se trouve dans la zone verbale (contrairement au sujet nominal qui se trouve dans la zone postverbale), tandis que le sujet préverbal, nominal ou pronominal, se trouve dans la zone préverbale. Skårup observe que, alors qu'un sujet nominal peut suivre un verbe qui n'est précédé d'aucun élément (au moins jusque la fin du 12ème siècle), cela n'est pas possible pour les sujets pronominaux, qui ne peuvent suivre le verbe que si celui-ci est précédé d'un autre élément[15]. Ce dernier se trouve dans la zone préverbale, où il occupe plus spécifiquement la place du 'fondement'. La postposition du sujet pronominal dans les déclaratives et les subordonnées indique donc que la place du fondement est occupée (par un élément qui peut d'ailleurs occuper d'autres positions dans la proposition). Skårup fournit (chapitre 4) une liste sinon exhaustive, en tout cas très complète, des éléments susceptibles d'occuper la place du fondement. D'une manière générale, Skårup constate, plus qu'il n'essaie d'expliquer, comme il en convient dans son introduction[16]. *In fine*, même si elle s'inscrit dans une vision *schématisée* de la proposition, l'approche de Skårup n'est pas très éloignée de celles qui ont été précédemment évoquées.

L'analyse proposée dans le cadre de la grammaire générative est différente dans la mesure où, s'il y a bien déplacement, ce n'est pas le sujet qui est affecté. Adams (1987) a proposé l'explication suivante pour rendre compte de l'inversion dans les propositions « racines » (= indépendantes/principales) : après avoir reçu sa flexion et ses traits sous I (Inflexion), le verbe ainsi enrichi remonte jusqu'à la position C de CP (Complementizer Phrase). En raison de la contrainte V2, la position Spec CP (Spécifieur de CP) ne peut rester vide (sous peine d'avoir le verbe en première position) : un élément se déplace donc pour venir l'occuper. Il peut s'agir d'un adverbial, d'un régime, du sujet, etc. (voir schémas dans Vance (1997 : 15-16).

[14] Dans TSV, T= topique/thème.

[15] Elément qui provoque en outre l'antéposition au verbe des pronoms régimes : on retrouve ici, formulées en d'autres termes, les conséquences de la loi Tobler-Mussafia.

[16] « Ainsi, nous dirons que tel membre peut occuper telle et telle place, nous n'examinerons pas les facteurs qui déterminent le choix entre les places qu'il peut occuper, ni les facteurs qui déterminent le choix entre les membres et (zéro) qui peuvent occuper chaque place. Notre essai ne constitue que les prolégomènes des études de ces facteurs et de leur importance relative. » (Skårup, 1975 : 6).

On retiendra prioritairement de cela que, dans ce modèle, la position postverbale du sujet est sa position de base, après que se soit produite la montée du verbe en CP ; et ce n'est que si c'est le sujet (*versus* adverbial, régime, …) qui vient occuper la position Spec CP que la linéarisation de surface correspondra à l'ordre SV(X)[17], avec sujet préverbal.

C'est la directionalité du gouvernement qui rend possible l'hypothèse avancée par Adams : on a ici un gouvernement qui s'opère de gauche à droite (et qui correspond à un choix paramétrique[18]), ce qui est permis par la richesse de la morphologie verbale. Lorsque le verbe perdra sa richesse verbale, et qu'il ne sera donc plus assez « fort » pour attribuer le cas par gouvernement, cette attribution se fera par accord ('agreement') : le verbe ne remontera plus jusqu'à C, il s'arrêtera à I (Vance 1997), tandis que le sujet remontera à sa gauche, dans Spec I, où son cas lui sera assigné.

L'explication de l'inversion du sujet pronominal comme un simple automatisme syntaxique dès lors qu'un élément occupe la première position de la phrase – ou du moins le constat que le sujet postverbal est toujours associé à la présence d'un élément initial – présente néanmoins des faiblesses. La première tient à ce que, si le verbe occupe très majoritairement la seconde position en ancien français, cela ne constitue par pour autant une règle absolue. Nous avons déjà évoqué les occurrences de verbes en première position, et surtout, pour ce qui nous intéresse ici, en troisième position. Nous en avons vu des exemples en 10(b-d) et en (16), en voici un nouvel exemple :

(17) Sire nos volons que vos aiez vostre conseil ; et ***devant vostre conseil*** **nos vos dirons** ce que nostre seignor vos mandent (Villehardouin, *Conqueste de Constantinople*, vers 1200)

On voit que l'inversion du sujet pronominal n'est pas systématique, et qu'elle ne constitue donc pas un automatisme syntaxique. Sans pour autant rejeter le rôle de la contrainte du verbe en seconde position, notre hypothèse est qu'elle ne suffit pas à expliquer l'ensemble des inversions, et que celles-ci répondent, souvent, à des motivations sémantico-pragmatiques. Cela signifie donc aussi que l'inversion n'est pas arbitraire.

[17] X : n'importe quel élément.

[18] On rappellera que, dans le cadre de la grammaire générative, les 'paramètres' sont des propriétés variantes entre langues ou états de langue, alors que les 'principes' sont des propriétés invariantes (entre langues ou états de langues).

Pour ce qui est du développement des séquences 'sujet-verbe', on rappellera tout d'abord que l'essor de l'ordre 'sujet pronominal-verbe' s'inscrit dans un mouvement général de fixation de l'ordre des mots, qui a d'abord affecté l'objet nominal (voir Combettes 1988 et Marchello-Nizia 1995), avant de toucher le sujet. Pendant longtemps, on a expliqué la fixation de l'ordre des mots, en particulier celle du sujet devant le verbe, par la chute de la déclinaison et le recul de la contrainte du verbe en seconde position. Assurément, cette explication n'est pas à rejeter, mais elle est lacunaire pour plusieurs raisons : outre l'existence de verbes en première et en troisième positions, il se trouve que bon nombre de substantifs féminins ne se sont jamais déclinés ; pour les autres, la déclinaison est déficiente dès le 13ème (époque à laquelle l'ordre des mots est pourtant encore peu contraint) ; à l'inverse la déclinaison des pronoms personnels s'est maintenue. Cette explication, insuffisante donc, a par la suite été complétée par le recours à un principe fonctionnel (voir en particulier Vennemann 1976, Combettes 1988). L'ordre des mots en français aurait connu un changement de principe organisateur : on serait passé d'un principe pragmatico-informationnel (formulé en termes de « topique-commentaire » ou de « thème-rhème »[19]) à un principe syntaxique. En effet, dans la mesure où le sujet, topique ou thème privilégié, occupait souvent la position initiale, il se serait peu à peu fixé en cette place.

D'un point de vue typologique, on observe donc le passage d'un ordre SXV en latin à un ordre SVX en français moderne, cela par l'intermédiaire de deux étapes : TXV puis TVX (aux 12ème-13ème siècles). Selon le principe informationnel en vigueur en ancien français, lorsque le sujet est postposé, cela signifie, soit qu'il n'est pas le topique, soit qu'il est porteur d'une charge informative élevée. J'ai rappelé en introduction les limites de cette analyse pour les pronoms personnels, en particulier *il,* ce qui conduit à réinterpréter de manière plus large la notion de charge informative élevée.

Mon postulat est qu'il existe depuis l'ancien français une spécificité des séquences à sujet pronominal inversé, spécificité qui a évolué au fil des siècles.

[19] La notion de *topique* est entendue comme « ce dont on parle », et celle de *thème* comme « élément porteur d'une faible charge informative », voir note 1.

1.6. L'expression du sujet

Plusieurs points sont à considérer en ce qui concerne l'expression ou la non-expression du pronom : la nature et la position de la forme non exprimée, les motivations de son expression (ou de son absence), et les raisons du développement de cette même expression.

1.6.1. Nature et position du sujet non exprimé

En ce qui concerne la nature des sujets non exprimés, on s'accorde à en faire, sur une base référentielle, des pronoms personnels. C'est un point que je ne contesterai pas, même si cela ne signifie pas pour autant que le référent associé au sujet non exprimé soit toujours immédiatement accessible (voir plus bas des exemples de discontinutié thématique).

Le relatif consensus qui prévaut en ce qui concerne la position du sujet non exprimé me semble en revanche beaucoup plus contestable. Il a en effet été proposé, dans des approches fort différentes, de les considérer presque toujours comme des pronoms postverbaux omis. C'est la position de Foulet (modulée néanmoins en cas de V1 : voir ci-après) :

> C'est là un point fondamental de syntaxe du vieux français : *l'inversion du sujet entraîne facilement dans le cas du pronom personnel l'omission du sujet* (Foulet 1930 : 313),

mais aussi de Skårup (1975), de Vance (1997), et de Buridant (2000), pour n'en citer que quelques-uns :

> Quand le sujet postposé est un pronom personnel, il est le plus souvent absent […] mais dans les textes versifiés, des facteurs métriques peuvent favoriser son expression (Buridant 2000 : 746)

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)[20], on en conclut que, s'il avait été exprimé, le sujet pronominal aurait suivi le verbe. C'est d'ailleurs au nom de ce même principe que Foulet (1930 : 322) postule un sujet préverbal lorsque le verbe est en position initiale.

[20] 'C' désigne tout complément de type objet direct ou indirect, attribut ou bien complément locatif essentiel.

Or, nous l'avons vu (10(b-d), 16 et 17), les séquences XSpV sont possibles en ancien français, même si elles sont rares.

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 : la rareté des séquences 'verbe-sujet pronominal' en fait au contraire des constructions marquées du point de vue quantitatif, et nous savons qu'elles le sont aussi qualitativement. Or les constructions à sujet non exprimé ne présentent pas un tel caractère marqué, ne serait-ce que par leur fréquence. Pour ces raisons, il me semble préférable de ne pas assimiler pronoms postverbaux et sujets non exprimés.

C'est d'ailleurs aussi, dans une certaine mesure, la position de Vance (1997 : chap. 5). Elle considère en effet que les sujets postverbaux et les sujets « nuls » sont *syntaxiquement* équivalents, mais qu'ils présentent des différences sur le plan pragmatique, lesquelles tiennent au rôle discursif de l'élément initial : CV n'est pas « marqué », alors que CVSp l'est[21], en ce que la séquence dénote un lien spécifique avec le discours environnant. Selon Vance, ce système de contraste serait une innovation de l'ancien français tardif, et se serait prolongé en moyen français, l'inversion pronominale étant plus rare dans les textes plus anciens[22] dans lesquels elle ne semble pas avoir de fonction particulière.

J'ai moi-même souligné (voir introduction pour un rappel de Prévost 2001) que des effets de « rupture », généralement d'ordre logico-pragmatique, accompagnent effectivement les séquences VSp en moyen français ; mais l'exploration, rapide, de quelques textes des 12ème et 13ème siècles a montré que, déjà à cette époque, on ne peut assimiler, sur le plan sémantico-pragmatique, les énoncés à sujet postverbal et ceux à sujet non exprimé.

1.6.2. Valeurs respectives de l'expression et de la non-expression

La non-expression du sujet se rencontre le plus souvent dans des contextes dans lesquels l'identification du référent ne fait pas difficulté, et qui dénotent une continuité

[21] Franzen évoque aussi une différence, à propos des verbes vicaires précédés de l'adverbe *si* : sans sujet la proposition dénote une opposition par rapport à ce qui précède (*si fait*), avec sujet elle traduit la conformité vis-à-vis de ce qui précède (*si fait il*).
[22] C'est à vérifier.

thématique. A l'inverse, l'expression du sujet signale généralement une discontinuité thématique (changement de temps ou de référent par exemple) ou une opposition, ou une insistance particulière (voir Buridant 2000 : 424-433). Cela n'est cependant pas systématique. On rencontre effectivement des situations de continuité thématique dans lesquelles le sujet pronominal est malgré tout exprimé, en particulier lorsqu'une subordonnée temporelle précède. On observe de ce point de vue des variations d'un texte à l'autre, mais aussi, de façon plus surprenante, au sein d'un même texte, comme en témoignent (18a) et (18b), extraits de *la Mort Artu* :

(18a) Quant Agravains se fu aperceüz de la reïne et de Lancelot, **il en fu liez** durement et plus por le domage que il cuida que Lancelos en eüst que por le roi vengier de sa honte (*la mort Artu*)

(18b) Et quant Agravains sot que Boorz s'en aloit et li chevalier avec lui et que Lancelos remanoit, **si pensa** tantost que c'estoit por la reïne ou il vouloit avenir, quant li rois s'en seroit alez. (*la mort Artu*)

A l'inverse, on peut trouver des sujets non exprimés dans le cadre d'une discontinuité thématique, même si cela reste rare :

(19) Puis **fist monter** ses compaignons
Et **portent** ostoirs et faucons. (*Bel Inconnu*, av.1214)[23]

(20) Dido la dame de Cartage
mar vit onques le suen ostage ;
il fist de li sa volenté ;
quant **el l'ot piece sejorné,**
si s'en torna[24] o son navire,
et el s'ocist a grant martire. (*Eneas*, v.3314)

Par conséquent, si la continuité thématique constitue un facteur explicatif assez probant pour rendre compte de l'expression du sujet, il apparaît que d'autres facteurs doivent néanmoins être pris en compte, tels que des effets de mise en relief du sujet, des contraintes métriques, etc. Il faut aussi admettre une part de variabilité intertextuelle ou intratextuelle qui ne se laisse peut-être pas appréhender en termes systématiques.

La question de la mise en relief est à vrai dire assez complexe : selon Moignet, l'expression de sujet dénote un effet d'insistance dans le très ancien français, mais cet

[23] Dans l'exemple cité, il se peut que le causatif *faire* entraîne une continuité thématique sur le sujet de l'infinitif.
[24] Le sujet de 's'en torna' est Enee, que Didon a hébergé ('séjorné' : vers précédent).

effet se perd dès le 13ème siècle, époque à partir de laquelle il considère que l'absence de sujet devient assez rare, au moins dans les textes en prose[25].

C'est un même effet d'insistance que met en avant Detges (2003) pour rendre compte du développement des sujets pronominaux exprimés. Il suggère que celui-ci se serait fait à des fins de stratégie discursive, dans des contextes de prise de parole et de prise de position, et donc, prioritairement, avec la première personne. La hausse de la fréquence dans de tels contextes aurait eu un effet de dévaluation rhétorique, qui aurait conduit à une généralisation de l'emploi, et donc à l'affaiblissement des pronoms[26]. Le mouvement serait donc parti de la première personne et se serait ensuite généralisé aux autres. De son côté, Foulet a pu observer une présence accrue de sujets pronominaux dans les œuvres dramatiques. Or, dans la mesure où ces dernières sont supposées plus proches de la langue parlée, il en conclut que : « en parlant on employait plus de pronoms personnels qu'en écrivant » (Foulet, 1930 : 327). Il convient néanmoins de rester prudent quant aux généralisations que l'on peut faire à partir des situations discursives (dialogues) des textes écrits.

Si l'idée que le développement des pronoms personnels sujets se serait fait à partir de la première personne est à la fois séduisante et plausible, il n'en convient pas moins de l'étayer en opérant des dénombrements systématiques des fréquences respectives des première et troisième personnes dans un nombre suffisamment important de textes pour essayer de dépasser les phénomènes de micro-systèmes propres aux textes.

Par ailleurs, l'idée d'une mise en relief du sujet par son expression a été critiquée très tôt par Franzen (1939), qui a montré que, dans la traduction anglo-normande d'une partie du *Livre des Psaumes* (manuscrit Arundel), ainsi que dans les *Quatre Livres des*

[25] Les sujets non exprimés représentent cependant encore plus de 50% de l'ensemble des sujets dans *la Queste du Graal*. Mais il y a probablement dans cette appréciation, possiblement variable, un effet de perspective comparable à la perception du verre à moitié plein ou à moitié vide. En effet, si le latin constitue le point de repère, on peut considérer que l'expression est fréquente dès le 12ème siècle, surtout comparée à de nombreux textes du siècle précédent. C'est certainement ce qui fait déclarer à Brunot (1905-1939, vol. 1 : 226), quelques lignes après avoir signalé l'absence encore fréquente, comme en latin, des pronoms, qu' « il est certain qu'à la fin du XIIe siècle, chez des écrivains consciencieux, comme Chretien de Troies, la régularité est déjà très grande ». En revanche, si l'on considère la situation depuis le français moderne, la fréquence de la non-expression semble encore bien élevée.

[26] On peut à cet égard établir une analogie avec ce que l'on a pu observer pour le latin, où les personnes 1 et 2 étaient exprimées pour des raisons d'affectivité ou d'insistance, ou d'opposition..., leur emploi devenant peu à peu mécanique dans la langue parlée dans certaines locutions (voir Ernout et Thomas, 1953).

Rois (traduction de la Vulgate)[27], le traducteur a multiplié l'expression des sujets par rapport aux textes d'origine[28], sans qu'aucun effet d'insistance ne le justifiât[29]. Selon Franzen, le traducteur a suivi en cela sa propre langue (il s'agit d'anglo-normand, mais il semble tout à fait possible, selon Franzen, d'extrapoler au français continental).

Effets d'insistance, mise en relief, expressivité : on sait qu'il s'agit de notions difficiles à manier, car souvent un peu floues, au moins quand elles se rapportent au plan sémantico-pragmatique. Elles revêtent en outre probablement des modalités différentes, selon les personnes affectées (1ère personne du locuteur / 3ème personne), et selon les époques considérées (très ancien français ou 12ème ou 13ème siècle).

Une approche différenciée des personnes 1 et 3 du pronom permettra peut-être, au moins pour la période des 12ème-14ème siècles, d'apporter quelques lumières.

Je terminerai par une remarque d'Herman, qui abonde dans le même sens que Franzen, en ce qu'il souligne que l'expression du sujet, dans ces textes, n'a rien à voir avec le besoin de distinguer les personnes :

> L'emploi presque régulier des pronoms sujets dans certaines positions, – d'une régularité rare dans nos textes – ne doit donc pas être attribué à un besoin accru de clarté dans la distinction des personnes. A plus forte raison serait-il exagéré d'attribuer l'extension de leur emploi à la décomposition des désinences verbales, encore bien vivantes, à notre avis, au XIIe siècle. (1954 : 87).

Cette remarque nous mène droit à la question de la richesse des désinences verbales.

1.6.3. Développement de l'expression du sujet : quelles explications ?

Deux explications ont été traditionnellement proposées pour rendre compte du déclin des sujets non exprimés (à laquelle il faut ajouter l'hypothèse pragmatique évoquée précédemment, formulée en termes de mise en relief, d'insistance). La première

[27] Franzen insiste sur le caractère moins fiable, pour sa démonstration, des *Quatre livres des Rois*, dans la mesure où il s'agit d'une traduction assez libre, et pour laquelle on n'est pas toujours bien sûr du manuscrit d'origine.
[28] Constat confirmé par (Herman 1954).
[29] Cependant, comme le souligne Herman, les *Psaumes* présentent un style assez emphatique, qui explique selon lui un emploi un peu plus fréquent dans ce texte des pronoms sujets comparé aux classiques latins. On ne peut donc complètement exclure que le traducteur ait voulu renchérir sur l'emphase du texte d'origine, et ait pour cela multiplié l'emploi des pronoms.

s'appuie sur l'influence germanique[30], qui aurait aussi joué un rôle dans le développement des sujets pronominaux dans les dialectes d'Italie du Nord ((Kuen 1970) et (Hilty 1975), cités par Detges (2003)). Toutefois, comme le souligne Detges (2003 : 311), cette explication perd de sa force si l'on considère l'actuel développement des pronoms personnels sujets en portugais brésilien et en espagnol de Porto-Rico, que l'on ne peut mettre sur le compte d'une quelconque influence germanique.

La seconde explication proposée, plus ancienne et plus répandue, est celle de la perte de la richesse morphologique verbale. L'argument souffre lui aussi de faiblesses. Ainsi, en dehors du français, on observe une relation non réciproque entre expression des sujets pronominaux et richesse verbale, comme l'illustrent l'italien du nord et le rhéto-roman (Detges 2003 : 311), qui associent richesse morphologique et expression du pronom, et à l'inverse, le japonais et le chinois, dépourvus d'une morphologie verbale riche mais qui se passent néanmoins facilement de l'expression du sujet (voir Dupuis 1989, chapitre 2).

Pour le français, plus spécifiquement, la question a été largement débattue sans aboutir à un consensus. Selon Foulet (1930 et 1935), c'est la richesse de la morphologie verbale qui autorisait la non-expression du sujet, et c'est conséquemment la perte de cette richesse qui, très tôt (dès le 12ème siècle) aurait conduit à son expression accrue (laquelle aurait entraîné sa cliticisation). C'est aussi la richesse de la morphologie verbale qui rend généralement[31] compte, dans le cadre de la grammaire générative, de la possibilité des sujets nuls : ceux-ci, en raison de la faiblesse de leur contenu sémantique, doivent être légitimés et identifiés par 'gouvernement'. C'est un changement dans l'assignation du cas (nominatif), lié à la disparition d'une morphologie riche, qui expliquerait le déclin des sujets nuls.

Explication ancienne, et proposée dans des approches différentes, comme on vient de le voir, l'explication morphologique n'en est pas moins critiquable, et critiquée, au moins dans sa version radicale. Dans son étude consacrée aux pronoms personnels sujets, Franzen montre que la chute des désinences verbales ne suffit pas à expliquer le développement de l'expression du sujet. La multiplication des sujets exprimés dans

[30] Le francique a ainsi influencé le proto-roman de Gaule à partir du 5ème siècle (invasion des Francs), sans toutefois le supplanter : c'est un *superstrat* du français.

[31] Dupuis (1989 : chap. 2), sans remettre en cause le principe général, émet néanmoins quelques réserves face à ce qu'elle juge être une caractérisation trop vague de la relation entre AGR (agreement=accord) riche et pro-drop.

des traductions de textes anciens est un premier argument contre cela (voir 1.6.2.). Franzen souligne en outre que le sujet est bien plus largement exprimé dans les propositions subordonnées… alors que les désinences verbales y sont les mêmes que dans les propositions principales. D'une manière générale, il observe que la distribution des pronoms exprimés ou non-exprimés est largement corrélée aux constructions dans lesquelles ils se trouvent, et que les éléments initiaux jouent à cet égard un rôle décisif. Franzen ne rejette pas complètement l'influence de l'amuïssement des désinences verbales, mais il en fait un facteur secondaire, qui n'aurait fait qu'accélérer et achever un mouvement déjà bien établi, et qui a abouti à l'expression obligatoire des sujets pronominaux. Buridant reprend la même idée, considérant que l'érosion phonétique n'a joué qu'un rôle de catalyseur (2000 : 438)[32].

Outre les différents arguments avancés, il faut bien aussi reconnaître que, aujourd'hui encore, nous ne maîtrisons pas suffisamment les chronologies respectives des différents phénomènes (développement de l'expression, érosion des désinences), qui restent encore largement discutées (voir Schøsler 1991).

Si cette étude vise à apporter quelques précisions pour ce qui est de la chronologie du développement des sujets pronominaux, elle n'essaiera pas de se prononcer sur les désinences verbales. On sait combien cette question est complexe, en particulier parce que l'enregistrement à l'écrit des changements phonétiques opérés à l'oral est souvent largement différé, et de toute façon grandement partiel (en témoignent nos graphies modernes). Cela signifie que les textes peuvent témoigner, pour certains aspects, d'un usage décalé dans le passé, tout en rendant compte pour d'autres aspects, d'un usage plus récent : il n'est pas extravagant de considérer que les textes se sont montrés conservateurs en ce qui concerne les aspects phonétiques (que nous pouvons en partie appréhender, pour les désinences, à travers les rimes), mais plus en prise sur la langue orale contemporaine pour ce qui concerne l'expression du sujet. Assurément il n'est pas aisé de déterminer une chronologie fine et précise de l'amuïssement des désinences verbales à une époque où les témoignages méta-linguistiques restent encore extrêmement rares.

[32] Toutefois, on peut contester l'argument avancé selon lequel, si la perte des désinences avait été décisive, les pronoms des personnes 4 et 5 auraient dû connaître un essor moindre de leur expression (ce qui n'est pas le cas : voir Schøsler (1991)). On sait en effet que la force de l'analogie peut être grande, en particulier quand elle trouve à s'exercer au sein d'un même paradigme, en l'occurrence celui des sujets pronominaux. Que les pronoms 'nous' et 'vous' se soient autant développés que les autres ne me semble pas être un argument très robuste contre l'hypothèse du rôle déterminant de la perte de la richesse verbale. Mais nous disposons par ailleurs d'autres arguments pour qu'un tel rôle puisse de toute façon être écarté.

Ce qui est avéré, en revanche, c'est le recul de la non-expression du sujet à partir du 13ème siècle, phénomène qui s'accélère en moyen français et au 16ème siècle. L'omission du sujet sera fortement condamnée par les grammairiens au siècle suivant.

On l'aura compris à la lecture des pages précédentes : il ne me semble pas souhaitable d'assimiler non-expression et postposition du sujet, et je préfère considérer qu'il existe en ancien français trois types de sujets pronominaux : ceux qui ne sont pas exprimés, ceux qui sont exprimés en position préverbale, et ceux qui le sont en position postverbale. Les derniers sont assurément les plus rares[33], les premiers sont les plus fréquents. Ce sont donc ces trois modalités du sujet qui seront examinées et confrontées entre elles par la suite, cela en fonction des deux paramètres qui les définissent – expression et position –, auxquels nous ajouterons celui de la personne. Nous présenterons en détail les critères d'analyse dans la partie 3.

Avant de présenter les spécificités méthodologiques de cette étude, il convient de revenir un peu sur le fait de travailler « en corpus », question évoquée dans la mémoire de synthèse.

[33] Mais ils sont néanmoins attestés, semble-t-il, dans tous les textes (à l'exception de *Eulalie* et des *Serments de Strasbourg*) : Price (1966) en relève ainsi des occurrences dans la *Passion du Christ*, la *Vie de Saint Légier* et la *Vie de Saint Alexis*.

Chapitre 2. Splendeurs et misères de la massification des données

J'ai évoqué ailleurs[34] les apports que constitue la massification des données, conséquence immédiate de la numérisation des textes. Il convient de revenir quelque peu sur cette question, en précisant en quoi consistent ces apports, et en évoquant aussi les possibles inconvénients de cette abondance des données.

2.1. Extraction et traitement des données

2.1.1. La collecte des données

Les textes numérisés ont permis une double massification des données : celle des données observées – les textes et leur contenu – et celle des données collectées pour une étude spécifique. La première autorise la seconde, mais elle ne l'implique pas nécessairement : pour qu'un corpus quantitativement important permette la collecte significative d'une construction, il faut évidemment que celle-ci soit fréquente[35], mais il faut aussi disposer des outils adéquats pour la repérer et l'extraire.

Pour évaluer la facilité du traitement des textes et des données, il convient de considérer trois paramètres. Le premier concerne le type de phénomène que l'on cherche : s'il s'agit de formes lexicales, on pourra, avec un outil de requête assez rudimentaire, les repérer et les extraire facilement. Si l'outil permet en outre l'utilisation d'expressions régulières[36], il autorise des requêtes plus larges qui, en substituant une saisie en intension à une saisie en extension, évitent d'omettre certaines graphies : rappelons en effet que la variation lexicale est forte en ancien et encore en moyen français. Ainsi, lorsque j'ai travaillé sur les expressions construites sur *propos* (*à propos, à ce propos* et *à propos de*), j'ai effectué une requête sur la chaîne <prop*.>, c'est-à-dire *prop* suivi de tout autre caractère. J'ai ainsi vu apparaître une longue liste de formes, dont j'ignorais pour certaines l'existence : *propos, propoz,*

[34] Mémoire de synthèse, 3.2. « Une tradition de travail sur les textes renouvelée par la numérisation des textes et la linguistique de corpus ».

[35] Il reste qu'une construction rare a plus de chance d'apparaître fréquemment dans un grand corpus que dans un plus petit, pour autant que les deux corpus soient comparablement diversifiés.

[36] Une expression régulière est une chaîne de caractères qui comprend des caractères spéciaux, qui permettent de décrire un ensemble de chaînes de caractères possibles. Par exemple, le point '.' représente n'importe quel caractère, le point d'interrogation '?' signifie que le caractère qui le précède est optionnel, l'étoile '*' indique que le caractère qui la précède peut être répété 0 ou plusieurs fois, etc.

propost, propous, proppoz, proupoz, propotz, propoux, propox. Le lexème /propos/[37] est relativement simple de ce point de vue : sa variation est concentrée dans sa partie finale, ce qui a permis que l'expression régulière ne génère pas trop de « bruit », ni de « silence »[38]. Lorsque la variation graphique d'un lexème est plus complexe, le recours aux expressions régulières est plus coûteux, dans la mesure où elles sont plus difficiles à élaborer. Il est à cet égard confortable de travailler sur des textes lemmatisés, mais peu le sont aujourd'hui. A défaut, pouvoir générer un lexique des formes permet de recenser assez rapidement les différentes graphies d'un lexème, et de ne pas en omettre.

Travailler sur des catégories (morpho-syntaxiques, syntaxiques, sémantiques…) n'est possible que si l'on dispose d'un corpus enrichi des informations pertinentes : le degré d'enrichissement du corpus, c'est-à-dire la présence d'informations morphologiques, syntaxiques… attachées aux constructions simples ou complexes, constitue le second critère d'évaluation de la facilité de traitement des données. Il est évidemment corrélé au critère précédent : plus la construction recherchée est abstraite, plus on a besoin d'un corpus enrichi d'informations pertinentes.

Enfin, troisième critère, liés aux deux précédents, les outils dont on dispose pour effectuer les requêtes sont décisifs : il est bien beau d'avoir un corpus annoté syntaxiquement, mais c'est inutile si l'on n'a pas les moyens d'effectuer des requêtes sur cette annotation, et d'extraire les catégories ou les relations que l'on recherche. Les types et les modes de représentation de l'information dans les textes conditionnent largement les outils à utiliser. Ainsi, actuellement, dans les deux projets d'enrichissement des textes auxquels je participe[39], nous utilisons des outils de requête

[37] La forme entre barres obliques indique le lemme.

[38] Ces deux termes, ont été initialement utilisés dans le domaine de la recherche d'informations. Le « bruit » réfère à la proportion, sur l'ensemble des résultats collectés, de ceux qui ne sont pas pertinents au regard de ce que l'on souhaitait obtenir (il faut alors, si possible, affiner la requête). Le « silence » réfère quant à lui à la proportion d'occurrences pertinentes non collectées (il faut alors essayer d'élargir la requête). La proportion de bruit peut s'évaluer aisément, par l'observation des données collectées, tandis que celle de silence ne peut être mesurée que par un examen du texte dont proviennent les données. De cette différence essentielle d'appréhension – le bruit s'impose, alors que nous pouvons ne pas être conscients du silence – il découle que la part de bruit est souvent indiquée, contrairement à celle de silence.

[39] Projet d'étiquetage morpho-syntaxique CATTEX de la BFM et projet ANR franco-allemand d'annotation syntaxique de la BFM et du Nouveau Corpus d'Amsterdam (NCA) « Syntactic Reference Corpus of Medieval French ».

et d'extraction différents : *TXM*[40] pour les catégories morpho-syntaxiques, et *TIGERSearch*[41] pour les annotations syntaxiques.

Avoir un corpus de textes numérisés n'est donc pas, en soi, garant d'une collecte fructueuse : le corpus doit bénéficier d'un niveau d'enrichissement adapté aux constructions que l'on cherche, plus ou moins abstraites, et il faut par ailleurs disposer des outils adéquats pour repérer et collecter les données.

Les textes du corpus de la présente étude présentent trois niveaux d'enrichissement : certains bénéficient d'un étiquetage morpho-syntaxique fiable (les étiquettes ont été systématiquement vérifiées), d'autres sont dotés d'un étiquetage non totalement fiable (les textes ont été étiquetés automatiquement selon une procédure par apprentissage, mais les étiquettes n'ont pas été vérifiées), enfin, l'un des textes est annoté syntaxiquement. Les constructions étudiées sont par ailleurs de deux types : il s'agit d'une part des constructions 'sujet pronominal-verbe' ou 'verbe-sujet pronominal', et d'autre part des constructions sans sujet exprimé. La conjugaison de ces deux paramètres – degré d'enrichissement des textes et type de construction recherchée – donne lieu à des degrés de difficulté différents pour la collecte des données. La palette des stratégies de repérage conséquemment mise en place sera détaillée en même temps que sera présentée la méthodologie.

2.1.2. Traitement des données

Pour que le support numérique facilite le traitement des données qui auront été extraites, mais aussi de leurs contextes, il faut des outils pareillement adaptés à la complexité des calculs que l'on souhaite opérer : simples comptages ou bien calculs statistiques plus complexes.

L'euphorie que suscite l'abondance de données fait en effet parfois oublier la difficulté que l'on peut rencontrer à les traiter. Il est merveilleux de collecter des milliers d'occurrences pour une construction, mais encore faut-il être en mesure de les

[40] *TXM* (qui remplace *Weblex*) est un logiciel open-source développé par Serge Heiden (ENS de Lyon, laboratoire ICAR), qui met en œuvre la méthodologie textométrique, et permet, entre autres, d'opérer des requêtes (à l'aide d'un moteur de recherche qui produit en sortie des concordances) sur les valeurs de propriétés des formes (pour autant qu'elles aient été préalablement encodées). Voir <http://textometrie.ens-lyon.fr/>.

[41] *TIGERSearch* est un logiciel libre qui permet d'interroger des textes annotés linguistiquement. Plus spécifiquement il permet d'interroger des bases de données structurées en arbres (graphes orientés, connexes et sans cycles).
<http://www.ims.uni-stuttgart.de/projekte/TIGER/TIGERSearch/oldindex.shtml>.

analyser ! La capacité que l'on a à le faire est liée tant au type d'étude que l'on souhaite mener qu'aux outils dont on dispose.

Ainsi, dans le cadre d'une étude réalisée sur *les aucuns /d'aucun(s)/ aucun(s)*[42], il avait été extrait 10 444 occurrences pour la seule période médiévale (16ème siècle inclus), et plus de 30 000 pour les siècles suivants. Il n'était pas possible de traiter l'ensemble des données avec la même finesse d'analyse, dans la mesure où nous ne disposions pas d'outils pour cela. Pour la dimension sémantico-référentielle, la plus complexe de toutes, nous avons effectué une sélection des données traitées en restreignant le sous-corpus des résultats, en particulier par le choix des caractéristiques morphologiques des formes (formes plurielles et pronominales par exemple), et en opérant une sélection en termes de genre parmi les textes dans lesquels apparaissaient les résultats. Sur ce même corpus de départ, j'avais mené auparavant une étude sur *aussi*[43] en position initiale, pour laquelle j'avais collecté moins de 700 formes : ceci illustre bien la différence qu'il peut y avoir, à partir d'un même corpus, entre les différents ensembles que constituent les données pertinentes (voir (Prévost 2005)) pour une plus longue discussion sur ce point). La présente étude constitue un cas emblématique de cette situation : les constructions à sujet postverbal sont relativement rares, celles à sujet préverbal sont beaucoup plus fréquentes, tandis que les structures à sujet non exprimé sont très nombreuses. Pour des raisons matérielles de temps, leur traitement ne peut être le même : alors qu'il est envisageable de traiter l'ensemble des données avec sujet exprimé, il s'avère nécessaire d'opérer une sélection parmi les occurrences d'énoncés sans sujet exprimé. Je préciserai la démarche qui peut être adoptée dans la section 3.2.

2.2. La difficile constitution du corpus

En amont de l'exploitation des données se pose la question de la constitution du corpus, qui revêt des dimensions pratiques et théoriques. Sur le plan pratique, la difficulté tient à l'accessibilité des textes : tous les textes ne sont pas numérisés, et ceux qui le sont sont, pour beaucoup, soumis à des droits non pas d'auteurs mais d'éditeurs, qui restreignent fortement leur utilisation, en particulier pour ce qui est de

[42] Prévost et Schnedecker (2004).
[43] Prévost (1999).

la taille (en nombre de caractères) des contextes accessibles à l'aide des moteurs de recherche.

En admettant cependant que soient résolus les problèmes d'accessibilité aux textes, surgit la question du choix des textes. Pour des raisons matérielles de temps, il n'est en effet souvent pas possible de prendre en considération, pour une étude, tous les textes dont on pourrait disposer[44] : il faut donc opérer une sélection qui conjugue maniabilité du corpus et représentativité de celui-ci au regard de l'étude visée. La représentativité du corpus, qui se décline sur les plans quantitatif et qualitatif, est particulièrement importante pour une langue dont nous n'avons pas la compétence, et pour laquelle nous ne pouvons donc nous servir de l'introspection comme garde-fou fiable aux analyses tirées des textes.

La représentativité du corpus est une question délicate, pour laquelle il n'y a pas de solution « clé en main » : il n'existe pas de normes, ni de recommandations absolues, les choix sont variables en fonction du phénomène étudié et de sa supposée fréquence.

Une mauvaise constitution du corpus peut conduire à une interprétation totalement erronée des résultats. Ainsi dans le cadre d'une première étude du marqueur de topicalisation *quant à* (réalisée en 1999, mais qui a paru tardivement en 2003), j'avais interrogé la *Base de Français Médiéval* (*BFM*)[45] pour l'ancien français, et la base du *Dictionnaire de Moyen Français* (*DMF*)[46] pour le moyen français. Ne trouvant aucune occurrence de l'expression dans la *BFM*, j'en avais conclu à son émergence plus tardive. Mais l'examen des textes de moyen français dans lesquels apparaissait l'expression m'a montré qu'il s'agissait très majoritairement de textes argumentatifs, en tout cas non littéraires ; or la *BFM* n'était à cette époque composée que de textes littéraires : il n'était donc guère 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. Par la suite, la diversification de la *BFM* m'a permis, pour des études ultérieures, de collecter quelques occurrences de *quant à* en ancien français :

[44] Le temps requis pour l'exploitation des données est conditionné par différents facteurs : taille du corpus mais aussi et surtout facilité de repérage et fréquence de la construction (voir supra 2.1.).
[45] <http://bfm.ens-lyon.fr/>
[46] <http://www.atilf.fr/dmf/textes2010.htm>

(22) Et **quant a Dieu** entre tricherie et larrecin a poi de disference (P. de Beaumanoir, *Coustumes de Beauvaisis*, vers 1283)

L'exemple ci-dessus illustre une distinction qu'il convient de faire d'une manière générale parmi les constructions (quel que soit leur degré de complexité) que l'on étudie, entre celles qui apparaissent dans tous les textes (selon des fréquences possiblement variables), et celles qui n'apparaissent pas dans tous les textes, comme c'est le cas pour *quant à* en ancien français, et encore en moyen français. Dans le second cas, la couverture d'un vaste corpus, sinon diversifié en tout cas bien ciblé (pour ce qui est des genres par exemple), s'impose tout particulièrement. Il est d'ailleurs intéressant, mais on le fait rarement, de comparer le nombre total de textes composant un corpus et le nombre de textes dans lesquels apparaît la construction recherchée. Ainsi, pour une étude des expressions formées sur « propos » (*à propos de*, *à ce propos*, et *à propos*)[47], j'ai utilisé comme corpus de travail la base du *DMF*, qui comprenait 218 textes (et près de 7 millions d'occurrences). Après élagage de toutes les occurrences indésirables, il est resté 295 occurrences de *propos*, qui n'apparaissaient que dans 29 des 218 textes, les occurrences étant en outre 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. Certes l'expression est encore bien rare, mais elle existe.

La situation est différente pour les sujets pronominaux : ils sont présents dans tous les textes, actualisés selon les trois modalités qui les caractérisent, à savoir l'expression, préverbale et postverbale, et la non-expression. Mais ils se réalisent dans des proportions différentes d'une époque à l'autre, d'un genre à l'autre et même d'un texte à l'autre, et ils le font selon des modalités pareillement variables : si le mauvais ciblage du corpus d'étude ne risque pas d'occulter l'émergence des constructions recherchées, il peut en revanche faire passer à côté d'étapes importantes dans leur

[47] Prévost (2007).

évolution, tant en ce qui concerne leurs fréquences que les contextes dans lesquels elles peuvent se trouver.

Il existe aujourd'hui plusieurs bases de textes numérisés de français médiéval, plus ou moins importantes, plus ou moins enrichies, plus ou moins outillées[48]. Ce sont des bases textuelles précieuses, au sein desquelles chacun peut constituer son « corpus » pour une étude spécifique. La contrepartie négative de l'existence de ces bases est la tentation – et la tendance observée – de ne plus travailler que sur les textes numérisés, pour des raisons pratiques. Or, dans la mesure où ces bases sont encore mal équilibrées (en particulier du point de vue du genre), on laisse nécessairement de côté l'étude de certains genres de textes. Il faut donc toujours garder à l'esprit que ces bases ne sont pas représentatives de la variété des textes dont nous disposons, et en tenir compte, sans hésiter à reprendre des éditions sur papier et le crayon. Il faut signaler à cet égard qu'un effort important a été fourni depuis quelques années pour diversifier les textes qui composent la *BFM*, d'une part au niveau des genres, mais aussi en y intégrant des textes antérieurs au 12ème siècle[49], période généralement mal représentée dans les bases textuelles, et donc dans les études.

2.3. Les apports des données numérisées outillées : des quantifications multiples

Disposer de données massives permet en premier lieu de découvrir de nouvelles constructions, de nouveaux phénomènes, mais aussi de réviser des interprétations et des analyses précédentes : des régularités observées à grande échelle peuvent mettre en cause une règle existante, ou permettre d'en énoncer une qui est inédite.

Cela suppose évidemment de quantifier les données et de disposer d'outils pour le faire. Le dénombrement précis et massif des données (et non une simple appréciation vague) constitue l'apport majeur des corpus numérisés outillés. En effet, même si l'on a la compétence de la langue que l'on étudie, l'introspection est insuffisante pour évaluer la fréquence plus ou moins grande d'une construction, et elle est incapable de

[48] *Nouveau Corpus d'Amsterdam* (NCA) à Stuttgart, *Base des Textes de Français Ancien* (*TFA*) à Ottawa <http://www.uottawa.ca/academic/arts/lfa>, et *Base de Français Médiéval* (*BFM*), à l'ENS-LSH de Lyon < http://bfm.ens-lyon.fr/>, et *Base du Dictionnaire de Moyen Français*, à l'ATILF, Nancy <http://www.atilf.fr/dmf/textes2010.htm>.

[49] Dans le cadre du projet ANR Corptef (Corpus représentatif des premiers textes français) sous la direction de Céline Guillot (ENS Lyon).

la quantifier précisément. Comme le rappelle Corbin (1980 : 121) « 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. », et plus loin il insiste sur « l'espoir chimérique d'atteindre par introspection un 'réel' linguistique ayant un ancrage social ». Nécessaire pour une langue dont on a la compétence, la quantification exacte des données l'est encore plus pour une langue pour laquelle on ne peut de toute façon pas recourir à l'introspection.

Déterminer la fréquence exacte d'une construction et des contextes dans lesquelles elle apparaît présente au moins deux intérêts majeurs. D'une part, cela permet de mesurer l'évolution de sa productivité, ce qui est important dans tous les phénomènes de changement, en particulier pour ceux qui relèvent de la grammaticalisation. On observe en effet souvent le schéma d'évolution suivant : 'A > Ab > aB > B', qui suppose, pour sa mise au jour, une quantification précise de A et de B[50].

Etablir une quantification précise permet d'autre 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 capital (trop souvent négligé selon moi), en particulier lorsque l'on travaille sur l'émergence d'une construction (voir ci-dessus l'exemple de *à propos de* en position initiale, et sa concentration chez un auteur).

Les outils statistiques, même assez simples, permettent en outre d'apprécier le caractère plus ou moins « surprenant » de la distribution d'une construction, d'évaluer la corrélation de phénomènes entre eux, et de croiser différents facteurs. Nous reviendrons sur cette question, en 3.3., en présentant les outils de calcul utilisés dans cette étude.

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

Il convient enfin de mentionner un autre type d'apport des outils informatiques: il s'agit de la régularité, de l'homogénéité de l'analyse effectuée, quelle que soit sa

[50] A et B peuvent correspondre à des constructions simples (des formes) ou complexes, mais aussi aux différentes valeurs d'une construction. Durant la première étape, A apparaît seule, puis, seconde étape, B apparaît mais est minoritaire, avant de devenir majoritaire devant A dans une troisième étape. Ultime étape possible : B supplante A.

nature (collecte de données, dénombrements, etc.). En effet, lorsque l'on cherche un objet linguistique à l'œil nu, on peut omettre des occurrences[51]. 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.

C'est un fait qui s'observe particulièrement dans le cadre des procédures d'enrichissement des textes, qu'il s'agisse d'étiquetage morpho-syntaxique ou d'annotation syntaxique : avec de mêmes critères et de mêmes recommandations, des annotateurs différents sont susceptibles de produire des résultats différents, et un même annotateur peut même ne pas être cohérent avec lui-même, en particulier lorsqu'une même forme ou une même construction peut revêtir des valeurs différentes selon les contextes, et que la délimitation entre elles repose sur des critères qui laissent une place à l'interprétation de l'annotateur. C'est le cas par exemple, dans le cadre du projet d'annotation syntaxique *SRCMF*[52], de l'interprétation de *ce* comme porteur ou non d'une valeur anaphorique, et donc sujet référentiel ou non. Dans le premier cas, il est annoté sujet personnel, dans le second il est annoté sujet impersonnel. L'apposition de l'étiquette « auxilié » peut susciter pareillement des divergences d'interprétation. Elle recouvre à la fois les infinitifs qui suivent un modal (*je* ***veux*** *partir*) et les participe-passés, qu'ils se trouvent dans des temps composés ou dans des constructions passives. Dans le dernier cas, le label « auxilié » est spécifié en « auxilié passif » lorsque l'agent est explicitement présent, ou facilement récupérable. Cette seconde possibilité ouvre des interprétations parfois variables d'un annotateur à l'autre, comme c'est le cas dans l'exemple (23). L'agent de « *arse* (‘brûlée') n'est pas explicité, mais il est facilement récupérable ; on peut aussi considérer que *ré* (‘bûcher') a non seulement une fonction de complément locatif, mais aussi d'agent (ce sont les flammes qui brûlent), dans ce cas explicite :

[51] Mais il est vrai que, à l'inverse, l'œil humain saura reconnaître une forme sous ses différentes variantes, alors que l'outil ne tiendra compte que des variantes qui ont été explicitement spécifiées.

[52] Projet ANR franco-allemand « Syntactic Reference Corpus of Medieval French » (SRCMF) que je codirige avec Achim Stein (université de Stuttgart) ; le projet est prévu pour 3 ans (décembre 2008-décembre 2011), et il a reçu, côté français, une dotation de 160 000 euros, qui prévoit l'emploi d'un post-doctorant ou d'un ingénieur de recherche pendant 3 ans. Le projet réunit douze personnes ; pour la partie française : Julie Glikman (post-doctorante), Céline Guillot, Serge Heiden, Alexei Lavrentiev, Christiane Marchello-Nizia, Sophie Prévost, Tom Rainsford et Bernard Victorri à titre d'expert ; pour la partie allemande : Béatrice Bischoff (post-doctorante), Nicolas Mazziotta (post-doctorant) et Achim Stein, ainsi que Fernande Dupuis (UQAM, Canada).Voir présentation du projet < https://listes.cru.fr/wiki/srcmf>.

(23) Donc voudroit miex morir que vivre,
Donc savra bien Yseut la givre
Que malement avra ovré :
Mex voudroit estre **arse** en un ré. (*Tristan* de Beroul)

Il en va de même dans le cadre des procédures de vérification humaine de textes enrichis, que l'enrichissement ait été réalisé manuellement ou automatiquement (grâce à une procédure d'apprentissage par exemple) : il y a des incohérences d'une personne à l'autre, mais aussi chez une même personne. Seul l'outil informatique, lorsqu'il existe, peut assurer la tâche avec une cohérence sans faille.

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

Il apparaît donc que, si le recours aux textes numérisés et à leur traitement informatique peut occulter des constructions inédites ou des phénomènes intéressants (voir dans le document de synthèse la discussion en 3.2.1. et 3.2.3. à propos du fait qu'on lit de moins en moins les textes de manière linéaire), il permet en revanche de percevoir des objets invisibles à l'œil nu, et d'accéder à de nouvelles couches de description. Je terminerai par un exemple assez éloquent : une requête élémentaire sur la fréquence des étiquettes morpho-syntaxiques des textes enrichis de la *BFM* m'avait permis de constater que c'est le verbe conjugué qui est le plus fréquent dans les textes d'ancien français, alors que c'est le nom commun qui le devient à partir du moyen français[53]. L'étiquetage morpho-syntaxique en cours de l'ensemble de la *BFM*, ainsi que l'annotation syntaxique de plusieurs textes dans le cadre du projet *SRCMF*, laisse entrevoir la possibilité d'explorer des champs encore largement en frîche.

Constitution du corpus, collecte des données et sélection parmi elles, choix des outils de calcul et d'analyse : ces différentes étapes, communes à bon nombre d'analyses sur corpus, posent un certain nombre de questions auxquelles j'ai proposé les réponses qui me semblaient les plus adéquates dans le cadre de l'étude présentée ici. Ce sont ces choix que je présente dans la partie méthodologique ci-après.

[53] Conclusion confirmée en comparant la première règle d'étiquetage générée par un étiqueteur procédant par apprentissage appliqué à un corpus d'ancien français et à un corpus de français moderne : dans le premier cas, cette règle stipulait que « tout mot contenant un 'e' est à étiqueter verbe conjugué », dans le second cas elle stipulait que « tout mot contenant un 'e' est à étiqueter nom commun ».

Chapitre 3. Choix méthodologiques

La démarche méthodologique qui est présentée ici résulte d'un certain nombre de contraintes, qui ont conduit, d'une part à revoir certaines ambitions initiales à la baisse, et d'autre part à développer certaines stratégies. Si les données analysées ne sont pas aussi nombreuses que ce que l'on aurait pu souhaiter, la démarche proposée ici n'en constitue pas moins une méthode qui pourra être étendue à des données complémentaires. Par ailleurs, en mettant au jour ce qui est possible et ce qui ne l'est pas en fonction du degré actuel d'enrichissement des textes et des fonctionnalités des outils utilisés, nous laissons entrevoir ce que les avancées, tant en ce qui concerne les informations attachées aux formes linguistiques que le développement des outils, pourront permettre dans un avenir que l'on peut espérer assez proche.

3.1. La constitution du corpus

Les textes rassemblés pour cette étude constituent un sous-corpus de celui qui a été établi pour le projet de *Grande Grammaire Historique du Français (GGHF)*[54]. En effet, alors que le corpus de la *GGHF* couvre toute l'histoire du français, du 9ème au 20ème siècle, notre étude ne concerne que la période qui s'étend du 9ème au 14ème siècle, pour laquelle nous avons opéré une sélection parmi les textes retenus pour la *GGHF*.

La question de la constitution du corpus s'est posée de manière particulièrement aiguë pour la *GGHF*, dans la mesure où les textes ne sont pas destinés à jouer un simple rôle de réservoir à exemples, mais sont la base sur laquelle se fonde, pour une large part, la description et l'analyse de l'évolution des différents phénomènes langagiers étudiés.

3.1.1. Les choix opérés pour la *Grande Grammaire Historique du Français*

Nous ne reviendrons pas ici sur les difficultés liées à la constitution d'un corpus, en particulier pour l'accès à certains usages : elles ont été évoquées en 2.2. Elles se sont posées pour le corpus de la *GGHF*. En adoptant une démarche sur corpus, il nous a fallu expliciter une difficulté tout simplement éludée lorsque le corpus n'est pas à

[54] Projet actuellement en cours, sous la direction de Christiane Marchello-Nizia, Bernard Combettes, Sophie Prévost et Tobias Scheer.

l'ordre du jour : la représentativité du corpus et du même coup l'aptitude de la « grammaire » à être généralisable au-delà des seuls textes qui constituent le corpus.
La représentativité du corpus a été envisagée des points de vue quantitatif et qualitatif. Sur le plan quantitatif, il a fallu décider si l'on travaillait sur des textes intégraux ou sur des échantillons, ou bien en combinant les deux, selon la taille des textes. On peut ainsi décider d'échantillonner les textes lorsqu'ils excèdent un certain nombre de mots.
Sur le plan qualitatif, il est apparu que, dans un tel projet, les paramètres suivants étaient décisifs : le domaine et/ou le genre textuel, la date, le dialecte, la forme du texte (vers/prose). L'objectif a été d'obtenir un corpus aussi représentatif que possible de l'objet « langue française », dans toute la diversité qu'on lui présuppose. Plus un corpus est jugé représentatif, plus il est légitime de généraliser au-delà des seuls textes qui le constituent. Mais il fallait aussi constituer un corpus qui reste manipulable, non seulement du point de vue de l'exploration des textes mais aussi du traitement des données extraites. Il a donc fallu trouver un compromis acceptable entre le corpus idéal (inaccessible, mais dont il faut se rapprocher autant que possible), le corpus souhaité, et le corpus possible et raisonnable. La constitution d'un corpus à géométrie variable a permis de résoudre en partie les difficultés liées à la variation des modalités d'exploration des corpus et de traitement des résultats.

3.1.1.1. Un corpus à géométrie variable

La *Grande Grammaire Historique du Français* s'est donné un corpus à géométrie variable, tant du point de vue de sa constitution que de son utilisation, mais c'est uniquement de la première que nous parlerons ici. Il a en effet été élaboré, pour chaque siècle, un double corpus : un corpus « noyau » et un corpus « complémentaire ». Le premier, dont sont tirés les textes de la présente étude, répond à des critères de composition assez stricts quant à la taille des textes et quant à leur diversité.
Pour ce qui est de la taille, nous avons fait le choix de retenir les textes dans leur intégralité lorsqu'ils n'excèdent pas 45 000 mots (ponctuation comprise, soit un peu plus de 40 000 mots sans ponctuation). Pour les textes dépassant ce seuil, nous avons sélectionné trois échantillons d'environ 15 000 mots (ponctuation comprise) en début, milieu, et fin de texte. Il s'est trouvé deux exceptions à cette procédure générale. Tout

d'abord, pour certains textes, jugés répétitifs du point de vue de leurs structures morphosyntaxiques, la taille de l'échantillon a été réduite à 22 500 (ponctuation comprise). C'est le cas, par exemple, des *Coutumes de Beauvaisis.* Pour d'autres textes, c'est la répartition des échantillons qui s'est faite un peu différemment. C'est le cas des *Miracles* de Coinci. Les miracles constituent en effet de petites entités indépendantes, il n'y a pas de trame textuelle : il n'a donc pas été opéré un échantillonnage en trois endroits du texte, mais on a pris un extrait du 2ème volume.
Pour chaque période le corpus noyau comprend entre 200 000 et 230 000 mots, hormis pour la période la plus ancienne, avant 1100, pour laquelle nous avons retenu la quasi-totalité des textes dont nous disposons, l'ensemble ne dépassant pas 10 000 mots. C'est dans le corpus noyau, dont certains textes bénéficient d'un étiquetage morpho-syntaxique et d'une annotation syntaxique, que sont prioritairement effectués les calculs de fréquence.
La constitution du corpus complémentaire n'a pas été soumise aux même contraintes : la taille des textes n'a pas été limitée, et les autres critères (genre, dialecte…) ont été appliqués avec une rigueur moindre.

Le corpus joue un rôle décisif dans la *Grande Grammaire Historique du Français*, même si l'approche adoptée n'est pas strictement « corpus-driven ». Une telle approche suppose en effet « qu'aucune position théorique *a priori* ne préside aux observations sur corpus, la théorie étant induite du corpus » (Léon 1998 : 12), or nous nous appuyons en partie sur ce que nous savons déjà. Mais le corpus ne constitue pas pour autant un simple réservoir à exemples : il permet de confirmer ou d'infirmer les hypothèses avancées.
Nous ne partons en effet pas de rien : de nombreux phénomènes linguistiques ont déjà été bien décrits, et il ne s'agit pas de tout réécrire. Nous exploitons donc plusieurs études, ainsi que les données quantifiées qui les accompagnent, le cas échéant. Elles sont parfois complétées par de nouveaux relevés, opérés dans notre corpus. Les études inédites, ou partiellement inédites, se sont beaucoup plus largement appuyées sur l'exploitation du corpus.

3.1.1.2. Les critères de sélection des textes

Nous avons retenu, pour la sélection des textes, différents critères. Certains, que nous appellerons les descripteurs, ont pour but de caractériser le contenu des textes.

D'autres relèvent davantage du point de vue que le locuteur moderne porte sur ces textes. Il nous a ainsi paru important que le corpus de la *Grande Grammaire Historique du Français* comprenne, pour chaque période, quelques textes de référence, à côté de textes moins connus (et souvent aussi – car les deux sont de fait liés – moins littéraires). Certes la notion de texte de référence est en partie subjective, mais l'on peut cependant identifier comme tels quelques textes, en particulier pour la période médiévale. Il n'était ainsi pas concevable que le corpus du 12ème siècle, par exemple, ne contienne pas la *Chanson de Roland* et un roman de Chrétien de Troyes ; pour le 13ème siècle, la *Queste du Graal* et le *Roman de la Rose* se sont d'emblée imposés. Cela ne signifie pas pour autant que nous considérons qu'il s'agit là des seuls textes de référence pour les périodes concernées, mais il fallait faire des choix.

La sélection s'est révélée plus difficile au fur et à mesure que l'on avance dans le temps et que se multiplie la production écrite… et donc les textes « incontournables ». Le choix a nécessairement été partial, mais néanmoins influencé par la prise en compte d'un autre paramètre : la qualité des éditions. Pour les textes les plus anciens, cette exigence nous a conduits à privilégier les éditions les moins interventionnistes (et pour la période suivante, du 16ème au 19ème siècle, des textes non modernisés, l'examen des graphies étant un bon indice).

Il est enfin un critère très « pratique » qui est intervenu dans nos choix, conjointement à ceux précédemment mentionnés et aux descripteurs qui vont être évoqués ci-après. Il s'agit de l'existence d'une version numérisée disponible des textes, au moins pour ceux qui appartiennent au corpus noyau et qui ont fait l'objet de quantifications. Pour la période médiévale, nous nous sommes très largement appuyés sur les textes de la *Base de Français Médiéval*, dont certains sont enrichis linguistiquement (pour la période suivante, nous avons majoritairement sélectionné les textes dans la base *Frantext*).

Les textes retenus l'ont été aussi, et prioritairement, parce qu'ils contribuaient à construire le corpus diversifié et représentatif que nous souhaitions. Cette représentativité a été établie en fonction des critères qui nous ont semblé les plus pertinents, à savoir la date, le domaine et le genre, la forme et le dialecte des textes. La prise en compte de ces critères a été grandement facilitée par l'important travail sur les descripteurs des textes qui a été mené par C. Guillot, A. Lavrentiev et C. Marchello-Nizia dans l'équipe de la *BFM* à l'ENS de Lyon, et plus spécifiquement dans le cadre

du projet ANR *Corptef* (Corpus représentatif des premiers textes français) dirigé par Céline Guillot. Ce travail a non seulement consisté à expliciter de manière systématique certaines informations liées aux textes, plus ou moins faciles à établir (date de composition, date du manuscrit, dialecte…), mais aussi à élaborer une classification en termes de domaines et de genre[55].

C'est en nous appuyant sur les valeurs de ces différents descripteurs que nous avons constitué le corpus de la période médiévale.

La date des textes :

La *Grande Grammaire Historique du Français* est organisée, en premier lieu, par grands domaines de la langue (phonétique, morphologie, sémantique,…) : c'est au sein de chacune des questions abordées qu'intervient la perspective chronologique.

Chaque phénomène a sa propre temporalité. Nous avons donc établi un cadre chronologique très général, en délimitant des périodes de manière arbitraire, suivant pour cela un simple découpage par siècles, cette division ne correspondant à aucune présupposition quant à la périodisation des évolutions individuelles. Elle est un simple cadre à la description des phénomènes et des changements, dont les périodisations respectives n'émergeront qu'à l'issue des études.

C'est donc par siècles que nous avons organisé notre corpus, en essayant de sélectionner des textes qui s'échelonnent du début à la fin de chaque siècle. La période qui précède le 12ème siècle fait exception : en raison du petit nombre de documents qui nous sont parvenus, et de leur brièveté, nous avons regroupé ensemble les quelques textes dont nous disposons. Pour eux, la mise en œuvre des autres critères n'est donc pas pertinente : la *Séquence de Sainte Eulalie* a été retenue non pas parce que c'est un texte en vers qui relève du domaine religieux, mais simplement parce que c'est, avec les *Serments de Strasbourg*, le seul texte en « français » du 9ème siècle.

Pour chaque siècle envisagé, nous avons choisi des textes en fonction de trois critères, en faisant en sorte que l'ensemble composé soit diversifié.

[55] Voir la présentation détaillée des descripteurs sur le site du projet Corptef : <http://w7.ens-lsh.fr/corptef/spip.php?rubrique60>

La forme des textes : vers /prose

La distinction entre textes en vers ou en prose recouvre des réalités différentes selon les périodes considérées. Ainsi, jusqu'au 12ème siècle, la grande majorité des textes sont écrits en vers, qu'il s'agisse de récits épiques, de « romans », de récits hagiographiques… Au fil des siècles l'écriture versifiée va reculer, conjointement au développement de la prose à partir du 13ème siècle, pour finalement se voir réservée aux textes de poésie et de théâtre. La place respective faite à la prose et au vers n'est donc pas la même dans notre corpus selon les siècles considérés : les textes en vers sont très largement majoritaires jusqu'au 12ème siècle, puis ils cèdent une place croissante aux textes en prose.

Les dialectes

Le critère dialectal a tenu une place un peu à part parmi nos critères de sélection. Tout d'abord, il n'est véritablement pertinent que jusqu'au 15ème siècle, et déjà bien moins discriminant à cette époque qu'au 12ème siècle. Par ailleurs il n'est pas toujours facile de définir le dialecte d'un texte, et il est fréquent d'opter pour un dialecte… 'non défini', ce qui signifie que le texte ne comporte pas de traits dialectaux marqués. De plus, un texte peut être plus ou moins marqué du point de vue dialectal : ainsi *La Conqueste de Constantinople* de R. de Robert de Clari, au début du 13ème siècle, comprend de nombreux traits picards qui affectent des morphèmes courants (déterminants, en particulier les démonstratifs, verbes, etc.), alors que la « coloration » picarde est moins nette dans les *Coustumes de Beauvaisis* de Philippe de Beaumanoir (fin du 13ème siècle).

Enfin, et cela résulte en grande partie de la remarque précédente, ce n'est pas sur la base de leur dialecte que nous avons prioritairement sélectionné les textes. Il se trouve néanmoins que les textes retenus présentent, pour les périodes où cette distinction est pertinente, une relative diversité. Sont en particulier bien représentés l'anglo-normand et le picard, dont on sait qu'ils présentent plusieurs traits linguistiques spécifiques. Il est vrai que le 12ème comprend une majorité de textes anglo-normands ou normands, tandis que le 13ème siècle tend à favoriser les textes picards, mais cette répartition, plus qu'un défaut de notre corpus, dénote les tendances de la production de ces deux siècles.

Les genres

Nous n'aborderons les genres qu'en relation avec la période médiévale. Comme nous l'avons indiqué plus haut, nous nous sommes appuyés, pour le choix des textes selon le critère « genre », sur la classification qui a été élaborée dans le cadre de la *BFM* et du projet *Corptef*. Celle-ci n'est que partiellement pertinente pour les siècles suivants, pour lesquels nous avons donc adopté une autre classification[56].

Pour la période médiévale, il a été établi une distinction entre « domaines » et « genres ». Pour les premiers, les valeurs sont les suivantes : Religieux, Litteraire, Juridique, Historique, Didactique. La liste des genres est en revanche ouverte : Fabliau, Chronique, Nouvelle, Roman,… Même si le descripteur « genre » est d'une certaine façon subordonné au descripteur « domaine », un même genre peut néanmoins se rencontrer dans différents domaines (par exemple le genre «dramatique » se rencontre dans les domaines « religieux » et « littéraire »).

Nous considérons le corpus de la *GGHF*, non pas comme un corpus parfait, mais comme un bon compromis, du point de vue de sa représentativité, entre le corpus idéal et le corpus matériellement possible et exploitable dans le cadre du projet *GGHF*.

3.1.2. Un sous-corpus de la *GGHF*

Il n'était pas possible, dans le cadre de la présente étude, d'exploiter l'ensemble de la partie médiévale du corpus *GGHF* (9ème-15ème siècle), constitué de 956 858 mots. La taille est trop importante, le traitement aurait été trop long, en particulier parce que nous ne disposons pas encore des requêtes permettant de ne sélectionner que les occurrences désirables (voir 3.4. pour la formulation des requêtes).

Il a donc été réduit de deux manières. D'une part je n'ai considéré que la période des 12ème-14ème siècles, en laissant de côté le 15ème siècle[57]. J'ai d'autre part effectué, pour chaque période, une sélection parmi les textes du corpus de la *GGHF*[58].

[56] Il est vrai aussi que toutes les classes de cette typologie ne sont pas pareillement pertinentes pour toutes les étapes de la période médiévale. Ainsi la « chronique » (récit historique) n'a guère de réalité au 12ème siècle, ce genre de récit se faisant alors encore en latin.

[57] L'évolution de la position du sujet a en partie été étudiée pour cette période dans (Prévost 2001).

[58] Rappelons que ce découpage par siècles n'est qu'une commodité, mais qu'il ne présuppose nullement une quelconque chronologie des faits.

Il est sûr que, la réduction du nombre de textes du corpus de la *GGHF* conduit à un corpus potentiellement moins représentatif : il devient par conséquent plus difficile d'évaluer le poids respectif des différents critères : j'aurai l'occasion de revenir sur la question du « parasitage » entre critères au moment de l'analyse et de l'interprétation des résultats.

Un corpus idéal permettrait de faire varier un seul critère à la fois : par exemple entre deux textes de date, forme et dialecte identiques, on ne ferait varier que le genre ; puis entre deux textes de date, forme et genre identiques, on ne ferait varier que le dialecte, etc. Il faudrait même prévoir au moins trois textes pour chaque configuration : deux dont les critères seraient strictement identiques, et un troisième dont l'un des critères varierait par rapport aux deux autres textes. Cela permettrait de limiter le risque que la variation observée tienne plus aux pratiques idiolectales d'un auteur qu'aux éventuelles spécificités du genre ou du dialecte, etc,.

La lourdeur du traitement d'un tel corpus n'était pas compatible avec le format de la présente étude[59].

En réduisant le nombre de textes, j'ai essayé de maintenir malgré tout, autant que possible, un relatif équilibre entre les caractéristiques des textes, en termes de domaines/genres, de forme et de dialecte. On pourra trouver que le 12ème siècle est peu diversifié (prévalence du domaine littéraire, de la forme vers, des textes anglo-normands ou normands) : cela tient largement à la moindre disponibilité de textes variés pour cette époque. De même, comme je l'ai dit plus haut, le 13ème siècle contient un nombre proportionnellement important de textes plus ou picardisants : c'est une réalité historique. Par ailleurs, la fin du 12ème siècle et le début du 13ème siècle sont particulièrement bien représentés, ainsi que, dans une moindre mesure, la fin du 14ème siècle. C'est un choix délibéré, qui a permis, sur ces créneaux chronologiques, de faire varier les autres paramètres plus systématiquement que sur les autres périodes, afin de tester, sur deux « échantillons », les effets d'une variabilité plus poussée.

Je suis bien consciente que la dimension contrastive ici défendue n'est que partiellement actualisée dans l'étude : je la considère comme l'ébauche d'une démarche qui sera à développer dans les travaux qui feront suite à celui-ci.

Voici ci-dessous la liste des textes retenus :

[59] La difficulté à trouver des textes constitue de toute façon un obstacle préalable ; c'est la raison pour laquelle le corpus de la *GGHF* lui-même ne répond pas à une telle exigence.

Titre (+ auteur)	Enrichisse-ment [60]	Date de composition	Dialecte	Forme	Domaine	Genre
12ème siècle						
Chanson de Roland	cattex +	vers 1100	anglo-normand	vers	littéraire	épique
Eneas	cattex -	Vers 1155	normand	vers	littéraire	roman
Tristan, Beroul	annotation syntax.	entre 1165 et 1200	traits normands	vers	littéraire	roman
Ami et Amile	cattex -	Vers 1200	non marqué	vers	littéraire	épique
13ème siècle						
Conquête de Constantinople, Robert de Clari	cattex -	après 1205	picard	prose	historique	chronique
Aucassin et Nicolette	cattex -	1ère moitié du 13è	traits picards	mixte	littéraire	récits brefs
Miracles de Nostre Dame, G. de Coinci	cattex -	vers 1218-1227	non marqué	vers	religieux	lyrique
Queste del saint Graal	cattex +	vers 1225-1230	non marqué	prose	littéraire	roman
Coutumes Beauvaisis, P. de Beaumanoir	cattex -	vers 1283	traits picards	prose	juridique	traité
14ème siècle						
Mémoires ou Vie de saint Louis, Joinville	cattex -	entre 1305 et 1309	non marqué	prose	historique	mémoires
Chroniques, Froissart	cattex -	entre 1369 et 1400	franco-picard	prose	historique	chronique
Estoire de Griseldis en rimes et par personnages	cattex -	1395	traits picards	vers	littéraire	dramatique
Manières de langage	cattex -	1396, 1399	non marqué	mixte	didactique	manuel
Quinze joies de mariage	cattex +	vers 1400	non marqué	prose	littéraire	nouvelle

Tableau 1 : Textes et descripteurs

[60] J'indique ici le type d'enrichissement des textes : 'cattex+', pour étiquetage vérifié, 'cattex-' pour étiquetage non vérifié, et 'annotation syntaxique'. Voir plus loin pour plus de précisions.

Voici par ailleurs la taille des textes en nombre de mots, signes de ponctuation compris. Le second chiffre correspond à la taille du texte étudié, c'est-à-dire, pour certains, après sa réduction selon les modalités présentées en 3.1.1.1.

Le corpus est environ deux fois plus petit que le corpus de la *GGHF* (457 736 mots *versus* 956 858 mots).

Texte	**Nombre de mots du texte**	**Nombre de mots après échantillonnage**
Roland (ca. 1100)	29 338	29 338
Eneas 1 (ca. 1155)	34 958	34 958
Tristan de Beroul (entre 1165 et 1200)	27 257	27 257
Ami et Amile, (1200)	25 283	25 283
Conquête de Constantinople, Clari (après 1205)	33 994	33 994
Queste del Saint Graal (ca. 1220)	104 762	45 000
Miracles de Coinci, (livre second) (1218-1227)	132 682	22 500
Aucassin et Nicolete (1ère moitié du 13ème)	10 009	10 009
Coutusme de Beauvaisis, Beaumanoir (ca. 1283)	142 507	22 500
Vie de saint Louis de Joinville (1305 ou 1309)	75 699	45 000
Chronique de Froissart (entre 1369 et 1405)	216 520	45 000
Estoire de Griseldis (1395)	16 243	16 243
Quinze Joyes de Mariage (ca. 1400)	39 404	39 404
Manières de langage (1396, 1399)[61]	20 282	20 282
Taille du corpus	908 938	**416 768**

Tableau 2 : taille des textes en nombre de mots, signes de ponctuation compris

3.2. Constitution des données

3.2.1. Choix des constructions

Toutes les constructions comprenant un sujet pronominal, exprimé ou non, n'ont pas été retenues. Il a en effet été opéré une double sélection.

En premier lieu, les personnes 2, 4 et 5 (« tu », « nous » et « vous ») ont été écartées, que le sujet soit exprimé ou non[62]. Ces sujets pronominaux apparaissant comme quantitativement marginaux[63] comparés à ceux des personnes 1 et 3, c'est sur ces

[61] Les deux textes de 1396 et 1399 ont été regroupés.

[62] « Il » impersonnel a d'emblée été exclu, du simple fait de son caractère non référentiel. Son expression en français médiéval, et particulièrement en ancien français, reste de toute façon marginale.

[63] Je ne dispose pas de données quantifiées ; il faudrait précisément opérer des relevés pour cela, or, aussi peu fréquentes que soient ces formes, leur collecte n'en demeure pas moins complexe, qu'il s'agisse de relever les sujets non exprimés (cette question sera abordée plus loin) ou bien les sujets exprimés « nous » et « vous », qui ont la même forme que les pronoms compléments. Il me semble

derniers que j'ai préféré me concentrer, et ce d'autant qu'il s'agissait d'évaluer prioritairement les différences d'évolution entre le pronom désignant la personne du locuteur et les pronoms excluant l'implication du locuteur (or, du point de vue référentiel, un « tu » suppose plus un « je » qu'un « il » ne le fait). Outre les pronoms des personnes 1 et 3, j'ai retenu ceux de la personne 6, « ils » étant, référentiellement, un véritable pluriel de « il », alors que « nous » ne consiste pas en l'addition de plusieurs « je » (Benveniste parle de personne « amplifiée » pour « nous » et « vous »)[64].

Les pronoms de 1ère personne sont désormais abrégés P1 et ceux de 3ème et 6ème personnes sont abrégés P3.

Par ailleurs, l'étude a porté sur les propositions déclaratives, indépendantes ou principales : on a écarté toutes les propositions subordonnées, ainsi que les propositions interrogatives et impératives.

3.2.2. Le recensement des formes des sujets pronominaux P1 et P3

Les textes du corpus de la *GGHF* ont été intégrés dans le logiciel *TXM* (développé par Serge Heiden : voir note 9) par Alexei Lavrentiev et Matthieu Decorde. *TXM* permet de créer des sous-corpus : j'ai ainsi constitué 14 sous-corpus correspondant à chacun des textes retenus, ainsi qu'un sous-corpus les réunissant tous. En effet, si la collecte des données a été réalisée texte par texte, non seulement parce que cela semblait plus simple, mais aussi parce que les textes ne bénéficient pas tous du même degré d'enrichissement, il s'avérait en revanche pratique de pouvoir obtenir certaines informations sur l'ensemble du corpus, sans avoir à procéder texte par texte.

Il a ainsi été possible d'établir, globalement, la liste des formes des pronoms, en faisant une requête sur l'index du corpus[65]. En effet, si l'inventaire des formes de sujet pronominal est simple et rapide en français moderne, il est moins évident en français médiéval, en raison de la variété des formes qui existent. Les formes masculines des

---

néanmoins légitime d'affirmer, en me fondant sur ma fréquentation des textes médiévaux, que les pronoms de 2ème, 4ème et 5ème personnes sont globalement moins représentés que les autres.

[64] Accessoirement l'exclusion des sujets pronominaux « nous » et « vous » a permis d'éviter la collecte de constructions indésirables telles que celle-ci : *Et quant eles furent a l'eve, et li chevalier les virent venir si comencierent a dire : « **Tornez** **vos**, veez ci la reïne »* (*Queste del Saint Graal*), tournure impérative dans laquelle *vos* est complément.

[65] Pour cette requête, on aurait pu pareillement utiliser le lexique : le lexique calcule la liste hiérarchique des valeurs d'une propriété de mot donnée tandis que l'index calcule la liste hiérarchique des **combinaisons** de valeurs de propriétés données. Ici le lexique aurait suffi.

personnes 3 et 6[66] sont peu nombreuses, et bien répertoriées par les grammaires : il s'agit de *il*, *ils*, *ilz* principalement. Buridant (2000 : 333-335) mentionne des occurrences sporadiques de *ill* (j'en ai relevé une seule occurrence dans la *Queste*), ainsi que de rares occurrences, en anglo-normand, de *eus* pour « il » au cas sujet pluriel. Je n'en ai trouvé aucune occurrence. Il évoque aussi la possible réduction, à partir du 12ème siècle, de *il* en *i*. Un survol des 1 799 occurrences de *i* n'a pas laissé paraître, en proposition indépendante, d'occurrences de ce type. Les formes que peut revêtir la personne 3 au féminin sont plus variées, et celles de la personne 1 le sont davantage encore, en raison, pour cette dernière, d'une large gamme de formes complexes qui résultent de l'enclise du pronom sujet sur un pronom complément (*jel* = je + le) ou sur la négation (*jen* = je + ne). Il fallait dresser la liste de toutes ces formes, sans en oublier. La possibilité d'utiliser des expressions régulières a permis de ne pas passer en revue tout l'index (33 504 formes !) mais de cibler les requêtes. On sait que les formes de 3ème personne commencent par « el » ou « il », et que celles de première personne commencent par « g » ou « j ». Une requête de ce type génère cependant beaucoup de bruit. Dans la mesure où les formes en « il » étaient identifiées avec certitude, on ne les a pas incluses dans la requête. De même, pour la première personne, on sait que, hormis pour les formes *j*, *j'*, *g*, et *g*, un 'e' ou un 'i' suit le 'g', et un 'e' ou un 'o' suit le 'j'.

Finalement, la requête a pris la forme suivante :

"el.*|je.*|ge.*|gi.*|jo.*"%c

Cela signifie que l'on cherche la forme 'el' ou 'je' ou 'ge' ou 'gi' ou 'jo' suivie de n'importe quel caractère (.) qui peut se répéter plusieurs fois sans limite (*), la requête n'étant pas sensible à la casse (%c).

Evidemment il reste du bruit : on a récolté une liste de 553 formes (que l'on peut classer par fréquence ou par ordre alphabétique. Pour la plupart il est aisé de décider si elles sont pertinentes ou non, mais pour quelques-unes il est nécessaire d'opérer un retour à leur contexte d'occurrences. Pour cela, il suffit de cliquer sur la forme pour accéder aux différentes occurrences sous forme de concordance :

[66] Sans spécification de ma part, la personne 3 désignera désormais les personnes 3 et 6.

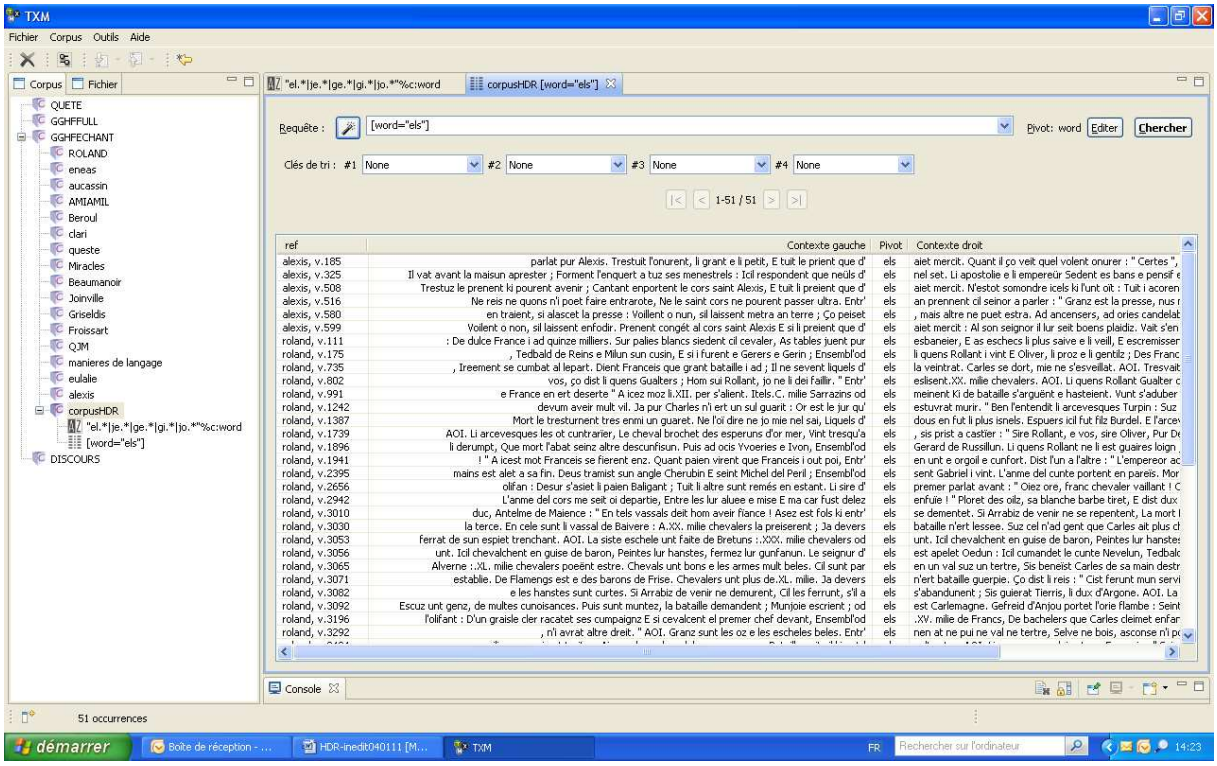


Figure 1 : Extrait de la concordance de *els* dans le corpus

Il apparaît ainsi que les 52 occurrences de *els* ne sont pas pertinentes, car elles correspondent toutes à une forme complément (‘eux’). C’est en revanche plus complexe pour *el* : sur les 586 occurrences, la plupart correspondent à l’enclise ‘en + le’, mais, dans certains textes[67], il s’agit parfois bien du pronom P3.

On a donc décidé d’inclure cette forme dans les requêtes, mais en restreignant ses occurrences à celles de pronom personnel. Cela ne posait pas de problème pour les textes dont l’étiquetage est fiable ; pour les autres en revanche rien ne garantissait a priori que la valeur attribuée à *el* (pronom personnel ou autre) était correcte. A l’aide des requêtes élaborées pour collecter les constructions à sujet préverbal et postverbal (voir 3.4. ci-dessous), on a vérifié qu’il n’y avait pas d’occurrences de *el* doté d’une valeur différente de ‘pronom personnel’ qui serait en fait un pronom. Il s’est avéré que toutes les occurrences non pronominales de *el* n’étaient effectivement pas, dans les contextes requis, des pronoms personnels : en restreignant la requête aux *el* pronoms personnels, on ne risquait pas de passer à côté d’occurrences pertinentes. On risquait simplement de collecter des occurrences indésirables (des *el* étiquetés pronom personnel à tort).

[67] Il s’agit des textes suivants : *Eneas*, *Tristan* de Beroul, les *Miracles* de Coinci, *Ami et Amil*, les *Quinze joyes de mariage.*

Voici la liste des formes de pronoms sujets, avec, à titre indicatif, leur fréquence totale dans le corpus, tous contextes confondus (subordonnées, interrogatives…) :

Forme	**Fréquence**		**Forme**	**Fréquence**
el	525[68]		G	1
El	61		g'	23
el'	1		G'	14
ele	414		ge	160
Ele	39		Ge	57
ELE	1		gel	29
elë	1		Gel	16
eles	45		Ges	4
Eles	2		ges	2
ell	1		Gié	1
Ell'	1		gié	1
elle	713		j'	302
Elle	22		J'	36
elles	96		je	2402
Elles	4		Je	316
total /ele/	*886*		Jel	12
			jel	4
il	8479		Jes	1
Il	507		jez	1
IL	1		jo	86
ils	98		Jo	55
Ils	6		jol	7
ilz	143		Jol	1
ilz	14		jon	1
total /il/	*9248*		jos	2
			jou	26
total	***10134***		Jou	1
			jous	1
			total	***3307***

Tableau 3 : Liste des formes de pronoms sujets

3.2.3. Des textes différemment enrichis

Alors qu'il est peu envisageable de repérer des constructions à sujet nominal sans que l'information soit syntaxiquement encodée[69], il est possible de repérer des constructions à sujet pronominal dans la mesure où les pronoms sujets constituent une liste finie de formes. Cela ne signifie pas pour autant que c'est une tâche aisée, cela pour plusieurs raisons. Tout d'abord, les sujets ne nous intéressent que s'ils se

[68] Ce chiffre inclut un nombre important de formes correspondant à « en + le ».
[69] Au moins en français médiéval où la position du sujet n'est pas fixe.

trouvent en proposition indépendante déclarative, or ils sont très nombreux en proposition subordonnée. Par ailleurs, si les formes /je/ et /il/ instancient toujours la fonction sujet, ce n'est pas le cas pour /elle/ qui peut aussi instancier une fonction de complément. Travailler sur des textes étiquetés morpho-syntaxiquement permet de dépasser en partie cette difficulté. En effet, la possibilité de repérer le verbe conjugué permet d'éliminer beaucoup de bruit, comme nous le verrons plus loin. Dernière difficulté, et la plus importante : l'étude porte aussi sur les constructions à sujet non exprimé, pour lesquelles il n'est évidemment pas possible de s'appuyer sur la forme des sujets pronominaux. Si l'étiquetage morpho-syntaxique permet de s'appuyer sur le verbe, nous allons voir qu'il convient de développer des stratégies de repérage assez fines, afin de limiter le bruit : ce ne sont évidemment pas tous les verbes que nous souhaitons collecter.

Les textes du corpus bénéficient de trois types d'enrichissement. Trois d'entre eux (*Roland*, la *Queste del Saint Graal*, abr. *Queste*, et les *Quinze Joyes de Mariage*, abr. *Quinze Joyes*) ont été étiquetés avec le jeu d'étiquettes morphosyntaxique Cattex, progressivement élaboré en collaboration avec les membres de l'équipe de la *BFM*[70]. Ce jeu comporte 60 étiquettes, qui sont structurées en deux champs principaux : <catégorie> et <type>, les valeurs étant composées de trois lettres, en majuscules pour la catégorie, en minuscules pour le type. Les catégories correspondent pour la plupart aux classiques parties du discours : VER (verbe), NOM (nom), ADJ (adjectif), PRO (pronom), etc. Les types correspondent à des sous-classes des catégories, s'il y a lieu. Les étiquettes ont donc un nom en 3 lettres quand elles sont composées de la seule catégorie (par ex 'PRE' pour préposition) ou, plus fréquemment, de 6 lettres quand elles sont composées de la catégorie et du type. Par exemple :

> 'chevalier' [<catégorie> = 'NOM', <type> = 'com' pour 'commun'] correspond dans notre étiquetage à : 'chevalier' NOMcom (nom commun)
>
> 'esgarda'[<catégorie> = 'VER', <type> = 'cjg' pour 'conjugué'] correspond dans notre étiquetage à : 'esgarda' VERcjg (verbe conjugué)
>
> 'il'[<catégorie> = 'PRO', <type> = 'per' pour 'personnel'] correspond dans notre étiquetage à : 'il' PROper (pronom personnel).

L'étiquetage des trois textes mentionnés ci-dessus est fiable : réalisé dans le cadre d'une campagne d'étiquetage automatique (par une procédure d'apprentissage) menée

[70] En particulier : Céline Guillot, Christiane Marchello-Nizia et Alexei Lavrentiev.

par Serge Heiden (à partir d'un texte qui avait été étiqueté semi-manuellement : *la Mort Artu*), il a ensuite fait l'objet d'une vérification systématique, étiquette par étiquette.

Les autres textes ont été étiquetés dans des conditions analogues, mais n'ont pas encore fait l'objet d'une vérification manuelle. Leur étiquetage n'est donc pas entièrement fiable. A partir des sondages qui ont été réalisés, on peut évaluer que le pourcentage d'erreurs d'étiquetage oscille entre 5 et 30%, la performance étant moins bonne pour certains textes (linguistiquement plus distants du corpus d'étiquetage, en particulier pour ce qui est du lexique), et pour certaines étiquettes. Je ne développerai pas davantage ce point. La poursuite de la vérification de l'étiquetage de l'ensemble des textes est en cours, et un projet visant à analyser les erreurs d'étiquetage est en gestation.

Enfin, l'un des textes du corpus, *Tristan* de Beroul (abr. *Beroul*), bénéficie d'une annotation syntaxique, réalisée dans le cadre du projet ANR franco-allemand *SRCMF* (*Syntactic Reference Corpus of Medieval French*, voir note 52). Grâce à cette annotation, de type dépendantielle, il est possible de repérer facilement les constructions avec un sujet préverbal, postverbal et même non exprimé, et il est possible de restreindre la requête aux propositions indépendantes/principales. L'information concernant le type (déclaratif, interrogatif ou impératif) de la proposition n'est cependant pas encodé à l'heure actuelle, mais le bruit lié à cette sous-spécification reste modéré, les propositions interrogatives ou impératives étant minoritaires, contrairement aux propositions subordonnées. Lorsque j'ai envisagé la présente recherche, je pensais que le nombre de textes annotés serait plus important qu'il ne l'est actuellement. Mais le processus d'annotation est long, en particulier parce qu'il comprend différentes étapes de vérification. Nous ne bénéficierons donc de l'annotation que d'un seul texte, mais celui-ci pourra servir de contrepoint aux autres quant à ses modalités d'exploration, et il permettra d'entrevoir les possibilités qu'ouvre l'enrichissement syntaxique, dont bénéficieront dans les mois qui viennent les autres textes, actuellement en cours d'annotation. On verra cependant que certaines limites actuelles ne rendent pas l'outil aussi idéal que l'on aurait pu l'espérer. Le traitement de *Beroul* sera présenté séparément de celui des autres textes, en raison de sa différence de traitement, l'annotation du texte étant exploitée non pas dans *TXM*, mais dans le logciel *TIGERSearch* (voir note 41).

Voici, résumées ci-dessous, les possibilités de requêtes (pour la présente étude) liées au degré d'enrichissement des textes :

Enrichissement des textes	*Requêtes possibles*
Textes étiquetés morpho-syntaxiquement	Verbes conjugués suivis ou précédés d'un sujet pronominal (bruit et silence)
Textes annotés syntaxiquement	- Verbes conjugués suivis ou précédés d'un sujet pronominal - verbes conjugués sans sujet exprimé (un peu de bruit) - propositions principales / indépendantes

On voit combien l'outil (ici de collecte) utilisé et la nature des données (plus ou moins enrichies) sur lesquelles il opère conditionnent notre démarche. Sans cesse se négocie un compromis entre le souhait de départ (collecter les sujets préverbaux, postverbaux et non exprimés) et ce que nous permet la mise en œuvre de la technique, avec ses forces et ses limites, même si l'on parvient parfois à les contourner, par exemple en développant des règles de repérage fines, ce qui suppose une bonne connaissance au départ de la langue que l'on interroge, afin de limiter le bruit, et surtout le silence.

3.3. La collecte et le tri des données : *TXM-Excel*

Si les stratégies de repérage des constructions ont varié en fonction du caractère plus ou moins fiable de l'étiquetage morpho-syntaxique des textes, la démarche générale a été la même. J'ai conçu des règles qui s'appuient à la fois sur la forme graphique des constructions, et sur leur valeur morpho-syntaxique. Les deux sont encodées dans *TXM* comme deux propriétés, respectivement *word*, et *pos* (pour « part-of-speech »).
Par exemple, si l'on veut collecter tous les *el* qui sont des pronoms personnels, on formulera la requête suivante :

```
[word= "el" & pos="PROper"]
```

Par ailleurs, si l'on a tenté de limiter le « bruit » autant que possible, avec plus ou moins de succès, l'on a surtout veillé à éviter le « silence ». Certes le bruit est coûteux, il est parfois très lourd en temps (et donc très énervant...), mais on en vient à bout, après avoir nettoyé les données, même si c'est parfois au prix de nombreuses heures que l'on aurait aimé passer à des choses plus constructives (je proposerai plus loin une évaluation quantitative du bruit, ainsi que de son coût de traitement). Il arrive

cependant que, en éliminant du bruit, on trouve moyen d'améliorer la requête et donc de limiter ce même bruit dans de futures collectes.

Le bruit est déplaisant donc, mais il n'est pas nuisible, contrairement au silence, pernicieux, contre lequel on ne peut rien (sinon essayer de l'éviter). Or le danger du silence est évidemment de passer à côté d'une construction inédite, inattendue, etc.

La démarche dans l'élaboration des règles de repérage des constructions a donc consisté à trouver le meilleur compromis possible entre la réduction du bruit et l'évitement du silence.

Je présenterai plus loin les règles qui ont été élaborées.

Les requêtes ont donc été effectuées grâce au moteur de recherche de *TXM*, qui utilise le langage CQP[71], lequel permet l'élaboration de requêtes élaborées, comme on le verra. Par ailleurs, on obtient, en aval, les concordances très maniables. On peut ainsi paramétrer la taille des contextes gauche et droit[72], ainsi qu'opérer différents classements, et les combiner : sur la référence (ordre d'apparition dans le texte) comme en figure 2, sur le mot pivot (classement alphabétique), sur les contextes droit et gauche (la figure 3 combine le mot pivot et le contexte gauche). On peut aussi, en cliquant sur une ligne de la concordance, retourner à la page de l'édition qui contient le pivot, surligné (figure 4).

[71] Pour "Corpus Query Processor" : implémenté par la technologie IMS Open Corpus Workbench <http://cwb.sourceforge.net>. Pour une présentation simplifiée de la syntaxe des requêtes, voir le manuel d'utilisation de TXM :
<http://netcologne.dl.sourceforge.net/project/textometrie/documentation/Manuel%20de%20Reference%20TXM%200.5_FR.pdf>

[72] J'ai opté pour des contextes de 40 mots à gauche et 30 mots à droite pour les textes en vers, et de 50 mots à gauche et 30 mots à droite pour les textes en prose, dont les phrases sont souvent plus longues. Mais les captures d'écran ont été faites sur des contextes plus réduits (20 mots de chaque côté) pour des raison de visibilité.

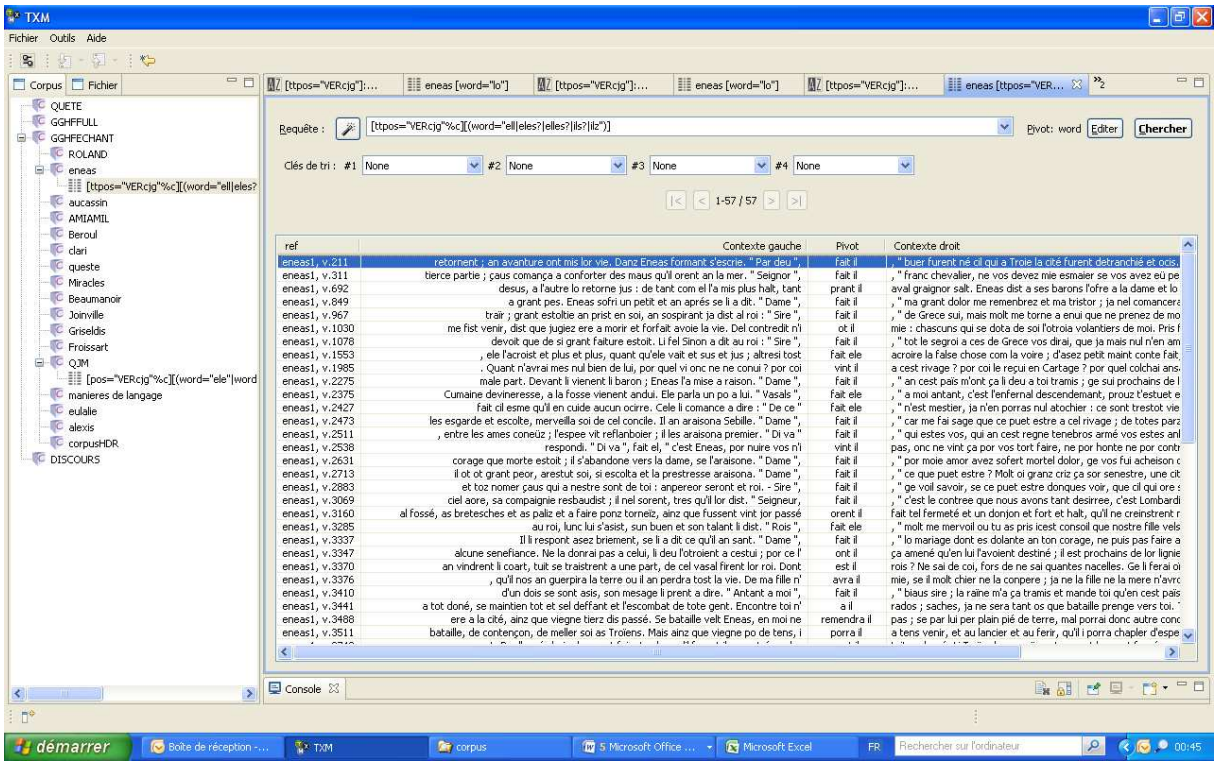


Figure 2 : Extrait de la concordance des sujets postverbaux dans *Eneas*, avec classement sur la référence.

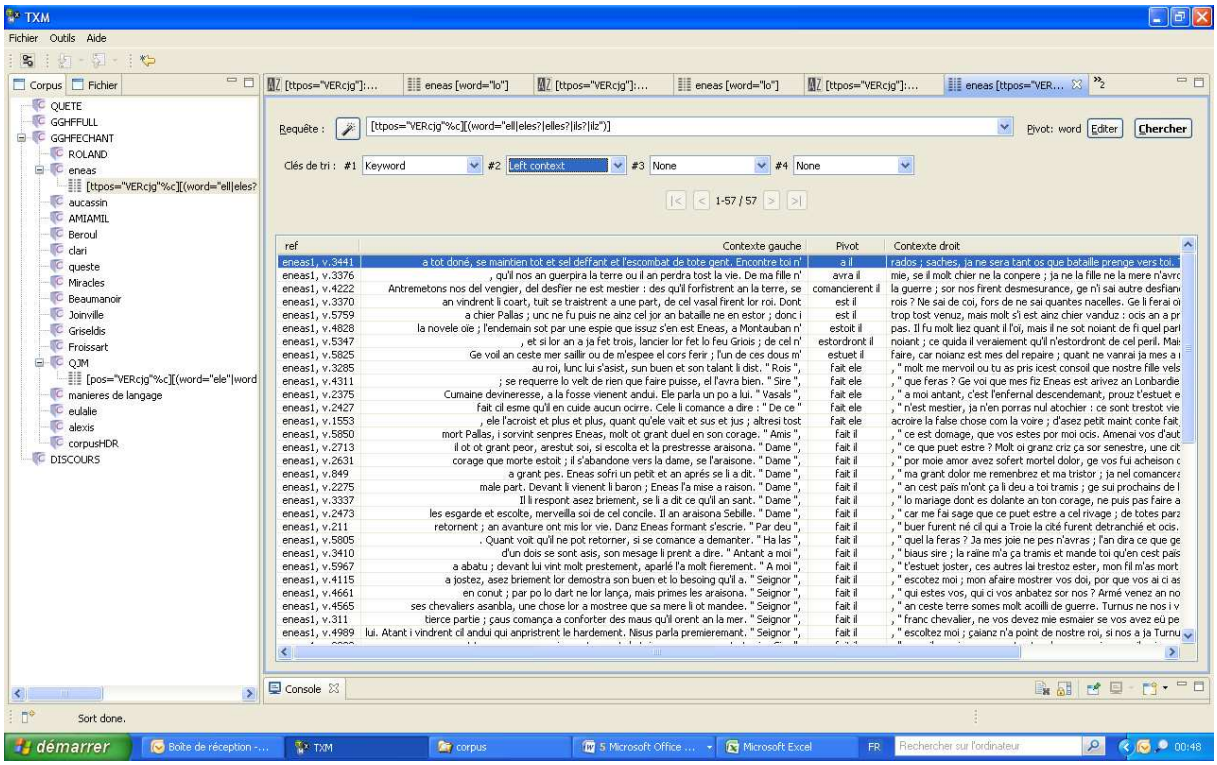


Figure 3 : Extrait de la concordance des sujets postverbaux dans *Eneas*, avec classement sur le pivot et sur le contexte gauche.

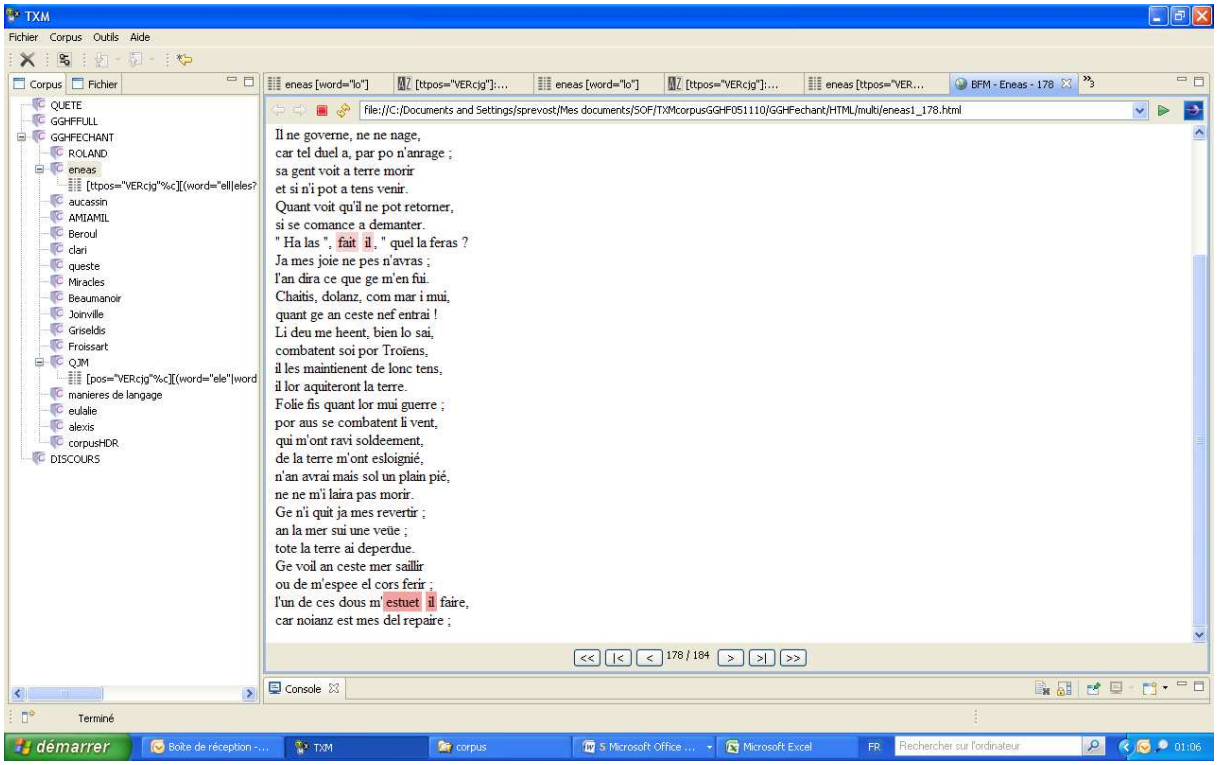


Figure 4 : Affichage d'une page de l'édition d'*Eneas* à partir de la concordance

Mais *TXM* ne permet pas de modifier la concordance générée, et donc d'éliminer le bruit (car il y en a… même si les requêtes sur les sujets postverbaux en produisent globalement peu). Il faut donc externaliser le nettoyage des données, et l'outil le plus immédiat s'est révélé être *Excel*, cela d'autant plus qu'il n'est pas possible, une fois le bruit écarté, de réintégrer les données dans *TXM*, en tout cas pour l'utilisatrice de base que je suis. C'est dommage, parce que l'on perd du coup de nombreuses possibilités d'analyse proposées par *TXM* (en particulier pour ce qui est des fonctionnalités textométriques), et que cela oblige à multiplier les outils (en se tournant vers *Excel*, qui offre une palette de fonctionnalités assez appréciable).

TXM m'a donc servi principalement comme moteur fin de requête, les données collectées ayant ensuite été exportées dans *Excel*. Elles l'ont été sous deux formes : avec un classement sur la référence, et avec un classement sur le contexte gauche. En effet, contrairement à *Excel*, *TXM* opère son tri alphabétique du contexte gauche sur le dernier mot de celui-ci, c'est-à-dire celui qui précède immédiatement le pivot de la requête. *Excel* opère son tri sur le premier mot de la colonne « contexte gauche », ce

qui présente assez peu d'intérêt dans une perspective linguistique[73]. *TXM* a été conçu pour travailler sur des données textuelles, pas *Excel*…

Dans un premier temps, les requêtes ont été conçues pour les textes dont l'étiquetage est fiable, puis elles ont été adaptées pour les autres textes. Les deux séries de tests ont été réalisés sur *Roland* et sur *Eneas.*

3.4. L'élaboration des requêtes

Il a été élaboré trois requêtes principales, pour collecter respectivement les constructions à sujet postverbal, à sujet préverbal et sans sujet exprimé.
Les requêtes pour appréhender les sujet préverbaux et postverbaux ont cependant d'emblée, chacune, été scindées en deux requêtes, afin de collecter séparément P1 et P3, et d'éviter un tri ultérieur.

3.4.1. Les constructions à sujet postverbal

3.4.1.1. *Roland*

La requête pour les sujets postverbaux est *a priori* assez simple, dans la mesure où le sujet pronominal est presque toujours immédiatement conjoint au verbe. Toutefois, nous savons qu'un pronom complément conjoint peut parfois s'intercaler entre le verbe et le sujet, et exceptionnellement un autre terme (voir note 10). Afin de ne pas risquer d'omettre de telles occurrences, j'ai élaboré une première requête prévoyant la présence de 1 à 3 mots (différents des signes de ponctuation) entre le verbe et le sujet. Cette requête a été appliquée à tous les textes : elle a généré du bruit, mais n'a repéré qu'une seule occurrence d'élément intercalé entre le verbe et son sujet (dans les *Miracles* de Gautier de Coinci). Dans la perspective de l'extension ultérieure du corpus, il faudrait renouveler cette opération.

Il a donc finalement suffi de formuler une requête « verbe conjugué immédiatement suivi de l'une des formes pronominales recherchées ». Afin de réduire la collecte des propositions interrogatives ou subordonnées, l'on a exclu la présence devant le verbe

[73] Pour cette étude, la fonctionnalité de *TXM* s'est cependant révélée moins utile que prévu. En effet la formulation de mes requêtes est telle que, dans la zone pivot de la concordance, le verbe ou le sujet pronominal est précédé d'un terme. Le classement sur le contexte gauche opère donc, non pas sur la forme qui précède immédiatement la construction, mais sur celle d'avant (la pénultième).

de tout mot interrogatif (.*int) ou subordonnant (.*sub), relatif (.*rel). Le signe ‘ !’ est un opérateur qui marque l’exclusion.

On aimerait pouvoir exclure aussi les cas où un ou plusieurs mots est/sont intercalé(s) entre le mot subordonnant/interrogatif (en particulier des pronoms compléments) et le verbe, mais il n’est pas possible de formuler une requête dans laquelle une forme exclue est suivie de la présence facultative d’une autre forme. Il faudrait donc procéder en deux temps : formuler une première requête permettant d’éliminer les mots subordonnants/interrogatifs directement suivis du verbe, puis formuler une seconde requête, appliquée au sous-corpus des occurrences collectées par la première : celle-ci repèrerait les cas où un ou plusieurs mots est/sont intercalé(s) entre le mot subordonnant/interrogatif et le verbe, et permettrait d’éliminer les mots subordonnants/interrogatifs restants. Mais il ne m’était pas possible, comme je l’ai déjà indiqué, de réinjecter dans *TXM* le résultat d’une collecte de données pour opérer dessus de nouvelles opérations.

Voici ci-dessous la requête élaborée pour P1.

Indiquons d’emblée que :
- tout ce qui se trouve à l’intérieur d’une paire de crochets dénote *une* occurrence,
- la succession d’expressions entre crochets dénote une succession d’occurrences,
- le signe ‘|’ marque la disjonction

```
 [!(pos=".*int|.*sub|.*rel")]
[pos="VERcjg"%c]
[word="g’|ge?|gié|gel|ges|gen|j'|je?|jel|jes|jen|jos?|jon|jol|jo
us"]
```
[74]

Voici la requête élaborée pour P3 :

```
[!(pos=".*int|.*sub|.*rel")]
[pos="VERcjg"%c]
[word="el|elë|el’|ell|eles?|elles?|ils?|ilz"&pos="PROper"]
```

Il convient de faire quelques remarques.

a) le moteur de recherche ne tient pas compte, par défaut, des limites de phrase : il faut le lui spécifier, ce qui n’a pas été fait ici. De plus, les signes de ponctuation sont traités comme les autres mots. Par conséquent, exclure devant le verbe tout mot subordonnant ou interrogatif (ce qui implique donc, dans le langage adopté,

[74] Les retours à la ligne ont pour seul but de faciliter la lecture.

l'occurrence d'une autre « mot », forme lexicale ou signe de ponctuation) n'empêche pas la collecte des verbes en première position. Le seul cas qui ne serait pas « reconnu » par la requête serait celui d'un verbe en début absolu de texte : il n'y en a pas.

b) la spécification, pour P3, de 'pos="PROper"' permet d'éviter les occurrences indésirables de *el* = « en + le » (voir 3.2.2. ci-dessus). Du même coup elle neutralise les « il » impersonnels (assez rares).

c) le signe '?' présent dans les requêtes ci-dessus est un opérateur qui spécifie que le caractère qui le précède peut être présent ou non. Ainsi "ils?" permet d'attrapper *ils* mais aussi *il*. La requête est ainsi plus économique dans sa formulation.

d) le signe '%c' est aussi un opérateur, qui neutralise la casse de ce qui le précède : la forme peut contenir des majuscules ou des minuscules. Nous avons utilisé cet opérateur pour le verbe, qui, s'il se trouve en début de proposition, débute par une majuscule (et les occurrences de verbe en 1ère position ne sont pas rares dans *Roland*). Il n'est en revanche pas nécessaire d'utiliser cet opérateur pour le sujet, qui, postverbal, ne peut commencer par une majuscule.

Pour *Roland*, on obtient 19 occurrences de P1, dont 1 n'est pas pertinente (pronom interrogatif suivi du pronom *en*), et 26 occurrences pour P3, dont 6 ne sont pas pertinentes (pronom interrogatif suivi d'un autre mot dans 3 cas, interrogation sans mot interrogatif dans 2 cas, et subordonnée sans mot subordonnant dans 1 cas : *Pui li* ***dites il*** *n'en irat, s'il me creit* (v. 2753).

Le bruit représente donc 5,3% des occurrences collectées pour P1, et 23% pour P3.

3.4.1.2. *Eneas*

Comme on a pu le voir pour *Roland*, les requêtes s'appuient prioritairement sur le verbe. Se passer du verbe est très coûteux. J'ai donc pris le parti de m'appuyer aussi sur les verbes pour les textes dont l'étiquetage n'est pas pleinement fiable. Afin d'évaluer la qualité de l'étiquetage, on a dressé l'index des formes ayant la propriété 'VERcjg'[75] et passé en revue les 141 formes (sur 1 742, qui correspondent à 6 032 occurrences) qui avaient une fréquence supérieure à 5 (seuil arbitraire en deça duquel

[75] Pour les textes dont l'étiquetage n'a pas été vérifié, la propriété qui encode Cattex s'appelle 'ttpos', et non 'pos', ce qui explique la variation dans les requêtes, en fonction des textes (étiquetage vérifié ou non).

j'ai estimé le bruit moins coûteux que la complexification de la requête). Celles ayant un étiquetage erroné se sont révélées assez peu nombreuses, 6 en tout (parmi les plus fréquentes : *lo* : 115 occurrences ; *Troie* : 39; *Eneas* : 33). Face à ces formes indésirables, deux options étaient possibles : les exclure systématiquement des requêtes, ou bien faire le pari que les contraintes de la requête (présence d'un sujet pronominal en particulier) les excluraient naturellement. J'ai choisi la seconde option, qui s'est révélée payante : ces formes ne se sont retrouvées que rarement dans les occurrences collectées.

Il a donc été établi que le bruit concernant les verbes était réduit. Il restait à apprécier le silence, c'est-à-dire la part de verbes non étiquetés comme tels, suivis ou précédés d'un sujet pronominal[76].

Pour cela j'ai élaboré une autre requête, moins contraignante que celle utilisée pour *Roland*, puisqu'elle autorise la présence de mots subordonnants ou interrogatifs, et vise donc tous les types de propositions. Après avoir dénombré les occurrences collectées[77], j'ai formulé une requête visant les seuls sujets pronominaux (fiables, puisque correspondant à une liste de formes finie), pour évaluer, grâce au différentiel éventuel entre les deux requêtes, les séquences sujet-verbe ou verbe-sujet qui seraient passées à travers le filet : dans *Eneas*, il n'y en avait pas[78].

Il restait ensuite à évaluer la fiabilité de l'étiquetage des mots subordonnants/relatif/interrogatifs. La liste dressée dans l'index ne comportait pas d'intrus, mais se posait à nouveau la question du silence (silence néanmoins moins dommageable dans ce cas, puisqu'il s'agit d'*exclure* ces formes des résultats des requêtes : qu'il y ait du silence dans leur repérage créera simplement du bruit dans les résultats des requêtes). Pour évaluer ce silence l'on a fait une requête pour P1 postverbal, en excluant devant le verbe les formes ayant une valeur de propriété [ttpos=".*int|.*sub|.*rel"], puis l'on a réitéré la requête en excluant non plus des valeurs de propriété, mais des formes : [word="qu.*|k.*|s'|se|com|coi|dont"]. La seconde formulation s'est avérée plus efficace, ce qui signifie que certaines des occurrences de ces formes (formes qui apparaissent dans l'index

[76] La démarche vaut donc pour les risques de silence concernant les séquences VSp ET les séquences SpV.

[77] Le test a été limité à P1 et P3 au masculin : P3 au féminin est en effet ambigu, il peut avoir une fonction de complément.

[78] A nouveau (cf. 3.4.1.1.) il apparaît qu'une fonction du logiciel serait particulièrement utile : pouvoir exclure des résultats d'une requête ceux d'une requête précédente, en l'occurrence pourvoir exclure d'une requête sur les formes pronominales les occurrences récoltées par une requête sur les constructions « sujet-verbe »/ « verbe-sujet ».

[ttpos=".*int|.*sub|.*rel"] sont passées à travers les mailles de l'étiquetage. Il a donc été décidé, pour le repérage de ces formes interrogatives et subordonnantes, de conjuguer chaînes de caractères et valeurs de propriété 'ttpos'.

Les requêtes ont finalement pris la forme suivante :

Pour P1 :

```
[!(word="qant|quant|q'|qi|qe.*|qoi|qu'|que.*|qui|quoi|k'|ke.*|ki
|koi|s'|se|com.*|coi|dont"%c|ttpos=".*int|.*sub|.*rel")]
[ttpos="VERcjg"]
[word="g'|ge?|gié|gel|ges|gen|j'|je?|jel|jes|jen|jos?|jon|jol|jo
us"]
```

Pour P3 :

```
[!(word="qant|quant|q'|qi|qe.*|qoi|qu'|que.*|qui|quoi|k'|ke.*|ki
|koi|s'|se|com.*|coi|dont"%c|ttpos=".*int|.*sub|.*rel")]
[ttpos="VERcjg"%c]
[(word="el"&ttpos="PROper")|word="elë|el'|ell|eles?|elles?|ils?|
ilz"]
```

On voit que pour P3, contrairement à la requête élaborée pour *Roland*, texte à l'étiquetage fiable, la restriction des formes pronominales à la valeur 'PROper' n'est appliquée qu'à *el*, des erreurs d'étiquetage pouvant affecter les autres formes (non vérifiées), en particulier le pronom *il* dont la distinction entre valeurs personnelle et impersonnelle est mal gérée par le moteur d'étiquetage automatique.

Il a ainsi été collecté 20 occurrences de P1, dont 8 ne sont pas pertinentes (structures interrogatives, sans mot interrogatif ou bien avec mots interrogatif suivi d'une autre forme, par exemples : *Por coi ne **sui ge** donc ocise ?*, (v. 1687*)*, c'est-à-dire que le bruit représente 40% des occurrences collectées.

Pour P3 il a été collecté 62 occurrences, dont 3 ne sont pas pertinentes (un pronom impersonnel, et 2 subordonnées : *mais bien avoit dit la semaine **que ce savoit il bien** sanz faille qu'il ne morroit pas an bataille* (v. 5067). Le bruit est donc assez réduit : 4.8%.

3.4.2. Les constructions à sujet préverbal

La requête destinée à repérer les sujets préverbaux est moins simple à concevoir : elle risque de générer du bruit, et surtout du silence. En effet, d'une part le pronom

préverbal n'est pas nécessairement conjoint au verbe : il peut en être séparé, comme en français moderne, par les pronoms compléments :

(24) Bel sire Guenes, ço li ad dit Marsilie, **Jo *vos* ai fait** alques de legerie (*Chanson de Roland*),

mais aussi par d'autres éléments :

(25) et gardez qu'il ne soit a nul home mortel conté que vos l'aiez veü en ceste voie, ne **ge *endroit moi* n'en parlerai ja**. (*La Mort Artu*).

3.4.2.1. *Roland*

La requête a été conçue de manière à :

a) éviter les contextes de propositions subordonnées, d'où l'exclusion devant le sujet de mot subordonnant, avec ou sans majuscule :

```
[!(pos=".*sub|.*rel"%c)]
```

b) prévoir l'insertion, entre le sujet et le verbe, de pronoms compléments ou de la négation, leur nombre pouvant varier entre 0 et 4 (ce qu'indique ci-dessous :`{0,4}`)[79] :

```
[(word="me|te|se|le|la|li|les|lor|lur|leur|m'|t'|s'|l'|nos|vos|n
oz|voz|nus|vus|en|i|ne|n'|nen|nel|nes|si|mie")]{0,4}
```

c) prévoir l'insertion entre le sujet et le verbe de 0 à 8 formes qui ne soient pas un signe de ponctuation ou un élément subordonnant (cela pour éviter un changement de proposition et que le moteur aille chercher un verbe conjugué hors de la proposition) :

```
[!(pos="PON.*|.*sub|.*rel")]{0,8}
```

Les requêtes ont finalement pris la forme suivante[80] :

[79] *A priori* on ne trouve pas plus de 4 éléments de ce type. Skårup (1975 :129) considère que les combinaisons de trois pronoms sont rares et qu'il se trouve toujours dans ce cas *en* ou *i* parmi eux. La présence possible, en outre, de la négation, conduit à un maximum de 4 éléments.

[80] Une fois les requêtes élaborées (et bien qu'elles soient assurément encore perfectibles), on est toujours un peu étonné, voire déçu, du temps que l'on a passé à les concevoir, à coups de tests successifs avec vérifications de centaines d'occurrences à la clé. Elles ont l'air d'une telle évidence… Un dilemme parfois se pose, en cours d'élaboration : une requête, que l'on pensait assez performante, dont on était assez satisfait, se révèle perfectible, et, ainsi améliorée, capable d'éliminer du bruit et/ou du silence : faut-il mettre à la poubelle les données déjà collectées, reprendre à zéro ? On retrouve ici un problème analogue à ceux déjà évoqués à deux reprises : l'impossibilité de rétroaction sur le corpus.

P1 :

```
[!(pos=".*sub|.*rel"%c)]
[word="g’|ge?|gié|gel|ges|gen|j'|je?|jel|jes|jen|jos?|jon|jol|jo
us"%c]
[word="me|te|se|le|la|li|les|lor|lur|leur|m’|t’|s’|l’|nos|vos|no
z|voz|nus|vus|en|i|ne|n’|nen|nel|nes|si|mie"]{0,4}
[!(pos="PON.*|.*sub|.*rel")]{0,8}
[pos="VERcjg"]
```

P3

```
[!(pos=".*sub|.*rel"%c)]
[word="el|elë|el’|ell|eles?|elles?|ils?|ilz"%c & pos="PROper"]
[word="me|te|se|le|la|li|les|lor|lur|leur|m’|t’|s’|l’|nos|vos|no
z|voz|nus|vus|en|i|ne|n’|nen|nel|nes|si|mie"]{0,4}
[!(pos="PON.*|.*sub|.*rel")]{0,8}
[pos="VERcjg"]
```

Le bruit est globalement réduit : pour P1, sur les 85 occurrences collectées, 83 sont pertinentes (2,3% de bruit); les 2 occurrences indésirables tiennent à la mise en relation du pronom sujet avec un verbe subséquent dont il ne dépend pas (ce qui est permis dans les deux cas du fait de l’absence de ponctuation et de mot subordonnant entre les deux). Par exemple : *Deus, dist li quens, or ne sai* ***jo*** *que* ***face***. (*Roland*, v.1982)

Pour P3, sur les 63 occurences collectées, 4 ne sont pas pertinentes, le bruit est de 6.3% (2 subordonnées, 1 *il* impersonel et 1 coordination de ‘vos’ et ‘il’ : *ne vos ne il n’i porterez*)[81].

3.4.2.2. *Eneas*

L’adaptation des requêtes pour *Eneas* a consisté, pour l’exclusion des mots subordonnants avant le sujet et entre le sujet et le verbe, à associer des formes et des valeurs de propriété ‘ttpos’ :

```
[!(ttpos=".*sub|.*rel"|word="k.*|q.*|s'|se|com.*"%c)]
```

En revanche on a gardé la valeur de propriété "PON.*", l’étiquetage des signes de ponctuation étant fiable (l’étiqueteur est performant dans la mesure où il s’agit d’une liste fermée de formes non ambiguës).

[81] Les cas de pronoms coordonnés n’ont été traités que lorsque le verbe est au pluriel de la personne 1 ou 3, c’est-à-dire égal à une personne 4 ou 6.

Par ailleurs, pour P3, on a restreint la spécification de la valeur « PROper » au seul *el (*voir 3.4.1.1.),, comme pour les sujets postverbaux.

Voici les requêtes.

P1 :

```
[!(ttpos=".*sub|.*rel"|word="qant|quant|q’|qi|qe.*|qoi|qu’|que.*
|qui|quoi| k’|ke.*|ki|koi|s'|se|com.*|coi|dont"%c)]
[word="g’|ge?|gié|gie|gel|ges|gen|j'|je?|jé|jë|jel|jes|jen|jeol?
|jos?|jon| jol|joul?|jous"%c]
[word="me|te|se|le|la|li|les|lor|lur|leur|m’|t’|s’|l’|nos|vos|no
z|voz|nus| vus|en|i|ne|n’|nen|nel|nes|si|mie"]{0,4}
[!(ttpos="PON.*|.*sub|.*rel"|word="qant|quant|q’|qi|qe.*|qoi|qu’
|que.*|qui|quoi|k’|ke.*|ki|koi|s'|se|com.*|coi|dont")]{0,8}
[ttpos="VERcjg"]
```

P3 :

```
[!(ttpos=".*sub|.*rel"|word="qant|quant|q’|qi|qe.*|qoi|qu’|que.*
|qui|quoi| k’|ke.*|ki|koi|s'|se|com.*|coi|dont"%c)]
[(word="el"&ttpos="PROper")|word="elë|el’|ell|eles?|elles?|ils?|
ilz"%c]
[word="me|te|se|le|la|li|les|lor|lur|leur|m’|t’|s’|l’|nos|vos|no
z|voz|nus| vus|en|i|ne|n’|nen|nel|nes|si|mie"]{0,4}
[!(ttpos="PON.*|.*sub|.*rel"|word="qant|quant|q’|qi|qe.*|qoi|qu’
|que.*|qui|quoi|k’|ke.*|ki|koi|s'|se|com.*|coi|dont")]{0,8}
[ttpos="VERcjg"]
```

Pour P1, sur les 92 occurrences collectées, 9 ne sont pas pertinentes (bruit : 9.8%) : 8 d’entre elles correspondent à des subordonnées, et 1 correspond à une interrogative, avec un étiquetage fautif de l’adjectif *chaitis*, analysé comme verbe : *Por coi m'en tornai **ge chaitis** ?* (v. 214)

Pour P3, le bruit est plus important (13.4%) : sur 268 occurrences, 36 ne sont pas pertinentes (occurrences de ‘ele’ complément : *chascune d'**eles esgarda** et longuemant les avisa* (v.123) ; occurrences en subordonnées et *il* impersonnel).

3.4.3. Les constructions à sujet non exprimé

Le traitement des constructions à sujet non exprimé est doublement complexe : il l’est parce qu’il s’agit d’une structure fréquente, et parce que la requête, aussi raffinée soit-elle, génère beaucoup de bruit, du fait même que l’on ne peut s’appuyer sur la présence du sujet pronominal, et que l’on collecte tous les sujets nominaux. On ne peut écarter ces derniers en excluant les noms communs ou les noms propres, car on

exclurait alors du même coup de nombreux compléments nominaux préverbaux ou postverbaux.

Voici ci-dessous la requête qui a été conçue pour collecter les constructions sans sujet exprimé, et les choix envisagés pour les traiter.

3.4.3.1. *Roland*

La requête a été conçue pour :

a) refuser la présence d'un mot subordonnant et /ou d'un pronom sujet devant le verbe :

```
[!(pos=".*sub|.*rel")]
[!(word="g'|ge?|gié|gie|gel|ges|gen|j'|je?|jé|jë|jel|jes|jen|jeo
l?|jos?|jon|jol|joul?|jous|elë|el'|ell|eles?|elles?|ils?|ilz"%c|
pos=".*sub|.*rel")]
```

b) exclure au mieux les verbes de personnes 2, 4 et 5. Pour cela une liste de désinences a été conçue pour *Roland*, progressivement enrichie au fil des textes :

```
[(pos="VERcjg")&!(word=".*(z|as|és|es|oms|om|uns|ons|un|on)")]
```

c) exclure les sujets pronominaux postverbaux, la restriction ne s'appliquant pas à /elle/, susceptible d'être complément postverbal du verbe :

```
[!(word="g'|ge?|gié|gie|gel|ges|gen|j'|je?|jé|jë|jel|jes|jen|jeo
l?|jos?|jon|jol|joul?|jous|ils?|ilz")]
```

La requête a donc pris la forme suivante :

```
[!(pos=".*sub|.*rel")]
[!(word="g'|ge?|gié|gie|gel|ges|gen|j'|je?|jé|jë|jel|jes|jen|jeo
l?|jos?|jon|jol|joul?|jous|elë|el'|ell|eles?|elles?|ils?|ilz"%c|
pos=".*sub|.*rel")]
[(pos="VERcjg")&!(word=".*(z|as|és|es|oms|om|uns|ons|un|on)")]
[!(word="g'|ge?|gié|gie|gel|ges|gen|j'|je?|jé|jë|jel|jes|jen|jeo
l?|jos?|jon|jol|joul?|jous|ils?|ilz")]
```

Elle a permis la collecte de 3 889 occurrences.

3.4.3.2. *Eneas*

L'adaptation de la requête pour un texte à l'étiquetage non fiable a consisté à conjuguer chaînes de caractères et valeur de propriété 'ttpos' pour l'exclusion des mots subordonnants. Il n'est pas possible d'exclure la conjonction *se/s'*, du fait qu'elle

a la même forme que le pronom réfléchi de 3ème personne (plus rarement, dans certains textes, elle peut correspondre à l'adverbe « si »). De même, on a gardé *ou* : bien que la forme corresponde le plus souvent au pronom relatif *où*, il s'agit parfois aussi de la conjonction de coordination. Des sondages ayant prouvé que l'étiquetage automatique ne s'était pas révélé très performant pour distinguer les deux valeurs, il a été jugé préférable de ne pas se fonder sur la catégorisation de *ou*, et de garder toutes les occurrences Cela a pu générer une part non négligeable de bruit dans certains textes, en particulier chez Joinville.
Pour chaque texte, les formes étiquetées « verbe conjugué » à tort (vérification grâce à l'index) ont été exclues de la requête lorsqu'elles dépassaient 5 occurrences (voir remarque à ce propos en 3.4.1.2.).

La requête a pris la forme suivante :

```
[!(ttpos=".*sub|.*rel"|word="qant|quant|q'|qi|qe.*|qoi|qu'|que.*
|qui|quoi|k'|ke.*|ki|koi|s'|se|com.*|coi|dont"%c)]
[!(word="g'|ge?|gié|gie|gel|ges|gen|j'|je?|jé|jë|jel|jes|jen|jeo
l?|jos?|jon|jol|joul?|jous|elë|el'|ell|eles?|elles?|ils?|ilz|qan
t|quant|q'|qi|qe.*|qoi|qu'|que.*|qui|quoi|k'|ke.*|ki|koi|s'|se|c
om|coi|dont"%c|ttpos=".*sub|.*rel")]
[(ttpos="VERcjg")&!(word=".*(z|as|és|es|oms|om|uns|ons|un|on)")]
[!(word="g'|ge?|gié|gie|gel|ges|gen|j'|je?|jé|jë|jel|jes|jen|jeo
l?|jos?|jon|jol|joul?|jous|ils?|ilz")]
```

La requête a permis de collecter 4 396 occurrences.

3.4.4. Le traitement spécifique de la non-expression dans Beroul : *NotaBene-TIGERSearch*

Le projet *SRCMF* (*Syntactic Reference Corpus of Medieval French*) est dédié à l'annotation syntaxique d'un certain nombre de textes d'ancien français de la *BFM* et du *NCA.* L'annotation, de type dépendantielle, est réalisée manuellement à l'aide du logiciel *NotaBene*, développé par Nicolas Mazziotta[82]. Voici ci-dessous un extrait de l'annotation de *Tristan* de Beroul, abr. *Beroul* (fenêtre centrale), avec à droite l'ontologie des catégories utilisées.

[82] *Nota Bene* est un logiciel libre d'annotation. Voir https://sourceforge.net/projects/notabene/

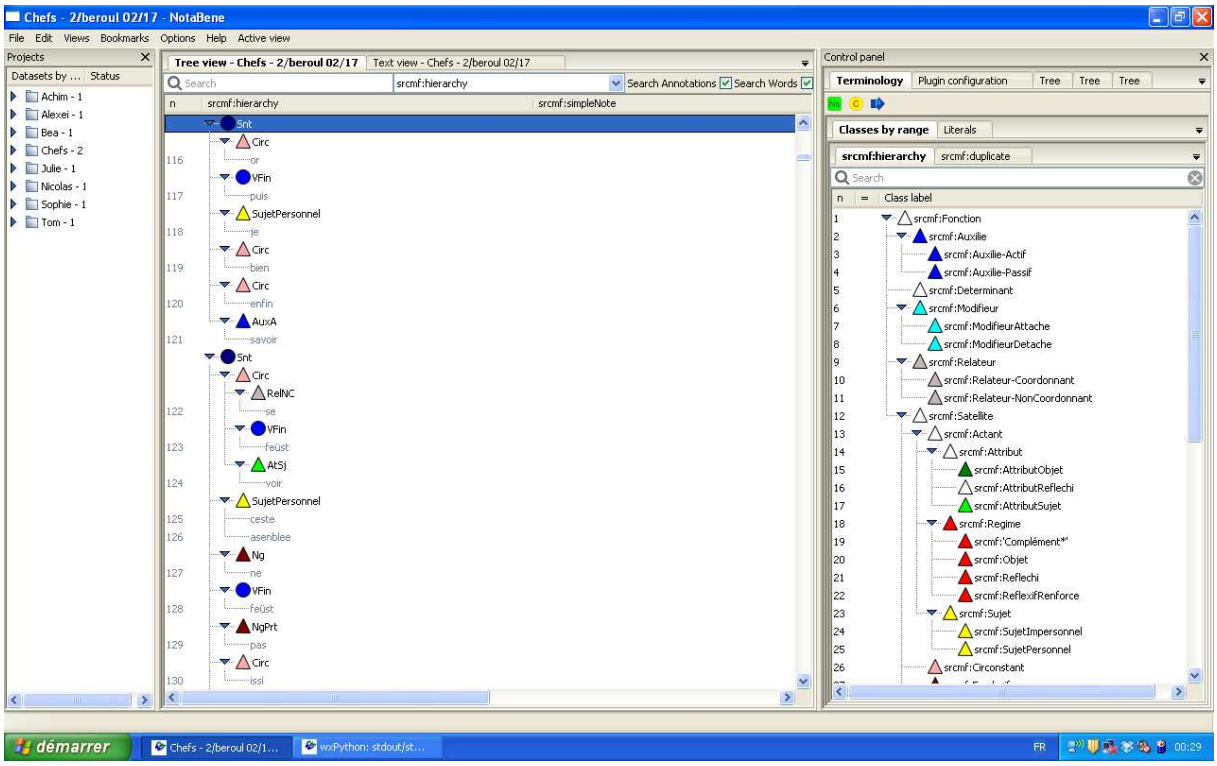


Figure 5 : Extrait d'annotation dans *NotaBene* de *Tristan* de Beroul.

NotaBene n'est pas un outil d'interrogation, et c'est donc un autre logiciel qui est utilisé pour effectué les requêtes, *TIGERSearch* (voir note 41), après que les textes annotés dans *NotaBene* y aient été exportés (exportation réalisée par Nicolas Mazziotta).

La syntaxe des requêtes dans *TIGERSearch* est différente de celle utilisée dans *TXM*[83]. Elle se fonde à la fois sur des relations de dominance entre fonctions, et sur la forme lexicale des mots, que l'on exprime par la propriété « word », qui correspond aux nœuds terminaux des arbres représentés dans *TIGERSearch*. Pour ce qui nous intéresse ici, les fonctions sont celles de verbe fini ('VFin') et de sujet personnel ('SjPer'), et les nœuds terminaux sont '#sujet' et '#verbe', que l'on peut définir plus précisément en spécifiant leur forme. Pour les requêtes sur les sujets exprimés on dresse ainsi la liste des formes, comme cela a été fait dans *TXM* :

```
[word=/[Gg]'|[Gg]e|[Gg]|[Gg]ié|[Gg]el|[Gg]es|[Gg]en|[Jj]'|[Jj]e|
[Jj]|[Jj]el|[Jj]es|[Jj]en|[Jj]os|[Jj]o|[Jj]on|[Jj]ol|[Jj]ous|[Ee
]l|[Ee]lë|[Ee]l'|[Ee]ll|[Ee]les|[Ee]le|[Ee]lles|[Ee]lle|[Ii]ls|[
Ii]l|[Ii]lz/]
```
[84]

[83] Je remercie grandement Achim Stein et Tom Rainsford de m'avoir largement aidée à formuler les requêtes dans *TIGERSearch*, la syntaxe de ces dernières ne m'étant pas encore bien familière.

[84] Liste qui a ensuite été scindée dans 2 requêtes distinctes pour collecter séparément les occurrences de P1 et de P3.

et pour les verbes sans sujet exprimé on peut éliminer certaines formes verbales (celles se terminant par les désinences correspondant aux personnes 2, 4 et 5) :

```
[word != /.*(z|as|és|es|oms|om|uns|ons|un|on)/]
```

Remarques :

a) les expressions regulières sont toujours exprimées entre barres obliques / /, et non pas entre guillemets " " comme dans *TXM*.

b) *TIGERSearch* est sensible à la casse, il faut donc lister toutes les majuscules (au moins pour les sujets préverbaux, susceptibles de se trouver en tête de phrase). Les crochets dans les expressions régulières indiquent qu'un choix s'opère entre les lettres qui se trouvent à l'intérieur : par exemple [Gg] signifie : "soit G, soit g, mais pas les deux".

Le premier avantage de l'exploitation des textes par *TIGERSearch* est que, les propositions indépendantes et principales ayant été annotées en tant que telles dans *NotaBene* (ce sont des 'phrases', labellisées 'Snt' pour 'sentence'), il est possible de restreindre la requête à leur occurrence, et donc d'éliminer les propositions subordonnées. L'information concernant le type (déclaratif, interrogatif ou impératif) de la proposition n'est certes pas encodé à l'heure actuelle, mais le bruit lié à cette sous-spécification reste modéré, les propositions interrogatives ou impératives étant minoritaires, contrairement aux subordonnées.

Voici la requête formulée pour la collecte des sujets pronominaux postverbaux :

```
#vfin:[cat="VFin"] > #sjper:[cat="SjPer"]
& #vfin >L#verbe:[ ]
& #sjper >L#sujet:[word =
/[Gg]'|[Gg]e|[Gg]|[Gg]ié|[Gg]el|[Gg]es|[Gg]en|[Jj]'|[Jj]e|[Jj]|[
Jj]el|[Jj]es|[Jj]en|[Jj]os|[Jj]o|[Jj]on|[Jj]ol|[Jj]ous|[Ee]l|[Ee
]lë|[Ee]l'|[Ee]ll|[Ee]les|[Ee]le|[Ee]lles|[Ee]lle|[Ii]ls|[Ii]l|[
Ii]lz/]
& #verbe.* #sujet
& [cat="Snt"] > #vfin
```

et celle formulée pour la collecte des sujets préverbaux :

```
#vfin:[cat="VFin"] > #sjper:[cat="SjPer"]
& #vfin >L #verbe:[]
& #sjper >L #sujet:[word =
/[Gg]'|[Gg]e|[Gg]|[Gg]ié|[Gg]el|[Gg]es|[Gg]en|[Jj]'|[Jj]e|[Jj]|[
```

```
Jj]el|[Jj]es|[Jj]en|[Jj]os|[Jj]o|[Jj]on|[Jj]ol|[Jj]ous|[Ee]l|[Ee
]lë|[Ee]l'|[Ee]ll|[Ee]les|[Ee]le|[Ee]lles|[Ee]lle|[Ii]ls|[Ii]l|[
Ii]lz/ ]
& #sujet.* #verbe
& [cat="Snt"] > #vfin
```

Remarque :

a) Le signe '>' indique une relation de dominance entre nœuds. Il est ainsi spécifié que, d'une part, le nœud 'verbe fini' domine le nœud 'sujet personnel', et que, d'autre part, ces nœuds dominent respectivement les nœuds terminaux 'verbe' et 'sujet', lesquels ont une relation « lexicale » ('L') avec leur nœud parent.

b) le signe '.' indique une relation de précédence linéaire entre nœuds : `'#verbe.* #sujet'` signifie que le verbe précède le sujet, tandis que `'#sujet.* #verbe'` signifie que le sujet précède le verbe.

Voici l'un des graphes résultant de la requête sur les sujets postverbaux :

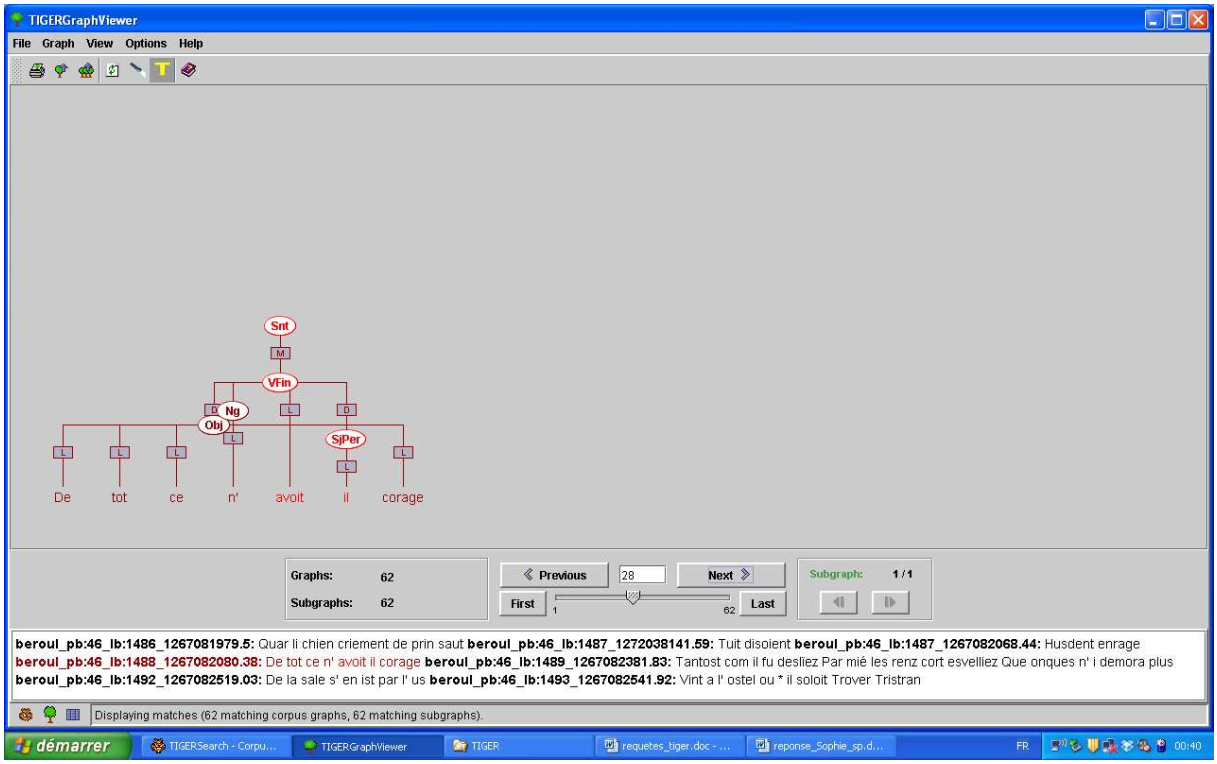


Figure 6 : Graphe de *TIGERSearch* résultant de la requête sur les sujets postverbaux.

Par défaut *TIGERSearch* ne peut pas repérer les constructions à sujets non exprimés. En effet, nous avons fait le choix théorique, dans *SRCMF*, de ne pas enregistrer au

niveau de l'annotation dans *NotaBene* les sujets « zéros ». Or *TIGERSearch* n'est pas capable d'interpréter une requête négative (=absence de sujet).

Pour pallier à cette lacune, Nicolas Mazziotta a mis en place un format de sortie *NotaBene* enrichi. Lors de l'exportation au format *TIGERSearch*, *NotaBene* recherche la présence d'un sujet dans toutes les structures où il est pertinent d'en chercher un (cela est codé « en dur » dans le programme) : celles dont le nœud principal est un verbe portant la catégorie de la personne. S'il n'en trouve pas, il reporte cette propriété de la structure examinée sur sa représentation *TIGERSearch* (dans ce que l'on appelle un « feature » dans le jargon du logiciel). Il devient dès lors possible de faire une requête directe sur ces absences, ce qui n'est possible que parce que les absences sont désormais artificiellement exprimées, et que l'on peut donc faire une requête « positive » sur elle[85].

Pour cela, il faut spécifier dans la requête que la propriété « nodom » (= no dominance) du nœud verbal est 'Sj' (le verbe ne domine pas un sujet).

Voici la requête pour collecter les verbes sans sujet exprimé (et éliminer autant que possible ceux des personnes 2, 4 et 5[86].

```
[cat = "Snt"] > #vfin:[cat = "VFin" & nodom = "Sj"]
& #vfin >L #verbe:[word !=
/.*(z|as|és|es|oms|om|uns|ons|un|on)/]
```

Et voici l'un des 1500 graphes générés par *TIGERSearch* :

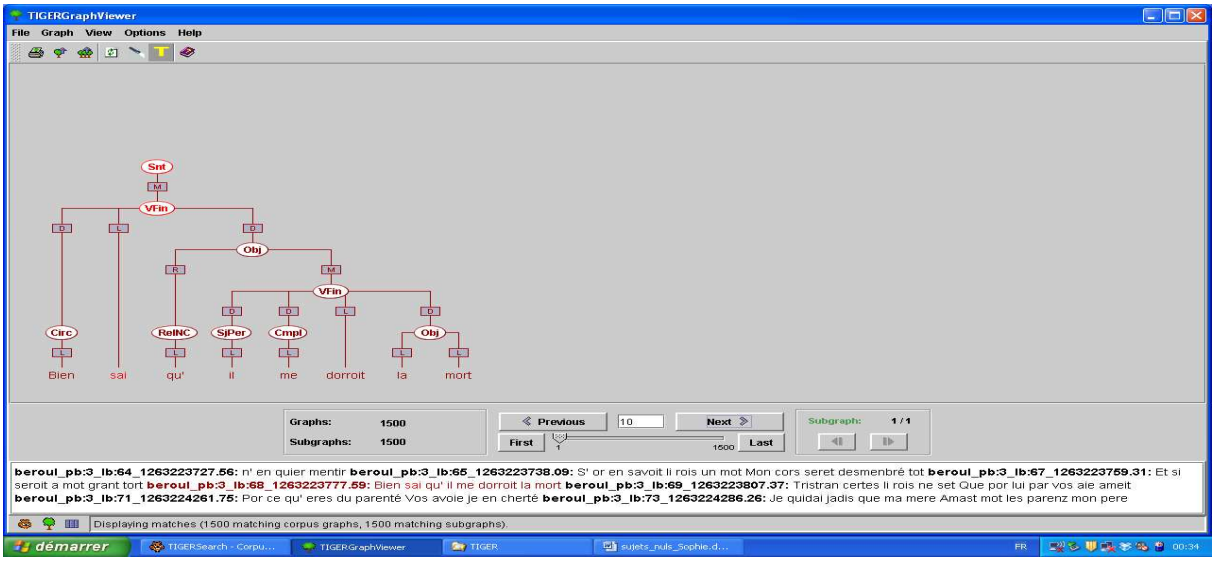


Figure 7 : Graphe de *TIGERSearch* résultant de la requête sur sujets non exprimés.

[85] Je remercie Nicolas Mazziotta de m'avoir « décortiqué » la procédure grâce à laquelle il a pu rendre accessible, dans *TIGERSearch*, les sujets non exprimés.
[86] Elle est due à Tom Rainsford.

Tout serait parfait, si *TIGER Search* ne présentait malgré tout un petit défaut. Il est possible d'afficher dans *TIGERSearch* un contexte gauche et droit de 3 phrases (maximum), comme cela apparaît dans les deux graphes des figures 6 et 7 ci-dessus, dans le bandeau blanc en bas de la fenêtre. Mais malheureusement, à l'heure actuelle, il n'est pas possible d'exporter ces contextes : on ne peut exporter que les phrases correspondant aux graphes affichés. C'est évidemment fort gênant dès lors que l'on veut dépasser l'analyse strictement quantitative.

Par conséquent, dans la mesure où le bruit qui apparaît dans les collectes des sujets exprimés à l'aide de *TXM* est assez réduit, et que ceux-ci sont de toute façon en nombre raisonnable (ce qui permet un « nettoyage »des données peu coûteux), j'ai décidé de conserver la procédure *TXM* pour collecter, dans *Beroul*, les sujets préverbaux et postverbaux. Le recours à *TIGERSearch* a donc été réservé à la collecte des constructions sans sujet exprimé.

Signalons que sur les 1 500 occurrences collectées, 1 277 se sont avérées pertinentes : le bruit est donc de 14.9%. Il s'explique par la collecte indésirable des structures interrogatives et optatives (annotées comme « Phrase » dans *Nota Bene*), et par celle des verbes impersonnels ou des verbes de personnes 2, 4 ou 5 (cela malgré l'exclusion de certaines désinences).

A titre de comparaison, la collecte, dans *TXM*, des verbes sans sujet exprimé, à l'aide de la requête utilisée pour les autres textes, s'élève à 3482 occurrences. Sachant que seules 1 277 occurrences sont finalement pertinentes, le bruit s'élève donc à 63% (et peut-être même un peu plus car il est probable que, en raison du silence, on ne trouve pas les 1 277 occurrences pertinentes parmi les 3482 collectées). D'un côté, bruit de 14.9% et absence de silence, de l'autre, bruit de 63% et silence probable : point n'est besoin d'insister sur l'avantage de l'option *NotaBene-TIGERSearch*, au moins en ce qui concerne l'efficacité de la collecte. Pour ce qui est du problème de l'exportation du contexte, on peut espérer qu'il sera prochainement résolu.

3.4.5. Le traitement des sujets non-exprimés

Il n'était pas envisageable d'analyser de manière exhaustive toutes les occurrences de sujet non exprimé pour l'ensemble du corpus (11 179 en tout) : il a donc été décidé de procéder en deux temps.

Le dénombrement des occurrences de P1 et de P3 a été opéré sur l'ensemble des occurrences collectées. Dans la mesure où le nombre de sujets non exprimés P1 et P3 est destiné à être mis en relation avec le nombre de sujets exprimés P1 et P3, et que le relevé de ces derniers a été opéré sur l'ensemble de chaque texte, il n'était pas envisageable de ne pas procéder de la même façon pour les sujets non exprimés. Extrapoler le nombre de ces derniers à partir d'un échantillon aurait été trop hasardeux, *a fortiori* pour les textes dont l'étiquetage des verbes n'est pas fiable (cela revenait à multiplier le risque que les chiffres avancés soient très éloignés de la réalité). Opérer des calculs statistiques, en particulier le calcul du khi2 en mêlant des valeurs absolues qui sont fiables et d'autres qui le sont beaucoup moins aurait fourni des résultats légitimement contestables.

Par conséquent, j'ai passé en revue l'ensemble des occurrences issues de la requête 'sujets non exprimés', en les codant. Pour *Roland*, j'ai codé l'ensemble des occurrences, pour évaluer la nature du bruit.

J'ai utilisé pour cela un jeu de 7 catégories : 'proposition indésirable (subordonnée, interrogative…)' ; 'pronoms non exprimés autres que P1/P3' ; 'P3 impersonnel non exprimé' ; 'SN' ; 'tout pronom exprimé' ; et enfin, ce que nous cherchions : P1 et P3 non exprimés. Voici les résultats :

P1 non exprimé	**166**
P3 non exprimé	**1341**
autre type de proposition	580
autre personnes non exprimées	51
P3 impersonnel non exprimé	167
SN	1469
P1/2/3/4/5/6 exprimé	115

Tableau 4 : valeurs des occurrences collectées à l'aide de la requête « sujet non exprimé » dans *Roland.*

Le nombre d'occurrences pertinentes (P1 et P3) s'élève à 1507. Le bruit représente donc 61% des occurrences collectées.

Il apparaît, et c'était prévisible au regard de la requête, que le bruit concerne très majoritairement les sujets nominaux.

Il était par ailleurs intéressant d'évaluer si la répartition de ces différentes catégories était homogène dans le texte, en particulier pour P1 et P3. Pour cela on a subdivisé le tableau des 3 889 résultats en trois parties, et dénombré les différentes catégories dans

chacune d'elle. La répartition s'est avérée très régulière pour les P3 non exprimés (402/472/467 respectivement dans chacune des trois parties) et pour les SN (un peu moins de 500 occurrences dans chacune des parties) ; elle l'est en revanche moins homogène pour les P1 (76/34/56), pour des raisons qui sont probablement à chercher du côté narratif.

Pour *Eneas*, et les autres textes, je n'ai codé que les occurrences de P1 et de P3 sans distinguer les occurrences indésirables[87]. Sur les 4 396 occurrences collectées, 1 687 se sont révélées pertinentes : le bruit est donc de 61,6%.

Le temps passé à opérer le codage des occurrences de *Roland* n'a pas été quantifié, dans la mesure où il s'agissait d'un codage spécifique à ce texte. Le coût du nettoyage de fichiers a en revanche été évalué pour les autres textes : en moyenne il faut une heure pour vérifier 700 occurrences d'un texte en vers dans *Excel*, et coder les P1 et P3[88]. Ainsi pour *Eneas*, la seule extraction des occurrences pertinentes de sujets non exprimés aura nécessité 6 heures (auxquelles s'ajoutent quelques pauses, évidemment…). On conçoit combien l'enrichissement des textes, y compris avec des informations sur les « absences », peut se révéler précieux. Dans l'un des textes du corpus, *Tristan* de Beroul, annoté syntaxiquement, les sujets non-exprimés sont encodés. Nous verrons plus loin combien ces informations ont pu alléger l'extraction des données.

Le dénombrement des occurrences de P1 et de P3 a porté sur l'ensemble de leurs occurrences, mais leur analyse qualitative ne concernera qu'un échantillon d'entre elles (et non les 11 179).

L'échantillonnage a été réalisé dans *Excel*, selon une méthode de tirage aléatoire. Il a été retenu, pour P1 et pour P3, un nombre d'occurrences équivalant à un pourcentage de leur nombre total d'occurrences. Pour chaque texte le pourcentage est le même pour P1 et P3, mais il peut varier d'un texte à l'autre. Cette variation est en partie due à la présence d'une « population » trop petite, comparativement aux autres, dans le cas des textes assez brefs, mais aussi, à l'inverse, d'une population trop importante, et donc

[87] Certes, une analyse systématique des erreurs serait intéressante et permettrait probablement d'améliorer la requête. Mais le temps m'a manqué pour réaliser une telle analyse.

[88] Le coût est plus élevé pour les textes en prose, environ 800 occurrences par heure. Cela est dû à ce que les phénomènes d'enchâssement sont plus complexes, les phrases plus longues et l'exploration du contexte gauche prend donc plus de temps.

trop lourde à traiter, dans le cas des textes longs. Par conséquent, l'échantillonnage, initialement prévu pour retenir 25% des occurrences, a pu osciller entre 15 et 40%.

J'ai conscience que l'échantillonnage est une procédure qui comporte des risques, en particulier lorsque l'on travaille en langue ancienne, et dans la perspective de repérer les lieux, les moments et les conditions du changement. Il est donc certain que les résultats qui seront présentés pour l'étude des sujets non exprimés ne présenteront pas le même degré de fiabilité que ceux présentés pour les constructions à sujet préverbal ou postverbal.

3.5. L'analyse quantitative des données : que dénombrer ?

3.5.1. Les objets et les phénomènes à quantifier et mesurer

Le but premier de cette étude est d'évaluer l'évolution de l'expression et de la position du sujet pronominal du 12ème au 14ème siècle (période assez courte, je le rappelle, pour des raisons matérielles de temps, mais qu'il conviendra d'élargir par la suite, en amont et en aval). Il s'agit de les quantifier séparément, mais il s'agit aussi d'essayer de déterminer si ces évolutions sont liées entre elles. Il est en effet avéré que la non-expression et l'inversion ont toutes deux reculé dès la période médiévale, mais il n'a pas été établi dans quelles proportions respectives elles l'ont fait, ni à quelle vitesse. De plus, s'il est souvent suggéré que les deux évolutions sont liées, cette hypothèse n'a pas été étayée de manière rigoureuse par les données chiffrées. C'est dans la perspective de préciser ces différents aspects que nous allons considérer les données chiffrées des deux évolutions, sachant que l'éventuel établissement d'une corrélation statistique entre les deux ne permettra pas nécessairement pour autant de déterminer les modalités exactes de cette corrélation : laquelle des évolutions a influencé l'autre ? Se sont-elles co-influencées ?

L'accent est mis ici sur les deux phénomènes qui se sont marginalisés – non expression et inversion – mais sans oublier, évidemment, qu'ils ne l'ont fait que conjointement au développement de l'expression et de l'antéposition au verbe[89].

Par ailleurs, j'ai évoqué plus haut l'hypothèse avancée par Detges (2003) selon laquelle l'expression du sujet se serait d'abord développée avec P1, pour des raisons

[89] J'emploie délibérément le terme neutre de « conjointement » : je laisse de côté la question de savoir si inversion et non-expression ont reculé *parce que* antéposition au verbe et expression se sont développées ou si au contraire ces deux dernières se sont imposées *parce que* les deux autres perdaient du terrain.

d'expressivité. Si tel est le cas, nous devrions observer, au moins dans les textes les plus anciens, une différence quantitative signifiante entre l'expression des personnes P1 et P3. Aucune hypothèse analogue n'a été avancée, à ma connaissance, pour la position du sujet. Toutefois, dans la mesure où nous testons une possible corrélation entre expression et position, et que nous évaluons par ailleurs le rôle moteur joué par P1, il est utile d'examiner si des différences signifiantes apparaissent entre personnes 1 et 3 au regard de la position.
Ce sont donc différents paramètres qui vont être croisés : expression, position, et personne, selon des modalités différentes.

Mettre en regard les fréquences de non-expression et d'inversion au sein d'un même texte n'a pas grand sens. Il est en revanche intéressant de comparer les fréquences respectives d'un texte à l'autre, de voir si une évolution le long de l'axe chronologique se dessine, si elle est régulière (le cas échéant), et de déterminer si le rapport entre les deux est constant ou au contraire variable.
En revanche, la prise en compte différenciée des personnes 1 et 3 se fera dans un premier temps au sein de chaque texte, pour l'expression d'une part, pour la position d'autre part. On peut en effet considérer que la distribution des personnes au regard de chacun de ces deux phénomènes constitue un micro-système (ce qui n'est pas le cas, *a priori* en tout cas, pour l'expression et la position entre elles). Dans un second temps, il sera intéressant d'observer comment ces micro-systèmes se comportent d'un texte à l'autre, et plus spécifiquement comment ils ont évolué sur l'axe chronologique. Pour cela on considèrera les évolutions respectives des personnes 1 et 3 au regard de l'expression et de la position, évolution qu'il conviendra de rapporter à l'évolution globale, personnes 1 et 3 confondues, de ces phénomènes. Plus précisément, on tentera d'établir s'il existe des convergences, voire des corrélations, d'une part entre l'évolution des personnes 1 et 3 au regard de chacun des phénomènes (dépassant en cela le caractère éventuellement signifiant de la répartition de P1 et P3 au sein de chacun des textes), et d'autre part entre expression et position au regard de chacune des personnes.

L'analyse des données se fera donc au niveau intra-textuel, et inter-textuel. L'approche intra-textuelle se justifie pour ce qui touche aux phénomènes de distribution : autant évaluer le caractère plus ou moins surprenant de la distribution

d'une ou plusieurs variables dans un texte qui constitue un tout homogène[90] a du sens, autant il paraît peu pertinent d'envisager la distribution d'une variable à travers différents textes en cherchant à déterminer parmi eux ceux qui sont « atypiques » au regard de la distribution générale de la variable dans le corpus. Ainsi étudier l'éventuelle spécificité quantitative du sujet postverbal P3 dans *Cligès* comparé à l'ensemble de l'œuvre de Chrétien de Troyes serait pertinent, autant il ne le serait guère de ramener la fréquence dans *Roland* de ce même phénomène à l'ensemble de notre corpus, composé de textes de dates, de genres, de formes, et de dialectes différents[91].

Pour ce qui est de la dimension inter-textuelle de l'analyse, l'interprétation des résultats se fera au regard des différents critères (date, genre, forme et dialecte), mais tous ne jouent pas le même rôle, le paramètre « temps » occupant une place privilégiée. A un premier niveau, le temps joue un rôle neutre, il est la toile de fond sur laquelle nous devons appréhender un phénomène, à différents moments, pour repérer une évolution ou au contraire une permanence[92] (ou bien encore une évolution en « dents de scie », sans directionalité perceptible). A un second niveau, et bien que les deux tendent à se confondre, le temps est plus spécifiquement associé à une évolution, il en est le « facteur » : c'est parce que le temps s'écoule que tel objet se modifie. Pour l'étude des phénomènes ici considérés, le temps joue non seulement le rôle de toile de fond sur laquelle nous les observons, mais il est aussi facteur de changement : la réalité des faits nous montre que, au fil des siècles, la non-expression et l'inversion ont reculé au profit de l'expression et de l'antéposition au verbe. Même si l'évolution ne s'est peut-être pas faite selon une pente continue et régulière (ce qu'il conviendra justement de déterminer), le mouvement général a été inexorable. A ce titre, le paramètre chronologique constitue un critère prévalent, et même si nous faisons l'hypothèse que d'autres critères (forme, genre, dialecte) ont pu intervenir et favoriser, ou au contraire contrecarrer cette évolution, rien ne prouve que ce soit le cas, et les éléments que je pourrai apporter dans la présente étude ne pourront constituer que des indices et non des preuves irréfutables. C'est la raison pour laquelle

[90] On peut évidemment discuter l'homogénéité d'un texte : j'entends qu'un texte constitue une unité rédigée par une même personne, sur un laps de temps assez court, ce qui est le cas dans mon corpus. Au-delà des possibles variations d'un passage à l'autre, il y a une unité « idiolectale ».

[91] Le calcul de l'écart réduit n'est donc pas pertinent pour rendre compte des phénomènes qui nous intéressent.

[92] Comme l'a fait remarquer E. Coseriu, même pour établir la permanence d'un objet linguistique, il faut se placer à deux moments distincts.

ces différents critères sont secondaires, et orthogonaux par rapport au paramètre principal, celui du temps.

3.5.2. Les outils de mesure utilisés.

Je ne suis pas statisticienne, ni spécialiste en textométrie : les méthodes que j'ai utilisées sont assez simples, dans leur application, et dans leur interprétation de base (reste ensuite la tâche du linguiste…). Après avoir envisagé la mise en œuvre, à l'issue des différentes analyses, d'une représentation multidimensionnelle (AFC ou ACP) de l'évolution de l'expression et de la position du sujet pronominal, il m'est apparu que cette démarche était encore prématurée au regard de mes capacités à appliquer ces méthodes et surtout à interpréter leurs résultats. Il m'a semblé plus sage de différer à une étude ultérieure une telle approche.

Les outils de mesure utilisés sont : les fréquences, le calcul du khi2 et l'indice de corrélation (avec des réserves que j'exposerai plus loin).

3.5.2.1. Les fréquences

Il a été établi :

a) des fréquences absolues pour :

- P1 et P3 respectivement exprimés en position postverbale
- P1 et P3 respectivement exprimés en position préverbale
- P1 et P3 respectivement non exprimés

b) des fréquences relatives pour :

- inversion de P1 + P3 par rapport à l'ensemble des P1 + P3 exprimés (et corollairement fréquence de l'antéposition)
- inversion de P1 par rapport à l'ensemble des P1 exprimés (et corollairement fréquence de l'antéposition)
- inversion de P3 par rapport à l'ensemble des P3 exprimés (et corollairement fréquence de l'antéposition)
- omission de P1 + P3 par rapport à l'ensemble des P1 + P3 exprimés ou non (et corollairement fréquence de l'expression)
- omission de P1 par rapport à l'ensemble des P1 exprimés ou non (et corollairement fréquence de l'expression)

- omission de P3 par rapport à l'ensemble des P3 exprimés ou non (et corollairement fréquence de l'expression).

3.5.2.2. Le calcul du khi2[93]

Le calcul du khi2 est utilisé pour évaluer le caractère plus ou moins « étonnant », et donc significatif, de la distribution des pronoms de 1ère et 3ème personnes au regard de leur position d'une part, et de leur expression (ou non) d'autre part, et donc la liaison plus ou moins forte entre les deux variables (position/expression et personne). Je rappelle que le test de khi2 « sert à apprécier en probabilité l'écart constaté entre une observation et un modèle théorique, quel que soit le nombre des variables » (Muller 1992 : 116)[94].

Le test du khi2, que je vais illustrer par deux exemples forgés à valeur exemplaire, comporte plusieurs étapes. La première consiste à dresser le tableau des effectifs réels (observés).

Considérons un texte fictif 1 dans lequel la répartition des personnes 1 et 3 en positions préverbales et postverbales serait la suivante :

Exemple 1

	Verbe-Sujet	Sujet-Verbe	total VS + SV
P1	22	47	69
P3	34	70	104
total P1 +P3	56	117	173

Tableau 5 : effectifs réels de l'exemple 1

A partir de ces effectifs réels, on crée un tableau des effectifs calculés (théoriques), en calculant pour les 4 cases ce que donnerait une répartition strictement homogène de P1 et P3 en position préverbale et postverbale. C'est ce qu'on appelle l'hypothèse nulle. Pour calculer ces effectifs théoriques, on répète pour chacune des 4 cellules une règle de 3 qui s'appuie sur les effectifs marginaux, c'est-à-dire que l'on divise le produit des 2 effectifs marginaux (ligne et colonne) par le total général. Par exemple, l'effectif théorique de P1 en position postverbale (effectif réel : 22) est égal à : 56*69/173, soit 22.34.

[93] Je remercie Bernard Victorri et Benoît Habert pour leur aide à la formulation, dans Excel, des différents calculs liés au khi2.

[94] La présentation ci-dessous est très largement inspirée de Muller (1992 : 116-121).

On peut dresser un second tableau avec les effectifs calculés, ou bien, pour plus de lisibilité, rassembler effectifs observés et effectifs calculés dans un même tableau. C'est l'option adoptée ici, effectifs observés et effectifs calculés étant indiqués dans une même cellule, les seconds en italiques.
Voici le tableau récapitulatif :

	Verbe-Sujet	Sujet-Verbe	total VS + SV
P1	22 *22.3*	47 46.7	69
P3	34 *33.7*	70 *70.3*	104
total P1 +P3	56	117	173

Tableau 6 : effectifs réels et théoriques de l'exemple 1

Il s'agit d'un tableau qui comporte '1 degré de liberté' (ddl) : cela signifie qu'en connaissant 1 seul nombre des 4 cellules, on peut, grâce aux effectifs marginaux, calculer ceux des trois autres cellules. D'une manière générale, avec k lignes et n colonnes, le ddl est de (n-1)*(k-1).
la reconnaissance du degré de liberté conditionne l'interprétation du khi2.

Le calcul du khi2 s'opère ainsi :
Pour chaque cellule, on calcule la différence entre effectif observé et effectif calculé, on l'élève au carré[95], et l'on divise chaque écart par l'effectif calculé (pour tenir compte de la taille des données). On répète l'opération pour chaque cellule, et on additionne les résultats :

$\chi^2 = \Sigma [(o-c)^2 / c]$

Dans l'exemple du texte 1, cela nous donne (colonne par colonne, de haut en bas) :
0.005 + 0.003 + 0.002 + 0.002 = 0.012
Le khi2 est donc de 0.012.

La probabilité que la valeur obtenue par le moyen de ce calcul puisse correspondre à une distribution aléatoire (si on admet l'hypothèse nulle) est fournie par une table de valeurs et de probabilités associées[96]. Elle dépend du degré de liberté du tableau.

[95] Pour éviter que les écarts négatifs et positifs ne s'annulent au moment où on en fera la somme.
[96] Voir (Muller 1992 :179). *Excel*, qui permet d'opérer le calcul du khi2, fournit des probabilités plus fines que les tables de Muller, en particulier pour les khi2 élevés. La table de Muller s'arrête en effet à une valeur de 10.827 pour le khi2 (ddl de 1), associé à une probabilité de 0.001. Au-delà de cette valeur

On admet généralement qu'une probabilité de 0.05 constitue un seuil. S'il apparaît que la répartition avait plus de 5% de chances d'être obtenue dans le cadre d'une répartition aléatoire, on considère que l'hypothèse nulle ne peut être écartée : la répartition peut être le fruit du hasard, elle n'est pas significative. A l'inverse, au-dessous de ce seuil, on admet que l'on peut rejeter l'hypothèse nulle (avec une force proportionnelle à la faiblesse de la probabilité) : la répartition observée n'a que peu de chances d'être le fruit du hasard, et elle mérite donc d'être étudiée de près.

Revenons-en à notre texte 1. Le khi2 est de 0.012. La valeur dont il se rapproche le plus dans la table d'interprétation du khi2 est 0.16, qui correspond à une probabilité de 0.9. Cela signifie que la distribution observée avait plus de 90% d'être obtenue dans le cadre d'une répartition aléatoire. La proximité, pour chaque cellule, des valeurs des effectifs observés et des effectifs calculés laissait d'ailleurs prévoir un tel résultat.
Face à un résultat aussi tranché, le linguiste peut se passer de chercher à comprendre la répartition des P1 et P3 selon leur position.

Considérons un autre exemple, forgé lui aussi. Le tableau ci-dessous donne les effectifs observés et les effectifs calculés.

Exemple 2

	Verbe-Sujet	Sujet-Verbe	total VS + SV
P1	55 *31.5*	14 *37.5*	69
P3	24 *47.5*	80 *56.5*	104
total P1 +P3	79	94	173

Tableau 7 : Effectifs réels et théoriques de l'exemple 2

Dans ce second exemple, le khi2 est égal à 53.6 (17.5 + 11.6 + 14.7 + 9.8).
La probabilité qui lui est associée est de $2{,}4343*10^{-13}$: elle est très largement inférieure à 0.01% : la liaison entre les variables 'position' et 'personne' mérite d'être commentée.
Dès lors commence l'interprétation des chiffres[97].
Il s'agit de préciser la nature de la liaison entre les deux variables, et il faut pour cela mesurer la contribution relative de chacune des cases au khi2 : on peut ainsi repérer

de khi2, on peut simplement dire que la probabilité est inférieure à 0.1%. Cela dit, passé ce seuil, il n'est pas certain que l'on peut interpréter les écarts de probabilité.
[97] Voir Badia, Bastida et Haït (1997 : chap.5) pour une présentation claire de la notion de contribution au khi2.

celles qui contribuent le plus à la liaison, que l'on peut dès lors caractériser par l'association des modalités des deux variables. Mais une même contribution peut avoir des interprétations différentes selon que l'écart entre effectif réel et effectif calculé est positif ou négatif. S'il est positif, et donc que l'effectif réel est supérieur à l'effectif théorique, on dira qu'il y a attraction entre les deux variables ; si au contraire il est négatif (effectif réel inférieur à l'effectif calculé), on dira qu'il y a répulsion.

Voici, pour l'exemple 2, la contribution absolue (en normal) et relative (en italiques) de chacune des cases à la liaison entre les variables exprimée par le khi2 (53.6). Est aussi mentionné le signe de la différence entre effectifs réels et effectifs théoriques. Enfin, on a surligné les cases du tableau qui contribuent le plus au khi2.

Par la suite, on n'indiquera plus la contribution absolue des différentes cases au khi2.

	Verbe-Sujet		Sujet-Verbe	
P1	17.5	*32.7*	14.7	*27.4*
	+		-	
P3	11.6	*21. 7*	9.8	*18.2*
	-		+	

Tableau 8 : Contribution au khi2 (53.6) des effectifs de l'exemple 2

Il ressort de ces données qu'il y a une attraction forte entre P1 et la position postverbale (contribution de 32.7% au khi2), et une répulsion assez marquée entre cette même personne et la position préverbale (contribution de 27.4% au khi2). La liaison s'opère donc principalement entre P1 et la position.

Il est intéressant de mettre ces éléments en relation avec les pourcentages correspondants. Ainsi, l'inversion de P1 + P3 est de 45.7% sur l'ensemble des sujets P1 + P3 exprimés. L'inversion de P1 est de 79.7% sur l'ensemble des P1 exprimés, et l'inversion de P3 est de 23.7% sur l'ensemble des P3 exprimés. Sans surprise, on retrouve les mêmes tendances que celles que le calcul du khi2 a mises au jour, mais ce dernier a l'avantage de quantifier précisément, et de qualifier, le caractère « étonnant » de la distribution.

3.5.2.3. Le coefficient de corrélation

Cet indice a été utilisé pour déterminer s'il existe un lien entre l'expression et la position du sujet pronominal, plus spécifiquement entre l'inversion et la non-expression, cela en tenant compte des personnes.

C'est plus précisément le 'coefficient de corrélation linéaire', noté *r*, qui a été utilisé. La présentation que j'en fais s'appuie très largement sur Champely (1994 : 127-129) et Muller (1992 :157-162).

Ce coefficient définit l'intensité et la direction d'une relation linéaire entre les données de deux variables quantitatives. Il est toujours compris entre 1 et -1. Son signe indique le sens de la relation, c'est-à-dire une attraction entre les deux classements quand le signe est positif, ou au contraire une répulsion quand il est négatif. La valeur absolue exprime quant à elle l'intensité de la liaison ou de la divergence : plus elle se rapproche de 1 ou de -1, plus la relation linéaire est forte, il y a une dépendance entre les deux variables ; plus elle s'en éloigne, plus les deux variables sont indépendantes. Proche de 0, le coefficient exprime l'absence de relation linéaire.

Il est rare que l'on calcule manuellement le coefficient de corrélation (c'est long et fastidieux), différents outils, dont *Excel*, permettant de le faire automatiquement. Il est néanmoins utile de rappeler en quoi consiste ce calcul.

Le coefficient de corrélation ne se calcule pas à partir des effectifs réels, mais à partir de données 'normalisées', qui correspondent à l' 'écart réduit' (*Z*). D'une manière générale, c'est-à-dire indépendamment du calcul de la corrélation, l'écart réduit consiste, pour une série de données quantitatives, en la mesure de la différence entre chacune des données et la moyenne, différence que l'on rapporte à l''écart type'. L'écart type couvre la dispersion autour de la moyenne, et il correspond à la racine carrée de la 'variance'.

Prenons la série de chiffres suivants : 5 ; 7 ; 9 ; 3 ; 21. La moyenne est de 9

Les écarts observés (=déviations) par rapport à cette moyenne sont respectivement : - 4 ; -2 ; 0 ; -6 ; +12.

La variance est la moyenne des carrés des déviations par rapport la moyenne[98].

Pour la série de chiffres ci-dessus, la variance est égale à :
(16 + 4 + 0 + 36 + 144) / 5 = 40

L'écart type correspond à la racine carrée de 40 : 6.3

L'écart réduit rapporte les écarts observés à cet écart type :

- 4 / 6.3 = - 0.63

[98] Elever les déviations au carré permet d'éviter les effectifs négatifs.

- 2 / 6.3 = - 0.32

etc.

Revenons-en au calcul du coefficient de corrélation, qui s'appuie donc sur les données normalisées que sont les écarts réduits. Reprenons la série de données, et considérons-la comme le nombre de mots des 5 phrases d'un texte. A cette série associons une autre série, qui correspond au nombre de lettres du premier mot de chacune des 5 phrases, le but étant de voir s'il y a une *corrélation* entre le nombre de mots des phrases et le nombre de lettres de leur premier mot.

Le calcul du coefficient de corrélation correspond à la moyenne du produit des écarts réduits.

Voici ci-dessous les effectifs réels, les données normalisées (Z) et le produit de ces dernières :

	nombre de mots (x)	Z (x)	nombre de lettre du 1er mot (y)	Z (y)	Z(x) * Z(y)
phrase 1	5	- 0.63	4	- 1.42	0.89
phrase 2	7	- 0.32	5	- 0.7	0.22
phrase 3	9	0	8	1.42	0
phrase 4	3	- 0.95	6	0	0
phrase 5	21	1.9	7	0.7	1.33

La moyenne du produit des écarts réduits est : (0.89 + 0.22 + 0 + 0 + 1.33) / 5 = 0.49

Le coefficient de corrélation est de 0.49. L'interpétation de ce coefficient se fait à l'aide d'une table qui tient compte du nombre d'unités mesurées (Muller 1992 : 180). Dans l'exemple ci-dessus cas présent, nous avons 5 unités, et donc 3 degrés de liberté (le nombre de degrés de libertés se calcule ici par soustraction de 2 au nombre d'unités). Il apparaît que la probabilité d'atteindre ou de dépasser, par le seul jeu du hasard, un coefficient de 0.49 est très largement supérieure à 0.1 (soit 10%) : l'hypothèse nulle ne peut être rejetée, il n'y a pas de dépendance entre les deux variables, elles ne sont pas corrélées (ce qui n'est pas très surprenant au regard du choix des variables).

Chapitre 4. Analyse quantitative des données

Le khi2 a été calculé pour évaluer la liaison plus ou moins forte entre la position ou l'expression du sujet pronominal et sa personne dans chacun des textes. Il s'avère que, dans la majorité d'entre eux, la probabilité que la distribution de la position puisse correspondre à une répartition aléatoire est supérieure à 0.05, qu'elle a donc plus de 5% de chances d'être obtenue dans le cadre d'une répartition aléatoire, et qu'elle n'est donc pas significative. Les résultats sont bien différents pour ce qui concerne l'expression de Sp, la répartition de celle-ci entre P1 et P3 correspondant très majoritairement à une très faible probabilité de distribution aléatoire. Afin de ne pas alourdir les données chiffrées, seules sont présentées les distributions qui sont significatives. Elles sont exposées en 4.3.2. et 4.4.2. dans le cadre des synthèses sur la position et sur l'expression de Sp.

4.1. Fréquences de la position et de l'expression du sujet dans chaque texte

Voici dans un premier temps les données propres à chaque texte, tant en ce qui concerne l'expression que la position du sujet (une synthèse des fréquences de l'inversion et de la non-expression de Sp est présentée en 4.3. et 4.4., mais il m'a semblé intéressant de grouper pour chacun des textes l'ensemble des données).
Pour chaque texte sont aussi rappelés le nombre de mots, la date, la forme, le domaine et le genre, et le dialecte quand il est connu, ainsi que la fiabilité de l'étiquetage.
J'indique en outre, en note, la part du bruit parmi les occurrences collectées pour chacune des requêtes. Quelques remarques à ce sujet sont faites à l'issue de la présentation des différents textes, en 4.2.
Les textes sont présentés par ordre chronologique, sachant que pour certains la date n'est pas facile à établir précisément (voir le tableau 1 en 3.1.2. pour une vision synthétique des dates des textes).

Remarque : j'ai décidé de ne pas inclure les incises (*dit-il, fait-il*) dans les différents calculs, dans la mesure où il s'agit, dans tous les textes du corpus, de structures relativement figées. J'indique néanmoins en note leur fréquence, d'ailleurs très variable selon les textes. Signalons que toutes les incises impliquent la troisième

personne, singulière ou plurielle : on ne rencontre pas d'incise avec la première personne ; on rencontre en revanche de rares cas de « ce di », constructions qui, elles, ont été prises en compte dans les calculs.

4.1.1. *LA CHANSON DE ROLAND*

Nombre de mots	29 338
Date	*ca* 1100
Forme	vers
Domaine	littéraire
Genre	épique
Dialecte	anglo-normand
Texte étiqueté avec vérification	

Tableau 9.1. Caractéristiques de *Roland*

4.1.1.1. Position du sujet[99]

Les fréquences absolues sont en 'normal', les fréquences relatives (exprimées en pourcentage) sont en italiques, et celles des séquences verbe-sujet sont en outre grassées.

Les fréquences relatives sont celles de la position préverbale ou postverbale de P1 et/ou P3 rapportées à l'ensemble des sujets exprimés de la ou des personne(s) correspondante(s) (total SpV+VSp).

J'indique en outre, entre parenthèses, le pourcentage de P1 sur l'ensemble des sujets (P1 et P3) préverbaux d'une part, postverbaux d'autre part. L'établissement de ce pourcentage a pour but de déceler une éventuelle relation entre cette proportion et celle de l'inversion de P1.

	Sp préverbal	Sp postverbal	Sp pré.+ post.
P1	83 *(58.4)* *82.2*	18 *(47.4)* ***17.8***	101
P3	59 *74.7*	20 ***25.3***	79
Total	142 *78.9*	38 ***21.1***	180

Tableau 9.2. Fréquences absolues et relatives de la position de P1 et P3 dans *Roland*

[99] Proportion de bruit : 2.3% pour les P1 préverbaux (c'est-à-dire que 2.3% des occurrences collectées à l'aide de la requête élaborée pour recueillir les constructions avec P1 préverbal ne sont pas pertinentes), 6.3% pour les P3 préverbaux, 5.3% pour les P1 postverbaux et 15.4% pour les P3 postverbaux. Les incises (2 occurrences), non comptabilisées dans les calculs mais non considérées comme du bruit, représentent 9% des P3 postverbaux.

***Exemples*[100] :**

(26) Puis od les ewes lavat les prez del sanc ;
Pur cel le fist ne fust aparissant.
Pur un sul levre vait tute jur cornant.
Devant ses pers **vait il** ore gabant.
(*Roland*, v. 1781)

(27) « Sire cumpain, faites le vos de gred ?
Ja est ço Rollant, ki tant vos soelt amer !
Par nule guise ne m'avez desfïet ! »
Dist Oliver : « Or vos **oi jo** parler
Jo ne vos vei, veied vus Damnedeu ! ...»
(*Roland*, v. 2003)

4.1.1.2. Expression du sujet[101] :

Les pourcentages (en italiques) indiquent la fréquence de l'expression ou de la non-expression de P1 et/ou P3 sur l'ensemble des sujets, exprimés ou non, de la/des personne(s) correspondante(s). Les pourcentages de sujets non-exprimés sont en outre grassés.

J'indique en outre, entre parentèses, le pourcentage de P1 sur l'ensemble des sujets (P1 et P3) exprimés d'une part, non exprimés d'autre part. L'établissement de ce pourcentage a pour but de déceler une éventuelle relation entre cette proportion et celle de la non-expression de P1.

	Sp exprimés	Sp non exprimés	Total Sp
P1	101 *(56.1)* *37.8*	166 *(11)* ***62.2***	267
P3	79 *5.6*	1341 ***94.4***	1420
Total	180 *10.7*	1507 ***89.3***	1687

Tableau 9.3. Fréquences absolues et relatives de l'expression de P1 et P3 dans *Roland*

***Exemples* :**

(28) Rollant saisit e sun cors e ses armes
E dist un mot : « Vencut est li niés Carles !
Iceste espee **porterai** en Arabe. »
En cel tireres li quens s'aperçut alques.
(*Roland*, v. 2282)

[100] Je donne des exemples des deux phénomènes qui se sont raréfiés : inversion et non-expression.
[101] Proportion de bruit à l'issue de la collecte des sujets non exprimés : 61.2%.

(29) « - E ! Deus ! dist Carles, ja sunt il ja si luinz !
Cunsentez mei e dreiture e honur ;
De France dulce **m'unt tolue** la flur. »
Li reis cumandet Gebuin e Otun, …
(*Roland*, v. 2431)

4.1.2. *ENEAS*

Nombre de mots	34 958
Date	*ca.* 1155
Forme	vers
Domaine	littéraire
Genre	roman
Dialecte	normand
Texte étiqueté sans vérification	

Tableau 10.1. Caractéristiques d'*Eneas*

4.1.2.1. Position du sujet[102]

	Sp préverbal	Sp postverbal	Sp pré.+ post.
P1	83 *(26.3)* *87.4*	12 *(35.3)* ***12.6***	95
P3	232 *91.3*	22 ***8.7***	254
Total	315 *92.3*	34 ***9.7***	349

Tableau 10.2. Fréquences absolues et relatives de la position de P1 et P3 dans *Eneas*

Exemples :

(30) l'en nos chace de tote terre,
au roi venons por consoil querre
et mostrer li nostre besoing ;
de mesfere **ne ai ge** soing.
Consoilliez nos, por deu, biaus sire (*Eneas*, 4680)

(31) ja n'i cuident a tens venir,
que li Troïen ne s'en fuient,
mais molt sont fol quant il lo cuident :
issi ne s'an iront il pas.
Bien **avoit oï** Eneas que
Turnus asanblot sa gent... (*Eneas*, v. 4243)

[102] Bruit : 9.8 % pour les P1 préverbaux, 13.4% pour les P3 préverbaux, 40% pour les P1 postverbaux et 4.9% pour les P3 postverbaux. Le nombre d'incises est élevé : 37 occurrences sur 59 P3 postverbaux, soit 62.7% de ces derniers.

4.1.2.2. Expression du sujet[103]

Remarque : pour ce texte, comme pour tous ceux dont l'étiquetage n'a pas été vérifié et n'est donc pas pleinement fiable, il convient de considérer avec précaution les chiffres donnés pour les sujets non exprimés.

	Sp exprimés	Sp non exprimés	Total Sp
P1	95 *(27.2)* *28.3*	241 *(14.3)* ***71.7***	336
P3	254 *15*	1446 ***85***	1700
Total	349 *17.2*	1687 ***82.8***	2036

Tableau 10.3. Fréquences absolues et relatives de l'expression de P1 et P3 dans *Eneas*

Exemples :

(32) voiz les, ges te nomerai toz,
mostrerai toi com il vandront,
tot an ordre, com il naistront,
en aprés les te **nomerai**
et les batailles te **dirai**,...
(*Eneas*, v. 2925-2929)

(33) de lui **firent** segnor et mestre.
Puis **ont gardé** devers senestre :
une estoyle **virent** levee
qui la voie lor a mostree...
(*Eneas*, v. 76-79)

4.1.3. *TRISTAN* DE BEROUL

Nombre de mots	27 257
Date	entre 1165 et 1200
Forme	vers
Domaine	littéraire
Genre	roman
Dialecte	franco-picard
Texte étiqueté sans vérification + annoté syntaxiquement	

Tableau 11.1. Caractéristiques de *Beroul*

[103] Bruit pour la collecte des sujets non exprimés : 61.6%.

4.1.3.1. Position du sujet[104]

	Sp-Verbe	Verbe-Sp	SpV + VSp
P1	125 *(56.6)* *83.9*	24 *(51)* ***16.1***	149
P3	96 *80.7*	23 ***19.3***	119
Total	221 *82.5*	47 ***17.5***	268

Tableau 11.2. Fréquences absolues et relatives de la position de P1 et P3 dans *Beroul*

Exemples :

(34) Puis dist itant : « Si je pooie
Husdent par paine metre en voie
Que il laisast cri por silence,
Mot l'avroie a grant reverence.
Et a ce **metrai je** ma paine
Ainz que ja past ceste semaine ;
(*Beroul*, v. 1593-1598)

(35) Amis Tristran, en grant error
Nos mist qui le boivre d'amor
Nos aporta ensenble a boivre,
Mex ne nos **pout il** pas deçoivre.
(*Beroul*, v. 2217-2220)

4.1.3.2. Expression du sujet[105]

	Sp exprimés	Sp non exprimés	Total Sp
P1	149 *(55.6)* *38.3*	240 *(18.8)* ***61.7***	389
P3	119 *10.3*	1037 ***89.7***	1156
Total	268 *17.3*	1277 ***82.7***	1545

Tableau 11.3. Fréquences absolues et relatives de l'expression de P1 et P3 dans *Beroul*

[104] Bruit : 10.7% pour les P1 préverbaux, 18.6% pour les P3 préverbaux, 17.2% pour les P1 postverbaux, 20.8% pour les P3 postverbaux. Les incises (34 occurrences) représentent 59.6% des P3 postverbaux.

[105] Bruit pour la collecte des sujets non exprimés : 14.9%. Rappelons que, pour ce texte, les sujets non exprimés ont été collectés avec *TIGERSearch*, à l'aide d'une requête permettant de repérer tous les sujets non-exprimés et uniquement les sujets non-exprimés (voir 3.4.4. ci-dessus). Rappelons aussi, à titre de comparaison, que la collecte des « sujets non exprimés » dans *TXM* s'élève à 3482 occurrences. Sachant que seules 1277 occurrences sont finalement pertinentes, le bruit s'élève donc au moins à 63%.

Exemples :

(36) Tristran, gardez en nule place
Ne me mandez por nule chose :
Je ne seroie pas tant ose
Que je i osasse venir
Trop **demor** ci, n'en **quier** mentir.
(*Beroul*, v. 60-64)

(37) Li rois qui sus en l'arbre estoit
Out l'asenblee bien veüe
Et la raison tote entendue.
De la pitié q'au cor li prist,
Qu'il ne plorast ne s'en tenist
Por nul avoir ; mout **a** grant duel,
Mot **het** le nain de Tintaguel.
(*Beroul*, v.258-264)

4.1.4. *AMI ET AMILE*

Nombre de mots	29 338
Date	*ca.* 1200
Forme	vers
Domaine	littéraire
Genre	épique
Dialecte	non défini
Texte étiqueté sans vérification	

Tableau 12.1. Caractéristiques de *Amile*

4.1.4.1. Position du sujet[106]

	Sp-Verbe	Verbe-Sp	SpV + VSp
P1	83 *(47.1)* *74.1*	29 *(52.7)* ***25.9***	112
P3	93 *78.2*	26 ***21.8***	119
Total	176 *76.2*	55 ***23.8***	231

Tableau 12.2. Fréquences absolues et relatives de la position de P1 et P3 dans *Amile*

Exemples :

(38) Or le voz voil bel et gent presenter,
Mais une chose **voz di je** par verté :

[106] Bruit : 1.2 % pour les P1 préverbaux, 11.4% pour les P3 préverbaux, 3.3% pour les P1 postverbaux et 4.6% pour les P3 postverbaux. Le nombre d'incises est très élevé : 78 occurrences sur 104 P3 postverbaux, soit 75 % de ces derniers.

Tant com je poi traïr et encuser,
Si m'ama Charles et si fui ses privéz.
(*Amile*, v.1614-1617)

(39) En terre fiche son roit espié forbi,
L'auberc ne l'iaume **n'a il pas deguerpi**,
Son bon escu avoit a son chief mis,
Car moult redoute Hardré, son annemi,
(*Amile*, v. 924-927)

4.1.4.2. Expression du sujet[107] :

	Sp exprimés	Sp non exprimés	Total Sp
P1	112 *(47.6)* *35.9*	200 *(19.7)* ***64.1***	312
P3	119 *12.7*	816 ***87.3***	935
Total	231 *18.5*	1016 ***81.5***	1247

Tableau 12.3. Fréquences absolues et relatives de l'expression de P1 et P3 dans *Amile*

***Exemples* :**

(40) Vostre proesce qu'est elle devenue ?
Tant soloit iestre et doutee et cremue.
Se tu le vaiz, touz jors **serai** ta drue.
Onques ne **fu** a tort si mescreüe
(*Amile*, v. 1527-1530)

(41) Amis le voit moult en **est** esperduz.
Or **se demente** et **dist** : « Las ! tant mar fuz
Que tu venis en terre. »
(*Amile*, v. 3040-3042)

4.1.5. ROBERT DE CLARI, *LA CONQUESTE DE CONSTANTINOPLE*

Nombre de mots	38 188
Date	après 1205
Forme	prose
Domaine	historique
Genre	chronique
Dialecte	picard
Texte étiqueté sans vérification	

Tableau 13.1. Caractéristiques de *Clari*

[107] Bruit pour la collecte des sujets non exprimés: 55.4%.

4.1.5.1. Position du sujet[108]

	Sp-Verbe	Verbe-Sp	SpV + VSp
P1	36 *(29)* *100*	0 ***0***	36
P3	88 *49.9*	92 ***51.1***	180
Total	124 *57.4*	92 ***42.6***	216

Tableau 13.2. Fréquences absolues et relatives de la position de P1 et P3 dans *Clari*

Exemple :

(42) Li vaslés fu molt esmaris, quan il oï ches nouveles, et tant qu'il vint avant, qu'il ne se peut estordre a nul fuer qu'il n'alast hors a chu balliu. Si ne **fait il** mais el, si **prent il** s'espee, si **le met il** sous sen surcot, **se s'en ist il** hors de le maison, **si vient il** devant le balliu, (Clari, p. 22)

C'est le seul texte du corpus dans lequel on observe une telle succession d'inversions du sujet pronominal (voir plus bas les exemples dans la *Queste* et *Quinze Joyes* avec 2 occurrences successives). C'est un phénomène globalement rare dans l'ensemble des textes d'ancien français. On notera que ce texte présente par ailleurs une absence totale d'inversion de P1 conjuguée à un pourcentage très élevé d'inversion de P3, qui dépasse largement celui de tous les autres textes. A différents égards, c'est donc un texte atypique, comparé aux autres textes du corpus.

4.1.5.2. Expression du sujet[109]

	Sp exprimés	Sp non exprimés	Total Sp
P1	36 *(17)* *75*	12 *(1.3)* ***25***	48
P3	180 *16.7*	900 ***83.3***	1080
Total	216 *19.2*	912 ***80.8***	1128

Tableau 13.3. Fréquences absolues et relatives de l'expression de P1 et P3 dans *Clari*

Exemples :

(43) si leur dist : « Seignor, cheste vile a molt meffait a mi et a me gent ; je m'en vengeroie volentiers. Si **pri** que vous me soiés en aiue. »

[108] Bruit : 0% pour les P1 préverbaux, 39.3% pour les P3 préverbaux, 23.2% les P3 postverbaux. On relève 14 incises, qui représentent 13.2% des 106 P3 postverbaux.
[109] Bruit pour la collecte des sujets non exprimés : 74.2%.

(*Clari*, p. 14)

(44) Quant li marchis vit que le kiertés fu si grans en le vile, et qu'il ne pooient avoir soulas ne confort de nule part, si **manda** tous chiaus de le vile et Genevois qu'il i avoit et uns et autres, si parla a aus et si leur dist... (*Clari*, p. 9)

4.1.6. *AUCASSIN ET NICOLETE*

Nombre de mots	11 679
Date	1ère moitié du 13ème
Forme	mixte
Domaine	littéraire
Genre	récits brefs
Dialecte	picard
Texte étiqueté sans vérification	

Tableau 14.1. Caractéristiques de *Aucassin*

4.1.6.1. Position du sujet[110]

	Sp-Verbe	Verbe-Sp	SpV + VSp
P1	54 *(34.8)* *83.1*	11 *(64.7)* ***16.9***	65
P3	101 *94.4*	6 ***5.6***	107
Total	155 *90.1*	17 ***9.9***	172

Tableau 14.2. Fréquences absolues et relatives de la position de P1 et P3 dans *Aucassin*

Exemples :

(45) « E ! Dix, fait Aucasins, ci fu Nicolete me douce amie, et ce fist ele a ses beles mains ; por le douçour de li et por s'amor me **descenderai je** ore ci et m'i reposerai anuit mais. » (*Aucassin*, p. 27)

(46) La nuit le laissent ensi,
tresqu'au demain par matin
que l'espousa Aucassins :
dame de Biaucaire en fist ;
puis **vesquirent il** mains dis
et menerent lor delis.
(*Aucassin*, p. 39)

[110] Bruit : 12.9% pour les P1 préverbaux, 5.6% pour les P3 préverbaux, 0% pour les P1 postverbaux, 3.2% les P3 postverbaux. La proportion d'incises est élevée : 24 occurrences, qui représentent 80% des 30 P3 postverbaux.

4.1.6.2. Expression du sujet[111]

	Sp exprimés	Sp non exprimés	Total Sp
P1	65 *(37.8)* *62.5*	39 *(9.8)* ***37.5***	104
P3	107 *23*	357 ***77***	464
Total	172 *30.3*	396 ***69.7***	568

Tableau 14.3. Fréquences absolues et relatives de l'expression de P1 et P3 dans *Aucassin*

Exemples :

(47) « - Sire, les deniers prenderons nos, mais ce ne vos **canterai mie**, car j'en ai juré ; mais je le vos conterai, se vos volés ... (*Aucassin*, p. 23).

(48) et il getent les mains de toutes pars, si le **prendent**, si le **dessaisisent** de l'escu et de le lance, si l'en **mannent** tot estrousement pris, et **aloient** ja porparlant de quel mort il feroient morir. (*Aucassin*, p. 10)

4.1.7. *MIRACLES* DE GAUTIER DE COINCI

Nombre de mots	22 831 (texte échantillonné)
Date	*ca.* 1218-1227
Forme	vers
Domaine	religieux
Genre	lyrique
Dialecte	traits picards
Texte étiqueté sans vérification	

Tableau 15.1. Caractéristiques de *Miracles*

4.1.7.1. Position du sujet[112]

	Sp-Verbe	Verbe-Sp	SpV + VSp
P1	37 *(37)* *80.4*	9 *(32.1)* ***19.6***	46
P3	63 *76.8*	19 ***23.2***	82
Total	100 *78.1*	28 ***21.9***	128

Tableau 15.2. Fréquences absolues et relatives de la position de P1 et P3 dans *Miracles*

Exemples :

(49) Laienz avec ces vielz provoires
Ne **veil je** plus laissier m'amie.

[111] Bruit pour la collecte des sujets non exprimés : 67.3%.
[112] Bruit : 14% pour les P1 préverbaux, 18.2% pour les P3 préverbaux, 0% pour les P1 postverbaux, 17.6% pour les P3 postverbaux. On relève 9 incises, soit 32.1% des 28 P3 postverbaux.

Demain au soir n'i sera mie :
(Miracles, v. 1961-1962)

(50) Plus a en aus borre que lainne
Venin et fiel que miel ne çucre.
Adez **quierent il** ou sepuchre
Nostre Seigneur, ce m'est avis.
(*Miracles*, v. 1194-1197)

Signalons l'exemple ci-dessous de disjonction entre le verbe et son sujet postverbal. C'est un hapax dans notre corpus, et un fait rarissime d'une manière générale (voir remarques en 1.1 et note 10). On peut y voir une licence poétique.

(51) Tex semences ont tost semees.
El feu d'enfer soient semé
Tuit mesdisant, tuit seursemé.
Por ce me tieng en petit cloistre
Que leur semence n'i puet croistre.
Fors de cloistre est ma damoisele ;
N'i *rentera* mais des mois *ele*.
Diex saut les moignes et l'abbé,
(*Miracles*, v.2124-2131)

[traduction : elle n'y rentrera sinon avant des mois]

4.1.7.2. Expression du sujet[113]

	Sp exprimés	Sp non exprimés	Total Sp
P1	46 *(36)* *29.5*	110 *(16.3)* ***70.5***	156
P3	82 *12.7*	563 ***87.3***	645
Total	128 *16*	673 ***84***	801

Tableau 15.3. Fréquences absolues et relatives de l'expression de P1 et P3 dans *Miracles*

Exemples :

(52) Qui sa provende ne desert
Diex est a lui et se le sert :
Diex est ses clers et ses vicaires.
Se Diex m'aït, n'en **voi** mais gaires
Qui les deservent bien a droit.
(*Miracles*, v. 967-971)

(53) Mais ele vit soudainnement
Une clarté deseur son lit

[113] Bruit pour la collecte des sujets non exprimés : 66.5%.

Si grant que nus si grant ne vit
Et **vit** venir une roÿne
Plus luisant, plus clere et plus fine
N'est Lucifer quant l'aube crieve.
(*Miracles*, v. 36-41)

4.1.8 *LA QUESTE DEL SAINT GRAAL*

Nombre de mots	45 000 (texte échantillonné)
Date	*ca.*1225-1230
Forme	prose
Domaine	littéraire
Genre	roman
Dialecte	non défini
Texte étiqueté avec vérification	

Tableau 16.1. Caractéristiques de *Queste*

4.1.8.1. Position du sujet[114]

	Sp-Verbe	Verbe-Sp	SpV + VSp
P1	129 *(30.8)* *66.9*	64 *(53.8)* ***33.1***	193
P3	290 *84.1*	55 ***15.9***	345
Total	419 *77.9*	119 ***22.1***	538

Tableau 16.2. Fréquences absolues et relatives de la position de P1 et P3 dans *Queste*

Exemples :

(54) - Sire, fet Lancelot, qui fu cil qui tant a parlé a vos ? Son cors ne **poi je** veoir, mes sa parole **oï je** bien qui est si laide et si espoantable qu'il n'est nus qui grant poor n'en deust avoir. (*Queste*, p. 122)

(55) et il ala parmi le palais tot entor les dois d'une part et d'autre, et tout einsi com il trespassoit par devant les tables **estoient eles** maintenant **raemplies** endroit chascun siege de tel viande come chascuns desirroit. (*Queste*, p. 15)

[114] Bruit : 11.6% pour les P1 préverbaux, 27.1% pour les P3 préverbaux, 13.5% pour les P1 postverbaux, 21% les P3 postverbaux. On relève 80 incises, qui représentent 59.2% des 135 P3 postverbaux.

4.1.8.2. Expression du sujet[115]

	Sp exprimés	Sp non exprimés	Total Sp
P1	193 *(35.9)* *78.5*	53 *(4.5)* ***21.5***	246
P3	345 *23.5*	1125 ***76.5***	1470
Total	538 *31.4*	1177 ***68.6***	1715

Tableau 16.3. Fréquences absolues et relatives de l'expression de P1 et P3 dans *Queste*

Exemples :

(56) Endementres qu'il parloient einsi si entra laienz uns vaslez qui dist au roi : « Sire noveles vos **aport** mout merveilleuses (*Queste*, p. 5)

(57) Et quant li rois vit qu'il avoient fait tel veu si en fu mout a malaise. Car bien **set** qu'il nes porra pas retorner de ceste emprise (*Queste*, p. 16)

4.1.9. *COUTUMES DE BEAUVAISIS de* PHILIPPE DE BEAUMAMOIR

Nombre de mots	22 500 (texte échantillonné)
Date	vers 1283
Forme	prose
Domaine	juridique
Genre	traité
Dialecte	traits picards
Texte étiqueté sans vérification	

Tableau 17.1. Caractéristiques de *Coustumes*

4.1.9.1. Position du sujet[116]

	Sp-Verbe	Verbe-Sp	SpV + VSp
P1	18 *(12.5)* *81.8*	4 *(14.8)* ***18.2***	22
P3	126 *84.5*	23 ***15.4***	149
Total	144 *84.2*	27 ***15.8***	171

Tableau 17.2. Fréquences absolues et relatives de la position de P1 et P3 dans *Coustumes*

Exemples :

[115] Bruit pour la collecte des sujets non exprimés : 69.2%.

[116] Bruit : 48.8% pour les P1 préverbaux, 43.7% pour les P3 préverbaux, 0% pour les P1 postverbaux, 51% les P3 postverbaux. Il n'y a pas d'incises.

(58) Mes se j'en oste plus du tiers, l'homages du tiers et du seurplus vient au seigneur. Et en tel maniere le **pourroie je** fere que je pourroie plus perdre, si comme se je retenoie les homages du plus du tiers (*Coustumes*, p. 237).

(59) S'il connoist les bons des mauvès, il pourra et devra les mauvès sarcler et essarter des bons, a l'essample que l'en oste les mauveses herbes des fourmens ; et a ce fere **est il tenus** (*Coustumes*, p. 23).

4.1.9.2. Expression du sujet[117]

	Sp exprimés	Sp non exprimés	Total Sp
P1	22 *(12.8)* *73.3*	8 *(12.7)* ***26.7***	30
P3	149 *73*	55 ***27***	204
Total	171 *73.1*	63 ***26.9***	234

Tableau 17.3. Fréquences absolues et relatives de l'expression de P1 et P3 dans *Coustumes*

Exemples :

(60) et se j'en donnai aucune chose en ce point pour la sauveté de mon cors ou pour le mien sauver, redemander le **puis** et le **doi** ravoir, car il apert que je le fis par paour. (*Coustumes*, p. 504)

(61) Et pour la franchise et l'aisement li oste de ses hommes i venoient sans fere envers leur seigneurs ce qu'il devoient de leurs masures, ainçois les **lessoient** gastes. (*Coustumes*, p. 491-492)

4.1.10. *MEMOIRES OU VIE DE SAINT LOUIS* DE JOINVILLE

Nombre de mots	45 000 (texte échantillonné)
Date	entre 1305 et 1309
Forme	prose
Domaine	historique
Genre	mémoires
Dialecte	non marqué
Texte étiqueté sans vérification	

Tableau 18.1. Caractéristiques de *Joinville*

[117] Le bruit est encore plus élevé en ce qui concerne la collecte des sujets non exprimés : 96.1%. Ces scores médiocres tiennent principalement à des erreurs d'étiquetage, qu'on peut expliquer par le lexique de ce texte, *a priori* assez différent des textes plus littéraires sur lesquels le moteur d'étiquetage avait été entraîné.

4.1.10.1. Position du sujet[118]

	Sp-Verbe	Verbe-Sp	SpV + VSp
P1	226 *(41.6)* *85*	40 *(52.6)* ***15***	266
P3	317 *89.8*	36 ***10.2***	353
Total	543 *87.7*	76 ***12.3***	619

Tableau 18.2. Fréquences absolues et relatives de la position de P1 et P3 dans *Joinville*

Exemples :

(62) Lors dit le roy : « Seigneurs, j'ai oÿ vostre avis et l'avis de ma gent. Or vous **redirai je** le mien, qui est tel que se je descent de la nef, que il a ceans tiex. (*Joinville*, p. 312)

(63) Dieu, en qui il mist sa fiance, le gardoit touz jours des s'enfance jusques a la fin ; et especialment en s'enfance **le garda il** la ou il luy fu bien mestier, si comme vous orrez ci aprés... (*Joinville*, p. 36)

4.1.10.2. Expression du sujet[119]

	Sp exprimés	Sp non exprimés	Total Sp
P1	266 *(43)* *78.9*	71 *(14.5)* ***21.1***	337
P3	353 *45.9*	417 ***54.1***	770
Total	619 *55.9*	488 ***44.1***	1107

Tableau 18.3. Fréquences absolues et relatives de l'expression de P1 et P3 dans *Joinville*

Exemples :

(64) et les hales sont faites a la guise des cloistres de ces moinnes blans, mes je croi que de trop il n'en soit nul si grant. Et vous **dirai** pour quoy il le me semble,... (*Joinville*, p. 48)

(65) En dit que ce preudomme qui ce enseignoit le roy gist a Marseille, la ou Nostre Seigneur fait pour li maint bel miracle. Et ne **voult** onques demourer avec le roy, pour priere que il li sceut faire, que une seule journee (*Joinville*, p. 28)

[118] Bruit : 7.7% pour les P1 préverbaux, 24% pour les P3 préverbaux, 11% pour les P1 postverbaux, 21.6% pour les P3 postverbaux. Il y a 40 incises, qui représentent 52.6% des 76 P3 postverbaux.
[119] Le bruit est de 65.9% pour la collecte des sujets non exprimés.

4.1.11. *CHRONIQUES* DE FROISSART

Nombre de mots	45 000 (texte échantillonné)
Date	entre 1369 et 1405
Forme	prose
Domaine	historique
Genre	chronique
Dialecte	picard
Texte étiqueté sans vérification	

Tableau 19.1. Caractéristiques de *Froissart*

4.1.11.1. Position du sujet[120]

	Sp-Verbe	Verbe-Sp	SpV + VSp
P1	87 *(23.4)* *94.6*	5 *(13.9)* ***5.4***	92
P3	284 *90.1*	31 ***9.9***	315
Total	371 *91.2*	36 ***8.8***	407

Tableau 19.2. Fréquences absolues et relatives de la position de P1 et P3 dans *Froissart*

Exemples

(66) « Ha ! tres chiers sires, puis que je apassai par deça la mer en grant peril, ensi que vous savés, je ne vous ai requis ne don demandet. Or vous **prie je** humlement et reqier en propre don que, pour le Fil a sainte Marie et pour l'amour de mi, vous voelliés avoir de ces siis honmes merchi. » (*Froissart*, p. 848)

(67) La dame respondi : « Dieus i ait part ! » Adont **prist elle** congiet au conte de Hainnau et a la contesse, et les remerchia moult doucement de la bonne et honnourable requelloite que fait li avoient. (*Froissart*, p. 70)

4.1.11.2. Expression du sujet[121]

	Sp exprimés	Sp non exprimés	Total Sp
P1	92 *(22.6)* *75.4*	30 *(3.4)* ***24.6***	122
P3	315 *26.8*	859 ***73.2***	1174
Total	407 *31.4*	889 ***68.6***	1296

Tableau 19.3. Fréquences absolues et relatives de l'expression de P1 et P3 dans *Froissart*

[120] Bruit : 7.4% pour les P1 préverbaux, 36% pour les P3 préverbaux, 0% pour les P1 postverbaux, 19.5% pour les P3 postverbaux. Il y a 2 incises, qui représentent 6% des 30 P3 postverbaux.
[121] Le bruit est de 74.9% pour la collecte des sujets non exprimés.

Exemples :

(68) « Madame, vechi vostre chevalier qui n'a pour le present que faire, ne a quoi entendre. Si **voel** estre en vostre service, et **n'entenderai** jamais a autre cose, si vous **averai remis** en Engleterre. (*Froissart*, p. 61)

(69) Or se perchut li dis messires Hues li Espensiers que on murmuroit sur lui ; si **doubta** trop fort que mauls ne l'en presist. (*Froissart*, p. 48)

4.1.12. *ESTOIRE DE GRISELDIS EN RIMES ET PAR PERSONNAGES*

Nombre de mots	18 909
Date	1395
Forme	vers
Domaine	littéraire
Genre	dramatique
Dialecte	traits picards
Texte étiqueté sans vérification	

Tableau 20.1. Caractéristiques de *Griseldis*

4.1.12.1. Position du sujet[122]

	Sp-Verbe	Verbe-Sp	SpV + VSp
P1	116 *(87.9)* ***88.6***	15 *(71.4)* ***11.4***	131
P3	16 *72.7*	6 ***27.3***	22
Total	132 *86.3*	21 ***13.7***	153

Tableau 20.2. Fréquences absolues et relatives de la position de P1 et P3 dans *Griseldis*.

Exemples :

(70) 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, p. 80)

(71) Son bon los de toutes pars sonne,
Et toujours croist sa renommee,
De bonne eure **feust elle nee**.

[122] Bruit : 9.4% pour les P1 préverbaux, 72.9% pour les P3 préverbaux, 31.8% pour les P1 postverbaux, 68.4% les P3 postverbaux. Il n'y a pas d'incise.

Quant a dame l'avons eüe,
Longuement l'avons congneüe ;
(*Griseldis*, p. 49)

4.1.12.2. Expression du sujet[123]

	Sp exprimés	Sp non exprimés	Total Sp
P1	131 *(85.6)* *34.9*	244 *(69.7)* ***65.1***	375
P3	22 *17.2*	106 ***82.8***	128
Total	153 *30.4*	350 ***69.6***	503

Tableau 20.3. Fréquences absolues et relatives de l'expression de P1 et P3 dans *Griseldis*

Exemples :

(72) LE QUINT CHEVALIER qui moult estoit ancien.
Messeigneurs, en la moye foy,
Simples homs **suis** et petit **say**
Et en moi petit d'avis **ay**
Pour ce devant vous proposer.
(*Griseldis*, p.10)

(73) Et depuis qu'il l'a congneü
En **a** deux beaux enfans **eü**
Qu'on ne scet qu'ils sont devenuz.
Trop en **est blamez** et **tenuz**
A rigoreux de ses subgez.
(*Griseldis*, p. 70)

4.1.13. *MANIERES DE LANGAGE*

Nombre de mots	20 282
Date	1396 et 1399
Forme	mixte
Domaine	didactique
Genre	manuel
Dialecte	non défini
Texte étiqueté sans vérification	

Tableau 21.1. Caractéristiques de *Manieres*

[123] Bruit pour la collecte des sujets non exprimés : 65.9%.

4.1.13.1. Position du sujet[124]

	Sp-Verbe	Verbe-Sp	SpV + VSp
P1	326 *(72)* *97*	10 *(50)* ***3***	336
P3	127 *92.7*	10 ***7.3***	137
Total	453 *95.8*	20 ***4.2***	473

Tableau 21.2. Fréquences absolues et relatives de la position de P1 et P3 dans *Manieres*

Exemples :

(74) Mon amy, estez vous jun unquore ?
Nonil, veraiment, sir, j'en ay dyné tresbien, Dieu merci.
Va donques a la dame de ciens et bevez ovesque ele, et puis aprés venez a vostre bosoigne.
Maister, si **frai je**. (*Manieres*, p. 35)

(75) Et quant il serra pres de son hostel, enquore **est il** si sot q'il ne sciet mye bien droit aler avant, qu'il en a un autre foiz demandé la voie. Donque **dit il** ainsi a primer homme qu'il encontre... (*Manieres*, p. 29)

4.1.13.2. Expression du sujet[125]

	Sp exprimés	Sp non exprimés	Total Sp
P1	336 *(61.9)* *95.5*	16 *(16.8)* ***4.5***	352
P3	137 *63.4*	79 ***36.6***	216
Total	473 *83.3*	95 ***16.7***	568

Tableau 21.3. Fréquences absolues et relatives de l'expression de P1 et P3 dans *Manieres*

Exemples :

(76) Dame, viendra il tost ?
Par ma foy, sir, je ne sçai. Vous purrés a moy dire vostre volunté.
Nonil, je lui vouldroy dire moy mesmes.
Sir, amontés et vous beverés, se vous pleast.
Non **fray** a ceste fois, par vostre congé.
(*Manieres*, p. 27)

(77) Lors dist l'escuier : 'Ma tresgentele dame, je vous remercie soveraigment de cuer de vostre amour et curtaisie. '

[124] Bruit : 2.4% pour les P1 préverbaux, 37.1% pour les P3 préverbaux, 67.7% pour les P1 postverbaux, 58.2% les P3 postverbaux. Il y a 18 incises, qui représentent 64.3% des 28 P3 postverbaux.
[125] Le bruit est très élevé (93.7%) pour la collecte des sujets non exprimés : les scores assez médiocres de ce texte s'expliquent par le lexique. Voir note 117.

Et puis aprés, quant l'escuier avoit congee de son seignour pur aler a sez amis, si **s'en ala** bien matyn a l'ajournant (*Manieres,* p. 14)

4.1.14. *QUINZE JOYES DE MARIAGE*

Nombre de mots	39 404
Date	*ca.*1400
Forme	prose
Domaine	littéraire
Genre	nouvelles
Dialecte	non défini
Texte étiqueté avec vérification	

Tableau 22.1. Caractéristiques de *Quinze Joyes*

4.1.14.1. Position du sujet[126]

	Sp-Verbe	Verbe-Sp	SpV + VSp
P1	306 *(42.8)* *96.5*	11*(25.6)* ***3.5***	317
P3	408 *92.7*	32 ***7.3***	440
Total	714 *94.3*	43 ***5.7***	757

Tableau 22.2. Fréquences absolues et relatives de la position de P1 et P3 dans *Quinze Joyes*

Exemples :

(78) Par Dieu, mon amy, voire ! Mes je ne semble que a une chamberiere emprés elle ; non **fais je** emprés ma seur et si **sui ge** aisnee d'elle, qui est laide chouse (*Quinze joyes,* p. 40).

(79) Jamés n'avra joye et est de mervoilles qu'il ne entre en desesperance, et si **feroit il** si n'estoit qu'il est sages homs. Si lui convient prendre en pacience, quar aultre remyde n'y **peut il** metre; (*Quinze Joyes*, p. 76)

[126] Bruit : 4% pour les P1 préverbaux, 18.7% pour les P3 préverbaux, 21.4% pour les P1 postverbaux, 6.8% pour les P3 postverbaux. Il y a 187 incises, qui représentent 85.4% des 219 P3 postverbaux.

4.1.14.2. Expression du sujet[127]

	Sp exprimés	Sp non exprimés	Total Sp
P1	317 *(41.9)*	58 *(8.9)*	375
	84.5	***15.5***	
P3	440	591	1031
	42.7	***57.3***	
Total	757	649	1406
	53.9	***46.1***	

Tableau 22.3. Fréquences absolues et relatives de l'expression de P1 et P3 dans *Quinze Joyes*

Exemples :

(80) Ainxin, regardans cestes paines qu'ilz prennent pour joies, considerans la repugnance qui est entre leur entendement et le mien et de pluseurs aultres, **me suy delicté**, en les regardant noer en la nasse ou ilz sont si bien embarrez, a escripre icelles [.XV.] joies de mariage a leur consolacion,... (*Quinze joyes*, p. 5)

(81) Puis vendra l'autre nuit, qu'elle se couchera ; et aprés qu'elle sera couchee, le proudomme escoutera si elle dort et **avisera** si elle a les braz bien couvers et la **couvrera** s'il est mestier. Lors **fera** semblant de s'esvoillier et le proudomme lui dit... (*Quinze joyes*, p. 10)

L'ensemble des données collectées pertinentes se répartit ainsi :

	Sp préverbal	Sp postverbal	Sp non exprimé	total
P1	1 709	252	1 488	3 449
P3	2 300	401	9 691	12 392
total	4 009	653	11 179	**15 841**

Tableau 23. Fréquences absolues des sujets préverbaux, postverbaux et non exprimés.

4.2. Remarques sur le bruit

Avant de proposer un essai de synthèse sur les fréquences d'inversion et de non-expression de Sp, il convient de faire quelques remarques sur le bruit qui apparaît à l'issue des différentes collectes.

Le bruit a deux origines. L'une réside dans le mauvais étiquetage des verbes. L'étiquetage erroné de certaines formes conduit en effet à collecter comme « verbe » des formes qui n'en sont pas. Ce problème reste néanmoins assez marginal, la vérification préalable de l'index ayant montré que les formes étiquetées « verbe » à tort et présentant plus de 5 occurrences sont rares (dans *Froissart*, où elles sont un peu

[127] Bruit pour la collecte des sujets non exprimés : 79.4%.

plus nombreuses, la requête a été adaptée afin d'éliminer ces formes). Certes les formes étiquetées « verbe » et qui apparaissent moins de 5 fois n'ont pas été vérifiées, or ce sont évidemment les plus nombreuses[128]. On collecte par ailleurs les *il* impersonnels, du fait que, dans les textes dont l'étiquetage n'a pas été vérifié, les requêtes pour les sujets exprimés n'ont pas été restreintes aux seuls sujets personnels, l'étiquetage automatique étant peu performant à distinguer les *il* personnels des *il* impersonnels. La part de bruit résultant de la collecte de *il* impersonnel est élevée dans *Clari*, *Coustumes* et *Griseldis*.

Malgré cela, on constate que le bruit dans les textes dont l'étiquetage est fiable (*Roland, Queste et Quinze Joyes*) n'est pas significativement moins élevé que dans les autres textes, en particulier pour ce qui concerne les sujets non-exprimés. Cela prouve, d'une part que le rôle joué par les formes mal étiquetées est marginal dans le bruit, d'autre part que, en l'état actuel des requêtes, l'avantage majeur des textes dont l'étiquetage est fiable est l'absence de silence, puisque tous les verbes sont repérés.

La source principale du bruit réside donc dans l'imperfection des requêtes, qui collectent des contextes d'occurrences indésirables. En tête de ceux-ci se trouvent les subordonnées, et, pour les sujets inversés, les propositions interrogatives. C'est ainsi la présence des séquences V-Sp dans des structures interrogatives qui explique que le bruit s'élève à 40% pour les P1 inversés dans *Eneas*, à 58% et à 67 % pour P3 et P1 inversés dans *Manieres* (précisons qu'il s'agit à chaque fois de faibles effectifs : le pourcentage élevé de bruit ne recouvre donc que quelques occurrences). Les structures subordonnées restent néanmoins, sur l'ensemble des textes, en tête des facteurs de trouble. Souvent la séquence SpV, ou VSp, ou V sans sujet exprimé, se trouve dans la subordonnée elle-même, qui n'a pas été reconnue comme telle parce que le mot subordonnant se trouve trop loin du verbe, alors que les requêtes prévoient l'exclusion des mots subordonnants relativement proches du verbe. Par exemple :

(82) Il acorderent aus amiraus en tel maniere **que** si tost comme en leur avroit delivré Damiete, **il deliverroient** le roy et les autres riches homes qui la estoient. (*Joinville*, p. 176)

En outre, certains mots subordonnants ne sont pas compris dans les formes que la requête est supposée exclure. C'est parfois involontaire (il n'est pas facile de penser à

[128] On a pu observer dans les langues un rapport inverse entre fréquence d'un mot et nombre de mots : les mots les moins fréquents dans leur usage sont bien plus nombreux que les mots fréquents. A titre d'exemple, l'index des verbes de *Beroul* comprend 1085 formes, dont 107 ont une fréquence supérieure à 5 et 1078 une fréquence inférieure à 5.

tous et à toutes leurs formes...), mais parfois volontaire aussi : ainsi de *ou*, qui, à côté de ses nombreux usages relatifs, présente quelques usages de coordination, qu'il ne fallait pas manquer. Pour ces quelques occurrences, on a donc collecté beaucoup de *ou* subordonnant (en particulier dans *Joinville*). Choix peu rentable, mais assumé, l'objectif était de minimiser le silence.

Certaines des occurrences indésirables résultent d'un problème de frontières entre propositions (la requête ne distingue pas les limites de propositions), le sujet et le verbe collectés appartenant à des propositions différentes, comme dans l'exemple ci-dessous, issu d'une requête VSp, dans lequel *aparut* appartient à la subordonnée initale, et *il* à la proposition principale qui suit :

(83) Et quant li jorz **aparut il** se leva et ala oïr le servise Nostre Seignor. (*Queste*, p.139)

Dans l'ensemble, le bruit associé aux requêtes portant sur les sujets exprimés oscille entre 0 et 25% : les quelques pics (voir les notes ci-dessus) tiennent pour la plupart à la présence importante, en plus de celle de propositions subordonnées, de *il* impersonnels et de propositions interrogatives.

Le bruit reste gérable, du fait de la conjonction de pourcentages globalement peu élevés et d'effectifs ne dépassant pas 500 occurrences par requête (et dans de nombreux textes les effectifs collectés ne sont que de quelques dizaines). Si du point de vue théorique ce bruit est déplaisant, il n'est pas gênant du point de vue pratique, le nettoyage étant assez rapide. De plus, lorsque l'étiquetage des textes aura été vérifié, il sera possible d'éliminer les *il* impersonnels des requêtes. Par ailleurs, les requêtes sont certainement perfectibles, au moins en ce qui concerne l'élimination des structures interrogatives (en tout cas pour celles dont le point d'interrogation n'est pas trop éloigné de la séquence verbe-sujet), ainsi que certaines subordonnées, ne serait-ce qu'en retravaillant la liste des mots subordonnants. L'élimination de subordonnées reste néanmoins plus complexe que celle des interrogatives.

Le bruit associé à la collecte des sujets non-exprimés est en revanche beaucoup plus important. Dans un seul texte (exception faite de *Beroul* et du recours au couple *NotaBene-TIGERSearch*), il est inférieur à 60% (*Amile* : 55.4%). Il oscille sinon entre 60 et 70% dans plusieurs des textes, score déjà bien médiocre, et ce score connaît – hélas – des pics dans quelques textes. Le pourcentage de bruit monte ainsi à 74% dans *Clari*, à 75% dans *Froissart*, à 79% dans *Quinze Joyes*, à 94% dans *Manieres*, et à 96% dans *Coustumes*. Les résultats sont à la fois *très peu* satisfaisants

scientifiquement et extrêmement lourds du point de vue du traitement[129]. Une part importante du bruit tient ici aussi aux occurrences en structures subordonnées (que l'on peut espérer réduire), mais surtout à la présence des sujets nominaux (voir la répartition du bruit dans *Roland*, 3.3.3.4., tableau 4) que la souplesse de l'ordre des mots ne permet pas d'éliminer. Seul un enrichissement en termes morpho-syntaxiques et syntaxiques, qui permet l'élimination des sujets nominaux, peut réduire cette source de bruit.

La solution idéale étant celle qui encode l'absence de sujet, comme c'est le cas dans le complexe *NotaBene-TIGERSearch,* avec néanmoins – car les choses sont rarement parfaites – le problème de la collecte des verbes impersonnels (et des verbes de personnes 2, 4, et 5, mais un travail sur les désinences doit permettre de réduire le bruit de ce côté). Rappelons que le bruit associé à la collecte des verbes sans sujet exprimé n'est que de 14.9%.

C'est sur les deux phénomènes qui ont reculé – inversion et non-expression de Sp – que sont concentrés les essais de synthèse qui sont présentés ci-dessous.

4.3. Synthèse et interprétation des fréquences de l'inversion de Sp

Les fréquences d'inversion de Sp sont présentées selon différents modes de visualisation : tableaux, graphiques en lignes et graphiques en barres. Sont indiquées systématiquement la fréquence d'inversion de P1, celle de P3, et celle de P1+P3.

Le tableau a l'avantage d'indiquer précisément les fréquences, mais il rend malaisée l'appréhension des comparaisons entre les différentes fréquences, et leurs évolutions. Les graphiques, en lignes ou en barres, permettent au contraire une vision plus claire des proximités, des distances, des évolutions. Ils créent cependant une « image », une représentation, qu'il convient d'interpréter avec prudence, comme on le verra plus bas.

La dimension chronologique reste première parmi nos critères, raison pour laquelle les données sont présentées, prioritairement, selon un ordre chronologique (autant qu'il se peut : cf. les problèmes de datation précise). Mais dans la mesure où nous souhaitons évaluer la possible influence des autres paramètres, il m'a semblé utile, pour la présentation en tableaux, de classer les données aussi selon les variables des autres

[129] Pour mémoire : en moyenne, entre 500 et 700 occurrences ont été codées par heure (selon un codage basique : indication du caractère pertinent ou non de l'occurrence, et, le cas échéant, mention de la personne P1 ou P3).

paramètres : dialecte, forme et domaine. Le lecteur pourra ainsi plus facilement confronter aux données les remarques qui seront faites plus loin. J'ai laissé de côté le critère du genre : si le roman est bien représenté, les autres genres sont plus éparpillés, et ordonner un tableau selon ce critère ne présente guère d'intérêt. Des remarques seront en revanche faites à ce sujet dans les développements qui suivent.

4.3.1. Tableau des fréquences d'inversion de P1 et P3

Les chiffres indiqués correspondent tous à des fréquences.
Les chiffres en 'normal' indiquent les pourcentages respectifs des inversions de P1+P3, P1 et P3 sur l'ensemble équivalent des sujets exprimés (P1+P3, P1, et P3).
Le chiffre en italiques indique la part des P1 inversés sur l'ensemble P1+P3 inversés (par soustraction de 100, on a le pourcentage concernant P3). L'établissement de ce pourcentage avait pour but de mettre au jour les variations d'un texte à l'autre de la proportion de P1 sur les sujets inversés, et, surtout, de déceler d'éventuels liens entre cette proportion et la fréquence de l'inversion de P1.

Texte + date	Dialecte	Forme	Domaine	Invers. P1+P3	P1		Invers. P3
					Invers. P1	P1/P1+P3 inversés	
Roland (1100)	anglo-normand	vers	littéraire	21.1	17.8	*47.4*	25.3
Eneas (1155)	normand	vers	littéraire	9.7	12.6	*35.3*	8.7
Béroul (1165-1200)	traits normands	vers	littéraire	17.5	16.1	*51*	19.3
Ami Amile (1200)	non marqué	vers	littéraire	23.8	25.9	*52.7*	21.8
Clari (1205)	picard	prose	historique	42.6	0	*0*	51.1
Aucassin (début 13ème)	traits picards	mixte	littéraire	9.9	16.9	*64.7*	5.6
Miracles (1220)	non marqué	vers	religieux	21.9	19.6	*32.1*	23.2
Queste (1230)	non marqué	prose	littéraire	22.1	33.1	*53.8*	15.9
Coustumes (1283)	traits picards	prose	juridique	15.8	18.2	*14.8*	15.4
Joinville (1307)	non marqué	prose	historique	12.3	15	*52.6*	10.2
Froissart (1369-1400)	franco-picard	prose	historique	8.8	5.4	*13.9*	9.9
Griseldis (1395)	traits picards	vers	littéraire	13.7	11.4	*71.4*	27.3
Manières (1396-1399)	non marqué	mixte	didactique	4.2	3	*50*	7.3
Quinze Joyes (1400)	non marqué	prose	littéraire	5.7	3.5	*25.6*	7.3

Tableau 24.1 : Fréquence d'inversion de P1, P3, et P1+P3 : classement chronologique.

Texte + date	**Dialecte**	**Forme**	**Domaine**	**Invers. P1+P3**	**P1**		**Invers. P3**
					Invers. P1	**P1/P1+P3 inversés**	
Roland (1100)	anglo-normand	vers	littéraire	21.1	17.8	*47.4*	25.3
Eneas (1155)	normand	vers	littéraire	9.7	12.6	*35.3*	8.7
Béroul (1165-1200)	traits normands	vers	littéraire	17.5	16.1	*51*	19.3
Clari (1205)	picard	prose	historique	42.6	0	*0*	51.1
Aucassin (déb. 13è)	traits picards	mixte	littéraire	9.9	16.9	*64.7*	5.6
Coustumes (1283)	traits picards	prose	juridique	15.8	18.2	*14.8*	15.4
Froissart (1369-1400)	franco-picard	prose	historique	8.8	5.4	*13.9*	9.9
Griseldis (1395)	traits picards	vers	littéraire	13.7	11.4	*71.4*	27.3
Ami Amile (1200)	non marqué	vers	littéraire	23.8	25.9	*52.7*	21.8
Miracles (1220)	non marqué	vers	religieux	21.9	19.6	*32.1*	23.2
Queste (1230)	non marqué	prose	littéraire	22.1	33.1	*53.8*	15.9
Joinville (1307)	non marqué	prose	historique	12.3	15	*52.6*	10.2
Manières (1396-1399)	non marqué	mixte	didactique	4.2	3	*50*	7.3
Quinze Joyes (1400)	non marqué	prose	littéraire	5.7	3.5	*25.6*	7.3

Tableau 24.2 : Fréquence d'inversion de P1, P3, et P1+P3 : classement selon le dialecte.

Texte + date	Dialecte	Forme	Domaine	Invers. P1+P3	P1		Invers. P3
					Invers. P1	P1/P1+P3 inversés	
Roland (1100)	anglo-normand	vers	littéraire	21.1	17.8	*47.4*	25.3
Eneas (1155)	normand	vers	littéraire	9.7	12.6	*35.3*	8.7
Béroul (1165-1200)	traits normands	vers	littéraire	17.5	16.1	*51*	19.3
Ami Amile (1200)	non marqué	vers	littéraire	23.8	25.9	*52.7*	21.8
Miracles (1220)	non marqué	vers	religieux	21.9	19.6	*32.1*	23.2
Griseldis (1395)	traits picards	vers	littéraire	13.7	11.4	*71.4*	27.3
Aucassin (déb. 13è)	traits picards	mixte	littéraire	9.9	16.9	*64.7*	5.6
Manières (1396-1399)	non marqué	mixte	didactique	4.2	3	*50*	7.3
Clari (1205)	picard	prose	historique	42.6	0	*0*	51.1
Queste (1230)	non marqué	prose	littéraire	22.1	33.1	*53.8*	15.9
Coustumes (1283)	traits picards	prose	juridique	15.8	18.2	*14.8*	15.4
Joinville (1307)	non marqué	prose	historique	12.3	15	*52.6*	10.2
Froissart (1369-1400)	franco-picard	prose	historique	8.8	5.4	*13.9*	9.9
Quinze Joyes (1400)	non marqué	prose	littéraire	5.7	3.5	*25.6*	7.3

Tableau 24.3 : Fréquence d'inversion de P1, P3, et P1+P3 : classement selon la forme.

Texte + date	Dialecte	Forme	Domaine	Invers. P1+P3	P1		Invers. P3
					Invers. P1	P1/P1+P3 inversés	
Roland (1100)	anglo-normand	vers	littéraire	21.1	17.8	*47.4*	25.3
Eneas (1155)	normand	vers	littéraire	9.7	12.6	*35.3*	8.7
Béroul (1165-1200)	traits normands	vers	littéraire	17.5	16.1	*51*	19.3
Ami Amile (1200)	non marqué	vers	littéraire	23.8	25.9	*52.7*	21.8
Aucassin (déb. 13è)	traits picards	mixte	littéraire	9.9	16.9	*64.7*	5.6
Queste (1230)	non marqué	prose	littéraire	22.1	33.1	*53.8*	15.9
Griseldis (1395)	traits picards	vers	littéraire	13.7	11.4	*71.4*	27.3
Quinze Joyes (1400)	non marqué	prose	littéraire	5.7	3.5	*25.6*	7.3
Clari (1205)	picard	prose	historique	42.6	0	*0*	51.1
Joinville (1307)	non marqué	prose	historique	12.3	15	*52.6*	10.2
Froissart (1369-1400)	franco-picard	prose	historique	8.8	5.4	*13.9*	9.9
Miracles (1220)	non marqué	vers	religieux	21.9	19.6	*32.1*	23.2
Coustumes (1283)	traits picards	prose	juridique	15.8	18.2	*14.8*	15.4
Manières (1396-1399)	non marqué	mixte	didactique	4.2	3	*50*	7.3

Tableau 24.4 : Fréquence d'inversion de P1, P3, et P1+P3 : classement selon le domaine.

Signalons en préalable aux remarques qui suivent que la proportion de P1 inversés sur l'ensemble P1 + P3 inversés connaît une importante variation d'un texte à l'autre : de 0% (*Clari*) à 71.4% (*Griseldis*). Dans 6 des 14 textes, les P1 inversés représentent environ la moitié des Sp inversés, dans 3 textes la proportion oscille entre 25 et 35%, dans 2 textes elle est inférieure à 15% (et un texte n'a pas d'occurrence), et elle dépasse les 60% dans 2 textes. De plus, il n'apparaît pas d'évolution, les fréquences plus ou moins élevés se répartissant sur toute la période considérée. Les autres critères (dialecte, forme et domaine) ne jouent pas plus de rôle discriminant.
Si l'on considère plus spécifiquement la relation entre la proportion de P1 sur les Sp (P1+P3) inversés et la fréquence d'inversion de P1, on constate que c'est l'absence de lien qui prévaut : une sous-représentation de P1 peut aussi bien être associée à un pourcentage d'inversion très faible (*Froissart*) que relativement élevé (*Coustumes*). De même, à une proportion d'environ 50% de P1 (sur les Sp inversés) peuvent correspondre des fréquences d'inversion variant du simple au double (*Joinville* : 15% et *Queste* : 33%). Le même constat vaut pour P3. Ainsi une relativement faible proportion de P3 sur l'ensemble des Sp inversés peut aussi bien être associée à une fréquence d'inversion de 27.3% (*Griseldis*), que de 5.6% (*Aucassin*).
Pour conclure sur ce point, il n'y a de toute évidence pas de corrélation entre la proportion de P1 ou P3 sur l'ensemble P1+P3 inversés et leur fréquence d'inversion.

Ce point réglé, le premier constat face à ces chiffres est que, si l'on met *Clari* à part, l'intervalle de variation des pourcentages d'inversion est relativement étroit, entre 3% et 33%, ces deux fréquences concernant l'inversion de P1. L'écart est légèrement moindre pour P3, de 5.6% à 27.3%. Globalement, l'inversion pronominale est donc peu élevée, toujours nettement inférieure à 50%, hormis dans *Clari*, texte tout à fait atypique puisqu'il associe absence d'inversion pour P1 et fréquence notablement élevée pour P3[130].
Notons aussi d'ores et déjà, et nous y reviendrons en considérant les représentations graphiques ci-dessous, que l'on n'observe pas de tendance évolutive nette.
Un autre point mérite d'être souligné. Si certains textes présentent des pourcentages d'inversion de P1 et P3 assez proches l'un de l'autre, d'autres au contraire accusent un écart notable. Ainsi en est-il de *Clari* bien sûr, mais aussi d'*Aucassin*, de la *Queste*, de

[130] Il est atypique dans notre corpus mais aussi au regard de bon nombre des textes d'ancien français : à titre d'exemple l'inversion pronominale dans *La Mort Artu* (1230) et dans *Tristan en prose* (1240) est inférieure à 30%.

Griseldis, de *Manieres* et de *Quinze Joyes*, avec un rapport supérieur à 2 entre les deux pourcentages, au profit de P1 ou de P3 selon les textes.
Cette divergence prouve l'intérêt qu'il y a à distinguer les différentes personnes, à ne pas perdre trace de leur singularité sous un pourcentage global d'inversion. Ainsi, sous les 22% d'inversion de la *Queste* se cachent des fréquences d'inversion de 33% pour P1 et de 16% pour P3. Nous verrons plus loin que la distinction selon les personnes s'impose plus encore pour l'expression du sujet, les écarts étant bien plus marqués.

Le calcul du khi2 permet de mieux apprécier le caractère éventuellement surprenant de la distribution de la position au regard de P1 ou P3.

4.3.2. Valeurs de khi2 significatives

Comme je l'ai indiqué plus haut, la distribution entre les variables 'position' et 'personne' est assez homogène dans la majorité des textes, ce dont témoignent des valeurs de khi2 qui oscillent entre 0.1 à 3.3. Ces chiffres correspondent à des probabilités qui se situent entre 0.07 et 0.74. Cela signifie que la distribution observée a entre 7% et 74% de chances d'être le fruit du hasard (voir tableaux 30.1. à 30.4 ci-dessous). En revanche, les 5 textes dans lesquels les pourcentages d'inversion de P1 et P3 sont assez divergents présentent des valeurs de khi2 beaucoup plus élevées. En ce qui concerne *Clari*, on ne peut accorder de valeur réelle au calcul en raison de la nullité de l'un des effectifs (absence de P1 postverbal)[131].

Voici ci-dessous les calculs de khi2 pour *Aucassin*, *Queste*, *Griseldis*, *Manieres* et *Quinze Joyes*. Pour chaque texte, un tableau présente, pour chacune des 4 cases : les effectifs réels (en normal) les effectifs théoriques (en italiques), la contribution relative (en italiques gras) à la liaison entre les variables exprimée par le khi2, et le signe de la différence entre effectifs réels et effectifs théoriques. Enfin, on a surligné les cases du tableau qui contribuent le plus au khi2.

[131] Un seuil de 5 occurrences par case est recommandé pour le calcul du khi2.

Aucassin et Nicolete

	Sp préverbal		Sp postverbal		Total
P1	54	*58.58*	11	*6.42*	65
	6.15	-	***56.1***	+	
P3	101	*96.42*	6	*10.58*	107
	3.73	+	***34.1***	-	
Total		155		17	172

Tableau 25 : Effectifs réels et théoriques et contribution au khi2 (5.81) dans *Aucassin*

Le khi2 est de 5.81; la probabilité qui lui est associée est de 0,0159 : la répartition observée a 1.6% de chances d'être le résultat d'une distribution aléatoire. Notons que l'un des effectifs dépasse tout juste le seuil critique de 5 (6 occurrences de P3 postverbal).

Il ressort des données du tableau 25 une attraction forte entre P1 et la position postverbale (contribution de 56% au khi2), et, conjointement, une répulsion entre cette même position et P3.

La Queste del Saint Graal

	Sp préverbal		Sp postverbal		Total
P1	129	*150.3*	64	*42.7*	193
	14.2	-	***49.9***	+	
P3	290	*268.7*	55	*76.3*	345
	7.9	+	***27.9***	-	
Total		419		119	538

Tableau 26 : Effectifs réels et théoriques et contribution au khi2 (21.3) dans *Queste*

Le khi2 est de 21.3, valeur bien plus élevée que celle de *Aucassin*, et à laquelle est associée une probabilité très faible, largement inférieure à 0.001 ($3{,}926*10^{-06}$). La répartition observée a donc moins de 0.1% de chances d'être le résultat d'une distribution aléatoire.

Il ressort des données du tableau 26 une attraction forte entre P1 et la position postverbale (contribution de 49.9% au khi2), et, conjointement, une répulsion entre cette même position et P3. La configuration est donc, de ce point de vue, analogue à celle observée dans *Aucassin*, en dépit d'une probabilité largement moindre.

Rappelons que *Queste* est le texte qui présente la fréquence d'inversion de P1 la plus élevée du corpus.

Estoire de Griseldis en rimes et par personnages

	Sp préverbal		Sp postverbal		Total
P1	116	*113.02*	15	*17.98*	131
	1.97	+	***12.4***	-	
P3	16	*18.98*	6	*3.02*	22
	11.75	-	***73.9***	+	
Total	132		21		153

Tableau 27 : Effectifs réels et théoriques et contribution au khi2 (3.98) dans *Griseldis*

Le khi2 est de 3.98 ; la probabilité qui lui est associée est de 0,046. La répartition observée a donc presque 5% de chances d'être le résultat d'une distribution aléatoire. Elle n'est que moyennement significative, ce dont témoigne l'écart assez faible entre effectifs réels et théoriques, sauf pour P3 en position postverbale.

Plus précisément, il ressort des données du tableau 27 une attraction très forte entre P3 et la position postverbale (contribution de 73.9% au khi2), et, conjointement, une répulsion, moindre, entre P3 et la position préverbale, et entre P1 et la position postverbale.

La configuration est donc différente de celles d'*Aucassin* et de *Queste*, textes dans lesquels on observe au contraire une répulsion entre P3 et la position postverbale, et une attraction entre P1 et cette même position.

Il est intéressant de noter que, alors que l'écart entre les fréquences d'inversion de P1 et P3 est important (respectivement 11.4% et 27.3%), le khi2 reste assez bas et la répartition observée n'est pas très éloignée du modèle théorique. Les deux textes suivants, *Manieres* et *Quinze Joyes,* présentent un cas de figure inverse : alors que l'écart entre les pourcentages d'inversion de P1 et P3 est moindre (rapport de 1 à 2) le khi2 est légèrement plus élevé, la répartition observée moins probable.

Manieres de langage

	Sp préverbal		Sp postverbal		Total
P1	326	*321.8*	10	*14.2*	336
	1.22	+	***27.7***	-	
P3	127	*131.2*	10	*5.8*	137
	3	-	***68***	+	
Total		453		20	473

Tableau 28 : Effectifs réels et théoriques et contribution au khi2 (4.49) dans *Manieres*

Le khi2 est de 4.49 et la probabilité qui lui est associée est de 0.034. La répartition observée n'a donc que 3.4% de chances d'être le résultat d'une distribution aléatoire. Comme dans *Griseldis*, et à l'inverse donc d'*Aucassin* et *Queste*, les données du tableau 28 traduisent une forte attraction entre P3 et la position postverbale (contribution de 68% au khi2). Elle est associée à une répulsion entre P1 et cette même position (contribution de 27.7% au khi2).
Notons qu'il s'agit, comme dans *Aucassin* et *Griseldis*, de faibles effectifs.

Quinze joyes de mariage

	Sp préverbal		Sp postverbal		Total
P1	306	*299*	11	*18*	317
	3.3	+	***54.8***	-	
P3	408	*415*	32	*25*	440
	2.4	-	***39.5***	+	
Total		714		43	757

Tableau 29 : Effectifs réels et théoriques et contribution au khi2 (4.97) dans *Quinze Joyes*

Le khi2 est de 4.97 ; la probabilité qui lui est associée est de 0.026 : la répartition observée a environ 2.6 % de chances d'être le résultat d'une distribution aléatoire.
La configuration est assez similaire à celle de *Manieres*, puisque l'on retrouve une attraction entre P3 et la position postverbale, et une répulsion entre cette même position et P1. Mais les contributions respectives sont en revanche différentes : c'est la liaison (répulsion) entre P1 et la position postverbale qui contribue le plus au khi2 (54.8%).

Rappelons que ces deux textes, *Manieres* et *Quinze Joyes*, présentent des pourcentages d'inversion très voisins : entre 3 et 3.5% pour P1 et 7.3% pour P3.

Les cinq textes présentent donc des configurations partiellement différentes. Dans deux d'entre eux (*Aucassin* et *Queste*), c'est l'attraction entre position postverbale et P1 qui contribue le plus à la liaison générale entre les variables, tandis que dans deux autres (*Griseldis* et *Manieres*) c'est au contraire l'attraction entre P3 et cette même position qui est importante ; enfin, dans *Quinze Joyes*, c'est la répulsion entre P1 et la position postverbale qui s'avère déterminante. Au-delà de ces divergences, il ressort que, d'une manière générale, c'est la valeur 'postverbale' de la variable 'position' qui joue un rôle déterminant, exerçant une force attractive ou au contraire répulsive vis-à-vis de P1 ou P3.

Que dire des relations qui s'établissent entre ces 5 textes à travers cette liaison forte entre P1 ou P3 et la position postverbale ? On notera simplement que quatre d'entre eux appartiennent au domaine littéraire, et que *Grisedis*, *Manieres* et *Quinze Joyes*, qui présentent une attraction forte entre P3 et la position postverbale, sont tous trois datés de la fin du 14ème siècle. C'est là sans doute le fait le plus marquant.

Voici ci-dessous, pour l'ensemble des textes, une récapitulation des valeurs de khi2 et des probabilités qui leur sont associées. Les valeurs significatives sont surlignées. Celles de *Clari* sont précédées d'un '?' en raison du caractère très peu fiable du calcul.
Comme pour les fréquences présentées en 4.3.1., les données sont ordonnées, dans des tableaux successifs, en fonction des différents critères : chronologie, dialecte, forme et domaine.
On pourra noter que, hormis la convergence temporelle entre *Griseldis*, *Manieres* et *Quinze Joyes*, aucune autre affinité ne semble se nouer entre textes présentant des khi2 voisins.

Texte + date	**Dialecte**	**Forme**	**Domaine**	**khi2**	**Probabilité**
***Roland* (1100)**	anglo-normand	vers	littéraire	1.49	0.22
***Eneas* (1155)**	normand	vers	littéraire	1.24	0.24
Béroul (1165-1200)	traits normands	vers	littéraire	0.47	0.49
***Ami Amile* (1200)**	non marqué	vers	littéraire	0.52	0.47
***Clari* (1205)**	picard	prose	historique	? 32	?$1.5*10^{-8}$
***Aucassin* (déb. 13ème)**	traits picards	mixte	littéraire	5.81	0.016
***Miracles* (1220)**	non marqué	vers	religieux	0.22	0.63
***Queste* (1230)**	non marqué	prose	littéraire	21.3	$3,9*10^{-06}$
***Coustumes* (1283)**	traits picards	prose	juridique	0.11	0.74
***Joinville* (1307)**	non marqué	prose	historique	3.3	0.07
***Froissart* (1369-1400)**	franco-picard	prose	historique	1.71	0.19
***Griseldis* (1395)**	traits picards	vers	littéraire	3.98	0.046
***Manières* (1396-1399)**	non marqué	mixte	didactique	4.49	0.034
***Quinze Joyes* (1400)**	non marqué	prose	littéraire	4.97	0.026

Tableau 30.1. Distribution de la position préverbale ou postverbale de P1 et P3 : valeur du khi2 et probabilité associée. Présentation selon le critère chronologique

Texte + date	**Dialecte**	**Forme**	**Domaine**	**khi2**	**Probabilité**
Roland (1100)	**anglo-normand**	vers	littéraire	1.49	0.22
Eneas (1155)	**normand**	vers	littéraire	1.24	0.24
Béroul (1165-1200)	**traits normands**	vers	littéraire	0.47	0.49
Clari (1205)	**picard**	prose	historique	? 32	? $1.5*10^{-8}$
Aucassin (déb. 13ème)	**traits picards**	mixte	littéraire	5.81	0.016
Coustumes (1283)	**traits picards**	prose	juridique	0.11	0.74
Froissart (1369-1400)	**franco-picard**	prose	historique	1.71	0.19
Griseldis (1395)	**traits picards**	vers	littéraire	3.98	0.046
Joinville (1307)	**non marqué**	prose	historique	3.3	0.07
Ami Amile (1200)	**non marqué**	vers	littéraire	0.52	0.47
Miracles (1220)	**non marqué**	vers	religieux	0.22	0.63
Queste (1230)	**non marqué**	prose	littéraire	21.3	$3,9*10^{-06}$
Manières (1396-1399)	**non marqué**	mixte	didactique	4.49	0.034
Quinze Joyes (1400)	**non marqué**	prose	littéraire	4.97	0.026

Tableau 30.2. Distribution de la position préverbale ou postverbale de P1 et P3 : valeur du khi2 et probabilité associée. Présentation selon le critère dialectal

Texte + date	**Dialecte**	**Forme**	**Domaine**	**khi2**	**Probabilité**
Roland (1100)	anglo-normand	**vers**	littéraire	1.49	0.22
Eneas (1155)	normand	**vers**	littéraire	1.24	0.24
Béroul (1165-1200)	traits normands	**vers**	littéraire	0.47	0.49
Ami Amile (1200)	non marqué	**vers**	littéraire	0.52	0.47
Miracles (1220)	non marqué	**vers**	religieux	0.22	0.63
Griseldis (1395)	traits picards	**vers**	littéraire	3.98	0.046
Aucassin (déb. 13ème)	traits picards	**mixte**	littéraire	5.81	0.016
Manières (1396-1399)	non marqué	**mixte**	didactique	4.49	0.034
Clari (1205)	picard	**prose**	historique	*32	$*1.5*10^{-8}$
Queste (1230)	non marqué	**prose**	littéraire	21.3	$3,9*10^{-06}$
Coustumes (1283)	traits picards	**prose**	juridique	0.11	0.74
Joinville (1307)	non marqué	**prose**	historique	3.3	0.07
Froissart (1369-1400)	franco-picard	**prose**	historique	1.71	0.19
Quinze Joyes (1400)	non marqué	**prose**	littéraire	4.97	0.026

Tableau 30.3. Distribution de la position préverbale ou postverbale de P1 et P3 : valeur du khi2 et probabilité associée. Présentation selon le critère de la forme

Texte + date	**Dialecte**	**Forme**	**Domaine**	**khi2**	**Probabilité**
Roland (1100)	anglo-normand	vers	**littéraire**	1.49	0.22
Eneas (1155)	normand	vers	**littéraire**	1.24	0.24
Béroul (1165-1200)	traits normands	vers	**littéraire**	0.47	0.49
Ami Amile (1200)	non marqué	vers	**littéraire**	0.52	0.47
Aucassin (déb. 13ème)	traits picards	mixte	**littéraire**	5.81	0.016
Queste (1230)	non marqué	prose	**littéraire**	21.3	$3{,}9*10^{-06}$
Griseldis (1395)	traits picards	vers	**littéraire**	3.98	0.046
Quinze Joyes (1400)	non marqué	prose	**littéraire**	4.97	0.026
Clari (1205)	picard	prose	**historique**	? 32	? $1.5*10^{-8}$
Joinville (1307)	non marqué	prose	**historique**	3.3	0.07
Froissart (1369-1400)	franco-picard	prose	**historique**	1.71	0.19
Miracles (1220)	non marqué	vers	**religieux**	0.22	0.63
Coustumes (1283)	traits picards	prose	**juridique**	0.11	0.74
Manières (1396-1399)	non marqué	mixte	**didactique**	4.49	0.034

Tableau 30.4. Distribution de la position préverbale ou postverbale de P1 et P3 : valeur du khi2 et probabilité associée. Présentation selon le critère du domaine

4.3.3. Représentations graphiques : ‘courbes’ et graphiques en barres

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 (voir graphiques 1-3 ci-dessous).

Elles dessinent aussi une évolution, dont la nature est définie par l'axe des abscisses. Dans le cas présent, cet axe comprend des textes, et les dates qui les accompagnent, et les textes étant ordonnés selon un ordre aussi chronologique que possible, il est tentant de considérer la 'courbe' comme témoignant d'une évolution temporelle. C'est en partie vrai : de *Roland* à *Quinze joyes*, il y a une progression des fréquences d'inversion, et celle-ci se situe bien dans le temps. Mais c'est en partie faux aussi, et cela pour deux raisons. La première est conjoncturelle, et tient à la datation des textes du corpus. Ainsi, de *Amile* à la *Queste*, il n'y a que 30 ans d'écart mais 5 textes (dans notre corpus). Or dans la mesure où, par défaut, l'espace sur l'axe des abscisses est le même entre chaque texte, la distance du premier au dernier de ces 5 textes est à peine plus courte que celle qui couvre les 170 ans et les 6 textes qui suivent. La représentation du temps sur l'axe des abscisses est donc en partie déformée, et partant celle de la 'courbe', la période du début du 13ème siècle étant étirée de façon disproportionnée. L'option qui consiste à adopter une gradation chronologique stricte de l'axe des abscisses conduit malheureusement à une mauvaise distinction visuelle des textes, *a fortiori* lorsque le graphique comprend plusieurs lignes. J'ai donc opté pour un compromis, et représenté chacune des trois courbes de deux façons : sans chronologie proportionnée de l'axe des abscisses, mais avec un repérage clair des textes, et avec un repérage fidèle de la chronologie, mas sans mention des textes, la seconde représentation permettant une appréhension plus juste du mouvement qui se dessine. Remarque qui nous mène au point suivant.

La seconde raison pour laquelle la représentation dite en courbe est en partie trompeuse est plus structurelle. En effet, même lorsque l'espacement entre les textes est proportionnel au temps qui les sépare, il n'en reste pas moins que 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[132]. Or l'exemple 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 nos courbes ci-

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

dessous n'auront plus du tout la même allure (de même, pour la représentation de la non-expression, si l'on supprime *Coustumes*).

Se pose donc à nouveau la question déjà abordée (dans le mémoire de synthèse et ici en 3.1.) de la représentativité du corpus, qui, plus précisément, se décline 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 plutôt les lignes, des représentations graphiques dessinent bien des évolutions, et non le simple parcours d'un texte à l'autre ? Cela n'a pas été fait à ma connaissance, et c'est complexe. Une autre question est en partie liée à celle-ci : faut-il fusionner les textes d'une même date, ou appartenant à une même période, et fournir des chiffres globaux ? Quand on voit 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[133] à une même date ne doit en tout cas pas occulter les fréquences propres à chaque texte : l'information doit rester disponible. Que la fréquence globale d'inversion de Sp dans les textes de 1200-1230 de ce corpus s'élève à 24% ne doit pas occulter que, derrière ce chiffre, se trouvent des fréquences qui oscillent effectivement autour de 20%, mais aussi une fréquence de 42% et une de 10% (sans parler de la distinction entre les différentes personnes).

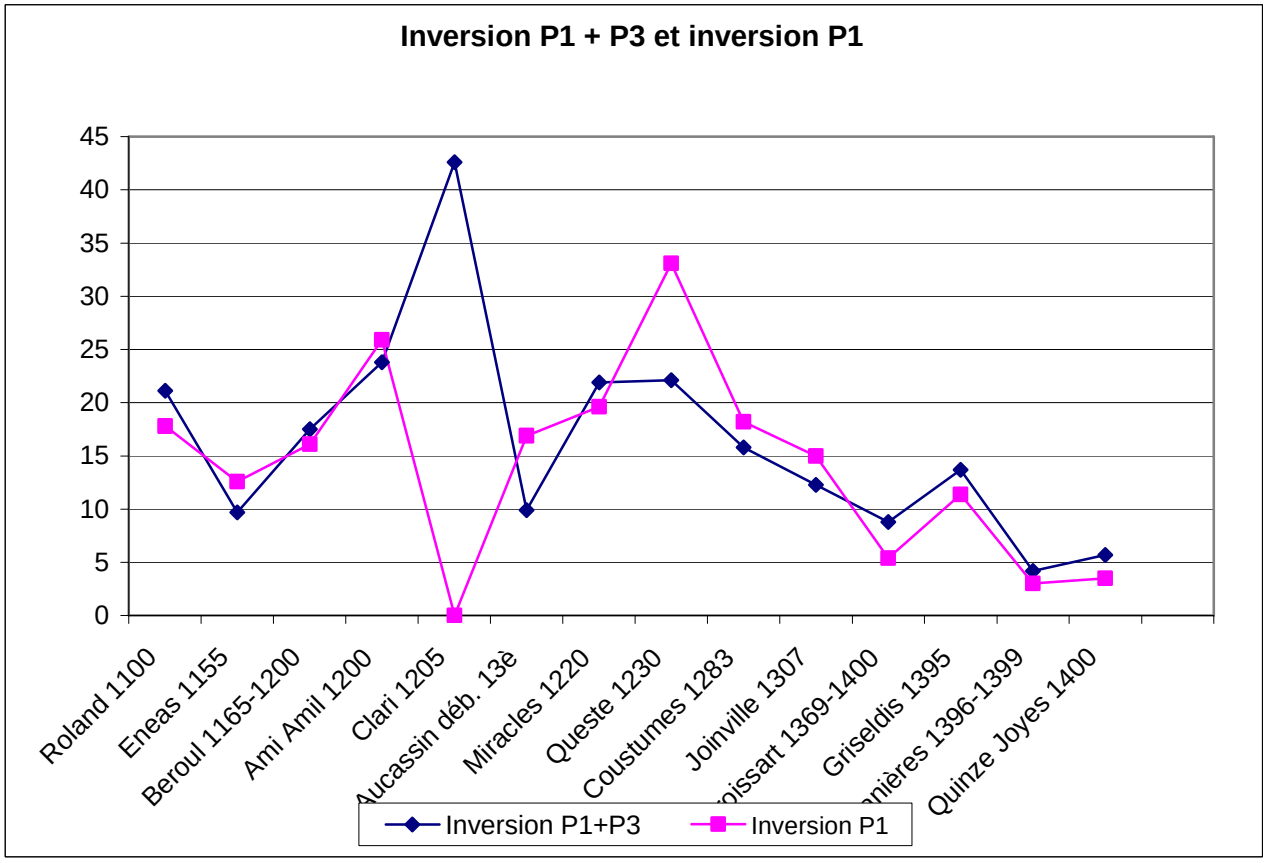
Se pose ici plus généralement la question de l'usage que l'on fait des moyennes, dont la vocation est précisément de niveler les différences. La mesure de la variance et de l'écart-type permet de dépasser le caractère réducteur de la moyenne en soulignant la spécificité de chacune des valeurs qui ont permis de la calculer. Etablir une moyenne suppose de regrouper des valeurs selon un certain critère. Celui de la date peut en être un. A ce titre, faire une moyenne des fréquences d'inversion dans l'ensemble du présent corpus n'aurait pas grand sens ; mais établir une moyenne pour les textes du début du 13ème serait pertinent. Il serait dès lors intéressant d'évaluer la dispersion des différentes fréquences autour de la moyenne. Mais cette étude n'étant pas

[133] En toute rigueur il faudrait aussi préciser la notion de contemporanéité. Quel écart temporel supporte-t-elle ? 2 textes de 1210 et 1240 sont-ils contemporains ? La réponse dépend pour une large part de la diachronie sur laquelle on travaille. Si la période étudiée s'étale de 1100 à 1400, comme c'est le cas ici, on peut les considérer comme contemporains (ce qui n'exclut d'ailleurs pas qu'un changement puisse intervenir entre les deux dates : la contemporanéité s'inscrit dans un cadre pour l'étude, cadre qui ne présuppose aucune périodisation). En revanche, si la période étudiée ne couvre que la période 1200-1250, on ne considèrera pas les deux textes comme contemporains. C'est une question de granularité.

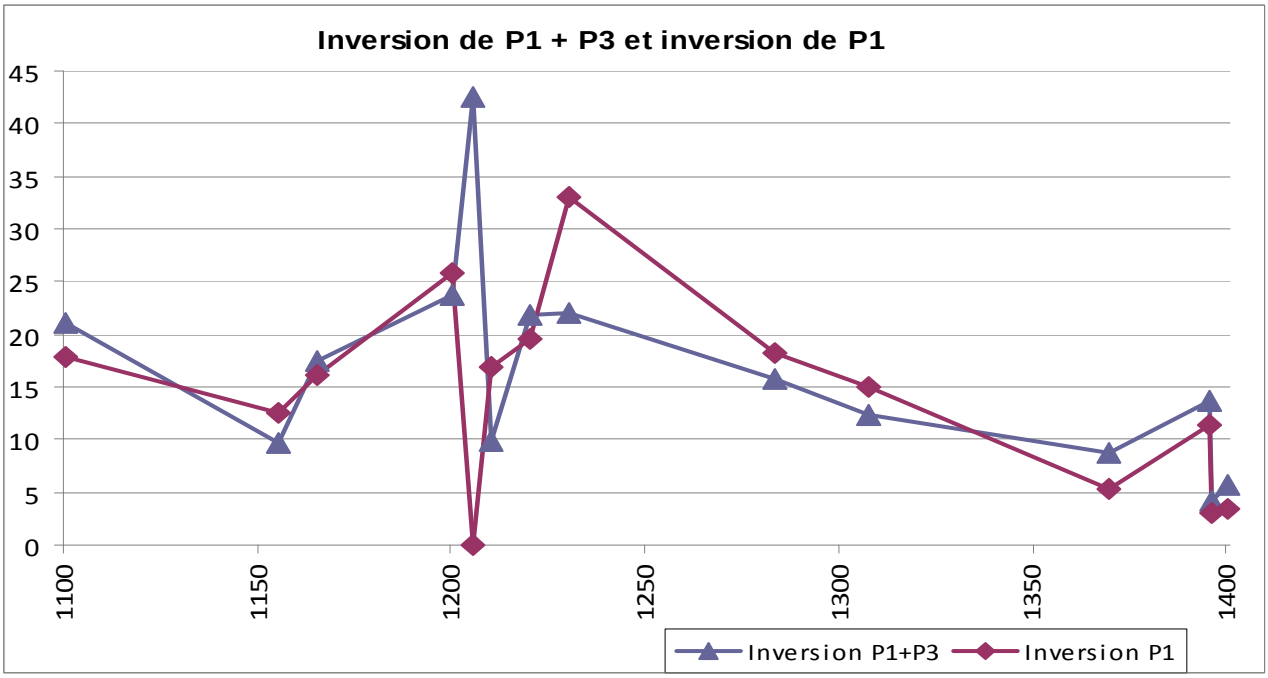
spécifiquement dédiée aux textes du début du 13ème siècle, je remets à plus tard de tels calculs.

Au-delà des seuls chiffres et de leur représentation, ces différentes questions renvoient à la difficulté qu'il y a à dégager des tendances générales sans pour autant écraser les spécificités propres aux textes. C'est d'autant plus complexe que, en langue ancienne, la variation syntaxique entre textes est bien plus importante qu'aujourd'hui. Nous reviendrons sur ce point plus loin.

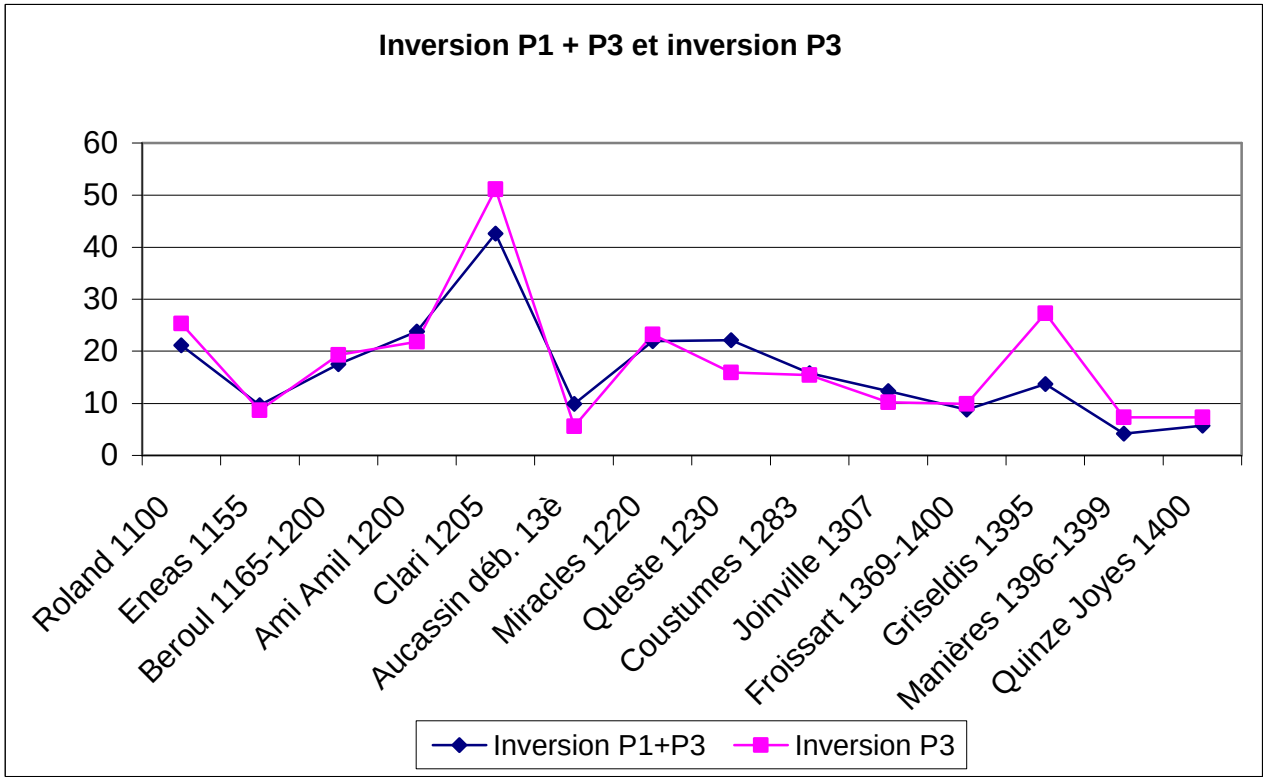
On trouvera ci-dessous les graphiques en 'lignes' qui traduisent les fréquences d'inversion de Sp dans les différents textes. Les fréquences d'inversion de P1 et de P3 sont successivement mises en perspective avec celle de P1 + P3 (graphiques 1 et 2), puis elles sont réunies dans le graphique 3. Pour chacun de ces trois graphiques il est proposé une double représentation, la première sans axe chronologique proportionné, la seconde avec un tel axe.
Le graphique 4 enfin propose une représentation en barres des fréquences P3+P1, P1 et P3. Il a l'avantage, comparé aux graphiques en lignes, de rendre plus lisibles les écarts des 3 fréquences pour chacun des textes, en particulier quand celles-ci sont assez proches.



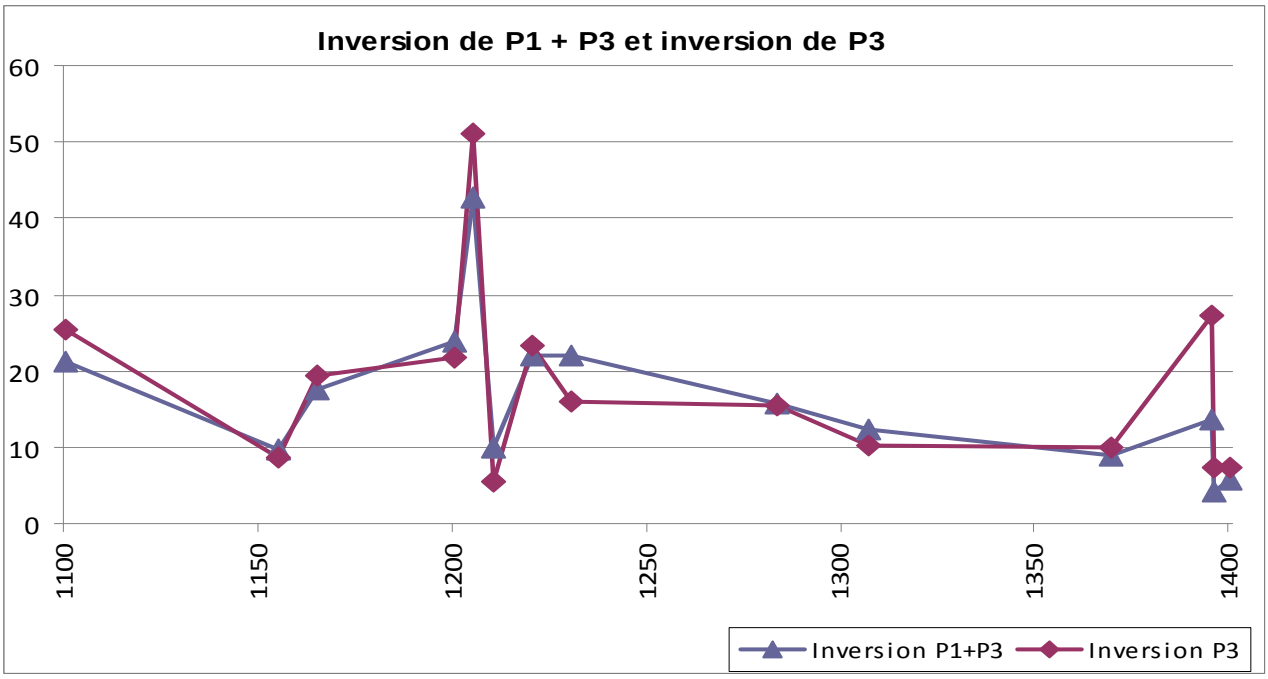
Graphique 1a : Inversion P1 + P3 et inversion P1 (axe des ordonnées : pourcentages)



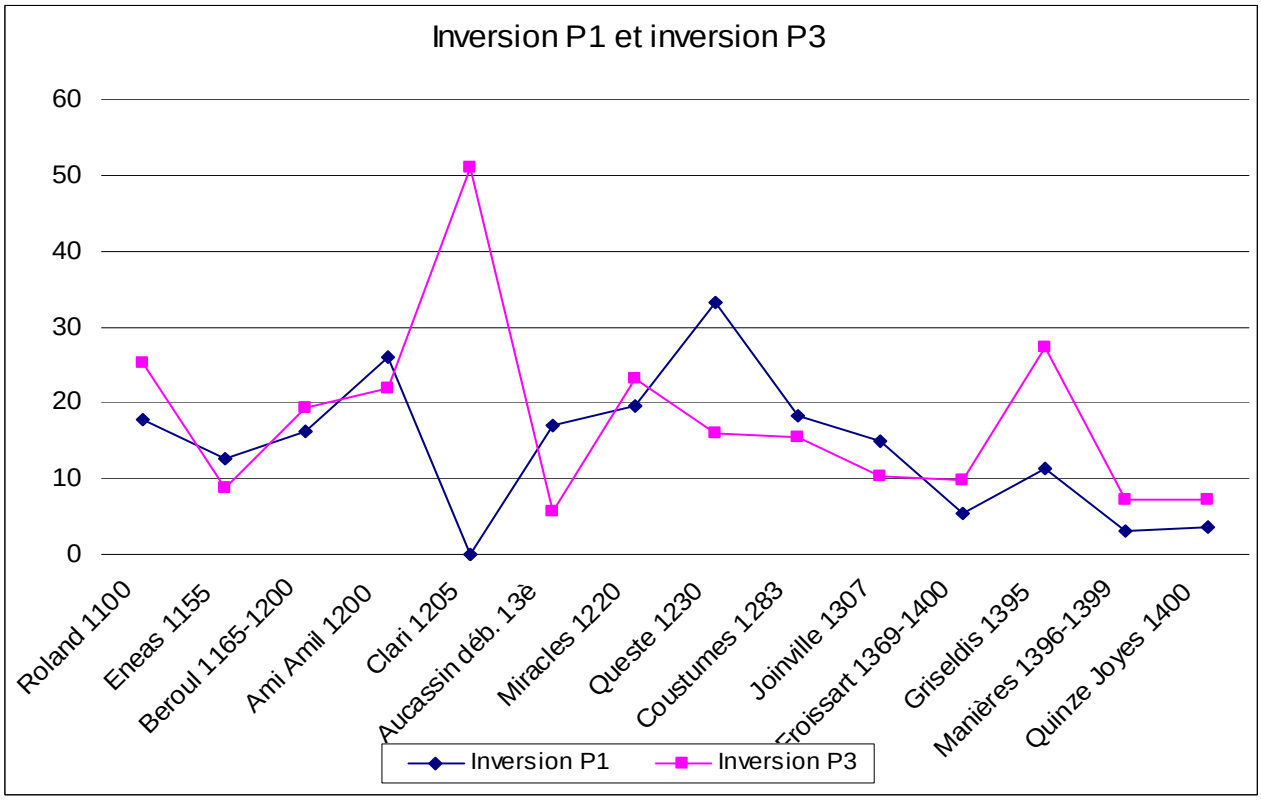
Graphique 1b : Inversion P1 + P3 et inversion P1. Chronologie stricte



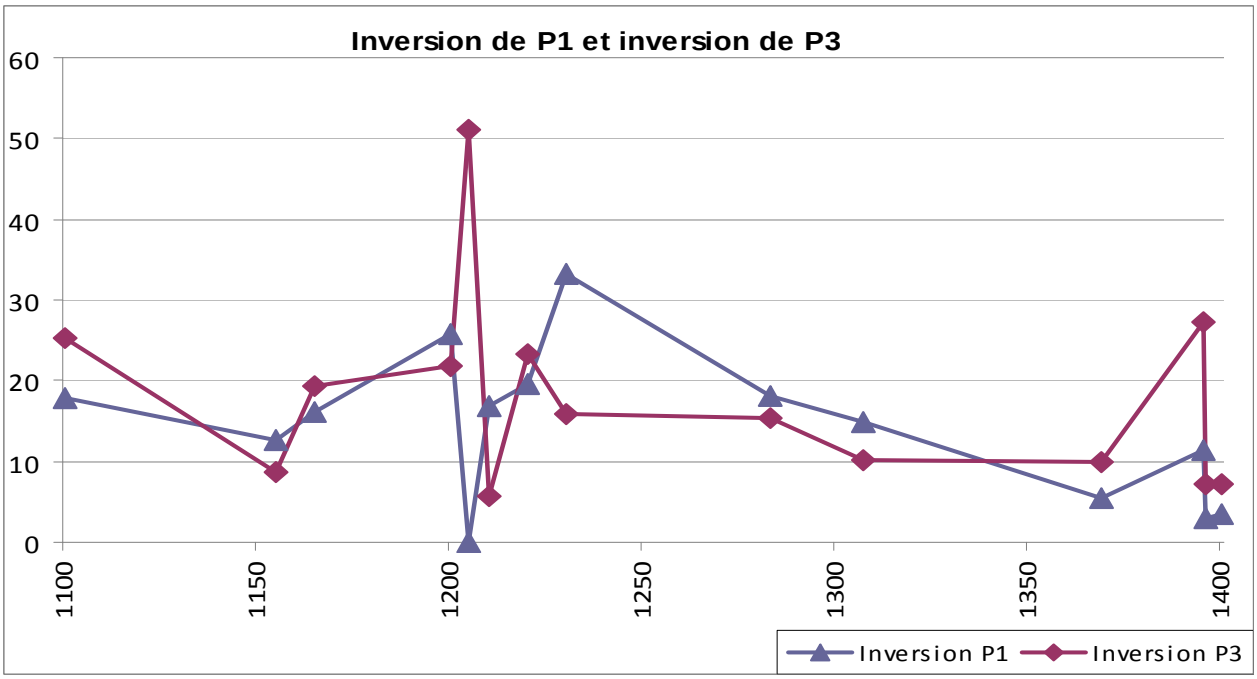
Graphique 2a : Inversion P1 + P3 et inversion P3



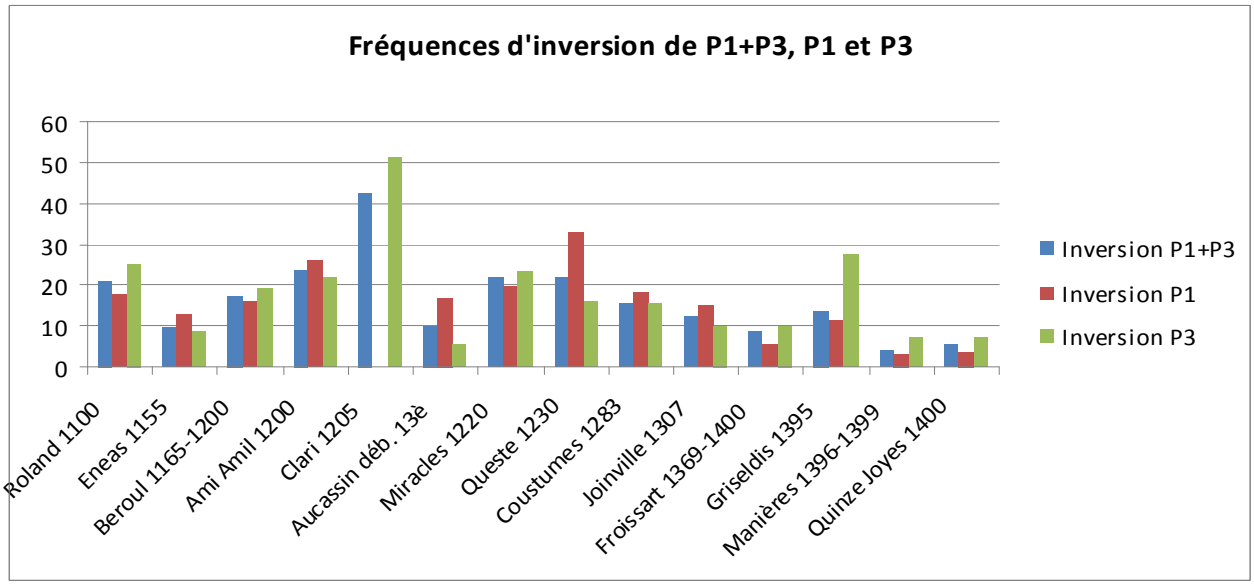
Graphique 2b : Inversion P1 + P3 et inversion P3 (chronologie stricte)



Graphique 3a : Inversion P1 et inversion P3



Graphique 3b : Inversion P1 et inversion P3 (chronologie stricte)



Graphique 4 : Fréquences d'inversion de P1+P3, P1 et P3.

Gardant à l'esprit les réserves qui ont été formulées ci-dessus concernant la portée de ces représentations, on peut proposer quelques éléments d'interprétation.

Premier constat général, l'amplitude de variation reste globalement faible sur l'ensemble de la période, *Clari* mis à part. De plus, si le texte le plus ancien (*Roland*) présente des fréquences d'inversion plus élevées que celles des textes les plus tardifs (*Griseldis*, *Manieres* et *Quinze Joyes*), la progression de ces fréquences à travers les différents textes n'a rien d'une ligne régulièrement descendante.

Par ailleurs, si l'on compare les graphiques 1 (P1+P3 et P1) et 2 (P1 + P3 et P3), on voit que le mouvement de la ligne de P1 + P3 est globalement plus proche de celui de la ligne de P3 que de celle de P1 (ce qui s'explique en partie par la supériorité numérique des effectifs globaux de P3). L'observation du graphique 4, qui réunit les 3 fréquences d'inversion, permet de déterminer précisément les points de divergence et de convergence. Seul un texte, *Griseldis*, présente une fréquence d'inversion P3 nettement divergente de celle de P1+P3, laquelle est en revanche assez voisine de celle de P1. La spécificité de *Griseldis* tient à la conjonction d'un pourcentage d'inversion P3 assez élevé (27%) et d'une faiblesse des effectifs de P3 (22 en tout, contre 131 P1).

Ce texte mis à part, les points de divergence concernent P1 : dans la *Queste* et dans *Griseldis*, la fréquence d'inversion de P1 dépasse celle de P3, mais aussi celle de P1+P3, en raison de la sur-représentation de P1 parmi les Sp inversés. Quant à *Clari*, la conjonction de l'absence de P1 inversé, d'un pourcentage très élevé d'inversion P3

(51%), et d'un effectif total de P3 5 fois supérieur à celui de P1 produit une convergence maximale entre P1+P3 et P3, et totalement nulle entre P1+P3 et P1.

Il n'apparaît donc pas d'évolution diachronique[134] spectaculaire. Certes la relative rareté du phénomène d'inversion pronominale dès les plus anciens textes fait que l'on part de fréquences assez basses et que la chute ne peut être vertigineuse.
Dans cette lente descente parsemée de sursauts, on peut néanmoins tenter de dégager quelques convergences au regard de différents critères : date, domaine et genre, forme et dialecte.

4.3.4. Essai d'interpétation des tableaux et des représentations graphiques

Considérons tout d'abord le **paramètre chronologique** : que se passe-t-il entre *Roland* et les *Quinze Joyes* ?
La datation des textes du corpus permet de distinguer nettement deux groupes : d'une part les 3 textes de la fin du 14ème, très resserrés dans le temps, auxquels on peut adjoindre *Froissart* (1369-1400) ; d'autre part les textes du premier tiers du 13ème siècle (*Ami*, *Clari*, *Aucassin*, *Miracles* et *Queste*) ; mais l'on pourrait aussi déplacer un peu la fenêtre temporelle vers la gauche, laisser les *Miracles* et la *Queste* de côté et inclure *Beroul* dans l'intervalle, qui couvrirait alors plutôt le tournant 12ème-13ème siècles. *Roland* est quant à lui nettement détaché au début du 12ème siècle, tandis qu'*Eneas* est à mi-chemin entre *Roland* et les textes du début du 13ème siècle. *Coustumes* et *Joinville* forment un autre groupe, au tournant du 14ème siècle.

Il est hélas assez malaisé de dégager des affinités au sein des regroupements, si ce n'est pour *Manieres* et *Quinze Joye,* qui présentent des fréquences d'inversion voisines pour P1 et pour P3. On peut rapprocher *Froissart* de ces deux textes : les fréquences, bien qu'un peu plus élevées, ne dépassent en effet pas la barre des 10%, P3 s'imposant devant P1, comme dans les deux autres textes. Ces trois textes mis à part, les affinités entre textes quant à leurs fréquences d'inversion ne sont guère liées à leur proximité temporelle. Le graphique 4 constitue, à mon sens, la visualisation optimale pour comparer conjointement les 3 fréquences d'inversion dans les différents textes.

[134] Evolution diachronique entendue comme restreinte aux textes du corpus.

Si l'on considère à la fois la valeur des fréquences et le rapport entre P1 et P3 (fréquences proches ou éloignées), les deux seuls textes qui présentent une certaine proximité sur ces deux points sont *Roland* et *Miracles* (fréquences autour de 20%), deux textes séparés de 130 ans, dont le seul point commun est qu'ils sont en vers. Si l'on se montre un peu moins exigeant, on peut rapprocher *Amile* de ces deux textes, dont il diffère cependant du point de vue de la fréquence d'inversion de P1, légèrement supérieure. On peut aussi trouver des affinités entre *Beroul* et *Coustumes*, dont les fréquences sont toutes légèrement inférieures à 20%, P3 l'emportant légèrement dans *Beroul*, alors que c'est P1 qui domine de peu dans *Coustumes*. Un siècle sépare les deux textes, qui ne présentent en outre aucune caractéristique commune : dialecte, genre, forme, tout les sépare. Constat similaire pour *Eneas* et *Joinville*, dont les fréquences tournent autour de 10%, avec une petite avance pour l'inversion de P1 : 150 ans entre eux, et rien en commun... *Griseldis* partage avec ces deux textes une fréquence d'inversion de P1 qui avoisine 10%, mais elle s'en distingue par un pourcentage d'inversion de P3 nettement supérieur (qui ne concerne cependant que quelques occurrences, d'où une moyenne P1+P3 qui reste proche de la fréquence de P1). Très éloigné d'*Eneas*, mais distant aussi de *Joinville* de presque un siècle, *Griseldis* n'a de traits communs qu'avec *Eneas* (deux siècles et demi les séparent par ailleurs), mais à vrai dire peu distinctifs : le fait d'être en vers et d'appartenir au domaine littéraire (mais à des genres toutefois bien différents : roman pour l'un, théâtre pour l'autre).

Les trois textes qui restent sont difficiles à relier à d'autres. *Clari*, tout d'abord, est simplement inclassable : impossible de le mettre en relation avec un quelconque texte du corpus, tant par la fréquence très élevée de l'inversion de P3 que par l'absence d'inversion de P1. Pour *Aucassin* et *Queste*, c'est différent. Ces deux textes, séparés d'une trentaine d'années[135], offrent un rapport analogue entre l'inversion de P3 et celle de P1, cette dernière dominant largement. Mais les fréquences de *Queste* sont nettement supérieures à celles d'*Aucassin.* De ce point de vue *Queste* se rapproche un peu des *Miracles* et d'*Amile*, tandis qu'*Aucassin* n'est pas très éloigné d'*Eneas*.

Il faut bien avouer qu'il ne se dégage rien de très probant de ces essais de rapprochements. Pour finir, tentons de regrouper les textes selon que la fréquence d'inversion de P1 ou P3 et celle de la moyenne P1+P3 se trouvent au-dessus ou en

[135] J'ai opté pour la datation qui situe *Aucassin* au début du 13ème siècle (et non pas entre entre le dernier quart du 12ème et la première moitié du 13ème).

dessous de 10% ou de 20%. Au regard de ce critère, on réunit ainsi *Roland*, *Amile*, *Clari*, *Miracles* et *Queste*, qui dépassent les 20%. On réunit par ailleurs *Beroul*, *Coustumes*, *Joinville* et *Griseldis* qui se situent entre 10% et 20%. A la limite des 10%, on trouve *Eneas*, *Aucassin* et *Froissat*, tandis que *Manieres* et *Quinze Joyes* se trouvent en dessous du seuil de 10%.
A l'issue de ces différentes observations, une seule certitude se dégage : à partir de *Coustumes*, et mis à part le pourcentage de P1 dans *Griseldis* (qui ne correspond cependant qu'à 6 occurrences), toutes les fréquences d'inversion de ce corpus sont inférieures à 20%.

Plusieurs remarques ont été faites dans les paragraphes précédents sur les caractéristiques externes sans que n'apparaissent d'affinités très nettes. Ainsi, si l'on considère le **critère dialectal** (tableau 30.2.), on a du mal à opérer des regroupements entre les trois textes anglo-normands ou présentant des traits normands du 12ème siècle. Il en va de même pour les textes picards ou picardisants : rien de commun entre *Clari* et *Aucassin*[136], ni entre *Coustumes*, *Froissart* et *Griseldis*.
Pour ce qui est du **domaine** et du **genre** (tableau 30.4), les trois textes historiques – *Clari*, *Joinville* et *Froissart* – présentent des configurations assez différentes. Parmi les textes littéraires, on peut considérer qu'il y a une certaine proximité entre *Roland* et *Amile*, bien qu'ils divergent sur l'inversion de P1. Rien de tel en revanche entre les 3 romans (*Beroul*, *Eneas*, *Queste*) et le recueil de nouvelles *Quinze Joyes*.

Il faut bien admettre, à l'issue de cette tentative modérément fructueuse de mise au jour de tendances associées à différents critères (temps, dialecte, et domaine et genre en particulier), qu'un désordre assez grand règne parmi nos textes. Même si les choses semblent commencer à s'ordonner – à la baisse – à la fin du 13ème siècle, la variation idiolectale apparaît encore comme la grande maîtresse de l'inversion, au moins en ce qui concerne les fréquences.

[136] M. Roques, dans son introduction à *Aucassin et Nicolete*, évoque une ressemblance de la prose de ce texte avec celle de *Clari* : ce n'est en tout cas pas vrai pour la syntaxe du pronom personnel sujet.

4.4. Synthèse et interprétation des fréquences de la non-expression de Sp

Les modalités de présentation des données de la non-expression de Sp sont les mêmes que celles de l'inversion : tableaux et représentations graphiques, khi2 significatifs et essai d'interprétation.

Pour la présentation en tableaux (synthèses des fréquences et des khi2), il est à nouveau proposé plusieurs classements des données, en fonction des différents paramètres : temps, dialecte, forme et domaine.

4.4.1. Tableau des fréquences de non expression de P1 et P3

Rappel : les chiffres indiqués correspondent tous à des fréquences relatives.

Les chiffres en 'normal' indiquent les pourcentages respectifs des non-expressions de P1+P3, P1 et P3 sur l'ensemble équivalent des sujets exprimés ou non (P1+P3, P1, et P3).

Le chiffre en italiques indique la proportion des P1 non-exprimés sur l'ensemble P1+P3 non-exprimés ; il a été établi afin de repérer d'éventuels liens entre cette proportion et la fréquence de la non-expression de P1.

Texte + date	**Dialecte**	**Forme**	**Domaine**	**non-expres. P1+P3**	**P1**		**non-expres. P3**
					non-expres. P1	**P1/ P1+P3 non expr.**	
Roland (1100)	anglo-normand	vers	littéraire	89.3	62.2	*11*	94.4
Eneas (1155)	normand	vers	littéraire	82.8	71.7	*14.3*	85
Beroul (1165-1200)	traits normands	vers	littéraire	82.7	61.7	*18.8*	89.7
Ami Amile (1200)	non marqué	vers	littéraire	81.5	64.1	*19.7*	87.3
Clari (1205)	picard	prose	historique	80.8	25	*1.3*	83.3
Aucassin (début 13ème)	traits picards	mixte	littéraire	69.7	37.5	*9.8*	77
Miracles (1220)	non marqué	vers	religieux	84	70.5	*16.3*	87.3
Queste (1230)	non marqué	prose	littéraire	68.6	21.5	*4.5*	76.5
Coustumes (1283)	traits picards	prose	juridique	26.9	26.7	*12.7*	27
Joinville (1307)	non marqué	prose	historique	44.1	21.1	*4.5*	54.1
Froissart (1369-1400)	franco-picard	prose	historique	68.6	24.6	*3.4*	73.2
Griseldis (1395)	traits picards	vers	littéraire	69.6	65.1	*69.7*	82.8
Manieres (1396-1399)	non marqué	mixte	didactique	16.7	4.5	*16.8*	36.6
Quinze Joyes (1400)	non marqué	prose	littéraire	46.1	15.5	*8.9*	57.3

Tableau 31.1. Fréquences de non-expression de P1, P3, et P1+P3. Classement chronologique.

Texte + date	Dialecte	Forme	Domaine	non-expres. P1+P3	P1		non-expres. P3
					non-expres. P1	P1/ P1+P3 non expr.	
Roland (1100)	**anglo-normand**	vers	littéraire	89.3	62.2	*11*	94.4
Eneas (1155)	**normand**	vers	littéraire	82.8	71.7	*14.3*	85
Beroul (1165-1200)	**traits normands**	vers	littéraire	82.7	61.7	*18.8*	89.7
Clari (1205)	**picard**	prose	historique	80.8	25	*1.3*	83.3
Aucassin (début 13ème)	**traits picards**	mixte	littéraire	69.7	37.5	*9.8*	77
Coustumes (1283)	**traits picards**	prose	juridique	26.9	26.7	*12.7*	27
Froissart (1369-1400)	**franco-picard**	prose	historique	68.6	24.6	*3.4*	73.2
Griseldis (1395)	**traits picards**	vers	littéraire	69.6	65.1	*69.7*	82.8
Ami Amile (1200)	**non marqué**	vers	littéraire	81.5	64.1	*19.7*	87.3
Miracles (1220)	**non marqué**	vers	religieux	84	70.5	*16.3*	87.3
Queste (1230)	**non marqué**	prose	littéraire	68.6	21.5	*4.5*	76.5
Joinville (1307)	**non marqué**	prose	historique	44.1	21.1	*4.5*	54.1
Manieres (1396-1399)	**non marqué**	mixte	didactique	16.7	4.5	*16.8*	36.6
Quinze Joyes (1400)	**non marqué**	prose	littéraire	46.1	15.5	*8.9*	57.3

Tableau 31.2. Fréquences de non-expression de P1, P3, et P1+P3. Classement selon le dialecte

Texte + date	Dialecte	Forme	Domaine	non-expres. P1+P3	P1		non-expres. P3
					non-expres. P1	P1/ P1+P3 non expr.	
Roland (1100)	anglo-normand	**vers**	littéraire	89.3	62.2	*11*	94.4
Eneas (1155)	normand	**vers**	littéraire	82.8	71.7	*14.3*	85
Beroul (1165-1200)	traits normands	**vers**	littéraire	82.7	61.7	*18.8*	89.7
Ami Amile (1200)	non marqué	**vers**	littéraire	81.5	64.1	*19.7*	87.3
Miracles (1220)	non marqué	**vers**	religieux	84	70.5	*16.3*	87.3
Griseldis (1395)	traits picards	**vers**	littéraire	69.6	65.1	*69.7*	82.8
Aucassin (début 13ème)	traits picards	**mixte**	littéraire	69.7	37.5	*9.8*	77
Manieres (1396-1399)	non marqué	**mixte**	didactique	16.7	4.5	*16.8*	36.6
Clari (1205)	picard	**prose**	historique	80.8	25	*1.3*	83.3
Queste (1230)	non marqué	**prose**	littéraire	68.6	21.5	*4.5*	76.5
Coustumes (1283)	traits picards	**prose**	juridique	26.9	26.7	*12.7*	27
Joinville (1307)	non marqué	**prose**	historique	44.1	21.1	*4.5*	54.1
Froissart (1369-1400)	franco-picard	**prose**	historique	68.6	24.6	*3.4*	73.2
Quinze Joyes (1400)	non marqué	**prose**	littéraire	46.1	15.5	*8.9*	57.3

Tableau 31.3. Fréquences de non-expression de P1, P3, et P1+P3. Classement selon la forme

Texte + date	Dialecte	Forme	Domaine	non-expres. P1+P3	P1		non-expres. P3
					non-expres. P1	P1/ P1+P3 non expr.	
Roland (1100)	anglo-normand	vers	**littéraire**	89.3	62.2	*11*	94.4
Eneas (1155)	normand	vers	**littéraire**	82.8	71.7	*14.3*	85
Beroul (1165-1200)	traits normands	vers	**littéraire**	82.7	61.7	*18.8*	89.7
Ami Amile (1200)	non marqué	vers	**littéraire**	81.5	64.1	*19.7*	87.3
Aucassin (début 13ème)	traits picards	mixte	**littéraire**	69.7	37.5	*9.8*	77
Queste (1230)	non marqué	prose	**littéraire**	68.6	21.5	*4.5*	76.5
Griseldis (1395)	traits picards	vers	**littéraire**	69.6	65.1	*69.7*	82.8
Quinze Joyes (1400)	non marqué	prose	**littéraire**	46.1	15.5	*8.9*	57.3
Clari (1205)	picard	prose	**historique**	80.8	25	*1.3*	83.3
Joinville (1307)	non marqué	prose	**historique**	44.1	21.1	*4.5*	54.1
Froissart (1369-1400)	franco-picard	prose	**historique**	68.6	24.6	*3.4*	73.2
Miracles (1220)	non marqué	vers	**religieux**	84	70.5	*16.3*	87.3
Coustumes (1283)	traits picards	prose	**juridique**	26.9	26.7	*12.7*	27
Manieres (1396-1399)	non marqué	mixte	**didactique**	16.7	4.5	*16.8*	36.6

Tableau 31.4. Fréquences de non-expression de P1, P3, et P1+P3. Classement selon le domaine.

Si l'on considère, comme nous l'avons fait pour l'inversion, la proportion de P1 non exprimés sur l'ensemble des P1 et P3 non exprimés, on constate que la variation est bien moindre et que, le plus souvent, P1 est largement minoritaire, à l'exception de

Griseldis (69.7%). Ce texte mis à part, la proportion de P1 ne dépasse pas les 20%, elle descend même sous la barre des 5% dans 4 textes, et elle est particulièrement faible dans *Clari.*

Certaines régularités se dégagent en ce qui concerne la relation entre la proportion de P1 sur l'ensembe P1+P3 non-exprimés et la fréquence de non-expression de P1. Il apparait en effet que, dans les textes dans lesquels la proportion de P1 est inférieure à 10%, la fréquence de sa non-expression ne dépasse pas 30%, à l'exception de *Aucassin*. Mais l'inverse n'est pas systématique : dans *Coustumes* et *Manieres* la fréquence de non-expression de P1 est inférieure à 30%, alors que la proportion de P1 est supérieure à 10%. On observe donc des tendances, et non des relations systématiques.

Plusieurs remarques s'imposent face aux fréquences rassemblées dans les tableaux 31.1-4. Tout d'abord la non-expression est globalement (P1 et P3 confondus) très fréquente dans tous les textes. A l'exception de 4 d'entre eux (*Coutusmes*, *Joinville*, *Manieres* et *Quinze Joyes*), sa fréquence dépasse 68%, et elle franchit même la barre de 80% dans 6 d'entre eux. Parmi les 4 textes plus réfractaires à la non-expression, *Joinville* et *Quinze Joyes* accusent néanmoins un pourcentage de non-expression proche de 45%. L'amplitude de variation de la non-expression est donc très large, entre 16.7% (*Manieres*) et 89.3% (*Roland*).

Toutefois, si l'on considère, non plus la non-expression globale de P1+ P3, mais les fréquences respectives de P1 et de P3 non exprimés, l'analyse des données se révèle plus complexe. Il apparaît tout d'abord que la fréquence de non-expression de P3 est bien plus élevée que celle de P1[137], à l'exception de *Coustumes*, texte dans lequel les pourcentages de P1 et P3 plafonnent tous deux à 27%. Dans la moitié des autres textes, le rapport entre la fréquence de non-expression de P3 et celle de P1 oscille entre 3 et 4. La non-expression de P3 est même 8 fois plus fréquente que celle de P1 dans *Manieres*. Mais l'amplitude de variation des fréquences reste élevée dans tous les cas : de 5% à 71.7% pour P1, et de 27% à 94% pour P3.

Insistons donc à nouveau sur la nécessité de ne pas écraser les spécificités de P1 et P3 sous les fréquences globales. Ainsi les pourcentages de non-expression de P1+P3 de *Froissart* et de *Griseldis* sont tous deux voisins de 69%, mais ils recouvrent des

[137] Et plus proche de celle de P1+P3 dans la mesure où P3 est nettement plus fréquent parmi les Sp non exprimés.

réalités bien différentes : une fréquence de non-expression de P1 de 24.6% pour *Froissart*, de 65% pour *Griseldis*.

Le calcul du khi2 permet d'évaluer précisément le caractère hétérogène de la distribution de P1 et P3 au regard de leur expression ou non-expression.

4.4.2. Valeurs de khi2 significatives

La conjonction de la moindre représentation de P1 parmi les Sp non exprimés et de sa fréquence de non-expression presque toujours inférieure à celles de P3 pouvait le laisser attendre : les khi2 calculés pour chacun des textes atteignent des valeurs très élevées et sont donc associés à des probabilités excessivement faibles que la distribution observée soit le fruit du hasard. Un texte cependant se singularise : *Coustumes*, pour lequel le calcul du khi2 met au jour une probabilité de 97% que la distribution observée soit le fruit du hasard. Ce texte mis à part, tous les autres présentent une configuration similaire, à savoir une attraction entre la non-expression et P3, et une répulsion entre cette même non-expression et P1. La régularité de ces liaisons recouvre néanmoins des attractions et des répulsions plus ou moins fortes, ainsi que des divergences variables entre effectifs réels et théoriques.
Voici donc la présentation du khi2, et des effectifs associés pour chacun des textes, ordonnés selon la valeur décroissante de leur khi2, et donc selon la valeur croissante de la probabilité d'une distribution aléatoire. Il est cependant évident que l'on est, pour la majorité des textes, face à des probabilités si infimes, largement inférieures à 0.001, que hiérarchiser les textes selon leur degré de probabilité n'a plus grand sens. On retiendra surtout de cet ordonnancement le caractère plus ou moins extrême des textes au regard d'une distribution caractérisée pour tous par son caractère non-homogène.

Pour chaque texte, un tableau présente, pour chacune des 4 cases, les effectifs réels (en normal) les effectifs théoriques (en italiques), la contribution relative à la liaison entre les variables exprimée par le khi2 (en italiques gras), et le signe de la différence entre effectifs réels et effectifs théoriques. Les cases du tableau qui contribuent le plus au khi2 sont surlignées.

La Queste del Saint Graal

	Sp exprimé		Sp non exprimé		Total
P1	193	*77.12*	53	*158.9*	246
	58.8	+	***26.8***	-	
P3	345	*460.9*	1125	*1009.1*	1 470
	9.8	-	***4.5***	+	
Total		538		1 178	1 715

Tableau 32 : Effectifs réels et théoriques et contribution au khi2 (296) dans *Queste.*

Le khi2 est de 296 : c'est une valeur extrêmement élevée. La probabilité qui lui est associée est de $2.4*10^{-66}$: cela signifie que la répartition observée n'a quasiment aucune chance de résulter d'une distribution aléatoire.

Il ressort des chiffres du tableau 32 une forte attraction entre l'expression de Sp et P1 (contribution de 58.8% au khi2) et, dans une moindre mesure, une répulsion entre cette même personne et la non-expression (contribution de 26.8%).

Rappelons que ce texte est aussi celui qui présente la fréquence d'inversion de P1 la plus élevée : 33%, et la distribution la plus marquée de P1 et P3 au regard de leur position (voir tableau 26). Il s'agit donc d'un texte dont la distribution tant de la position que de l'expression de P1 et P3 est surprenante.

Chanson de Roland

	Sp exprimé		Sp non exprimé		Total
P1	101	*28.5*	166	*238.5*	267
	75.2	+	***9***	-	
P3	79	*151.5*	1341	*1268.5*	1420
	14.1	-	***1.7***	+	
Total		180		1507	1687

Tableau 33 : Effectifs réels et théoriques et contribution au khi2 (245.4) dans *Roland*

Le khi2 est de 245.4, valeur qui reste très élevée. La probabilité ($2.54*10^{-55}$) que la distribution soit le fruit du hasard est toujours quasiment nulle.

On observe à nouveau une forte attraction entre expression de Sp et P1 (contribution de 75.2% au khi2), mais aussi, dans une moindre mesure, une répulsion entre expression et P3.

Quinze joyes de mariage

	Sp exprimé		Sp non exprimé		Total
P1	317	*201.9*	58	*173.1*	375
	33.8	+	***39.5***	-	
P3	440	*551.1*	591	*475.9*	1031
	12.3	-	***14.4***	+	
Total		757		649	1406

Tableau 34 : Effectifs réels et théoriques et contribution au khi2 (193.8) dans *Quinze Joyes*

Le khi2 est de 193.8**.** La probabilité qui lui est associée est de $4{,}6*10^{-44}$, c'est-à-dire à nouveau quasiment nulle. On retrouve une configuration proche de celle de *Queste*, mais avec une inversion de la force des liaisons : c'est la répulsion entre P1 et sa non-expression qui contribue le plus au khi2 (39.5%), suivie (de peu) par l'attraction entre P1 et son expression.

***Tristan* de Béroul**

	Sp exprimé		Sp non exprimé		Total
P1	149	*67.48*	240	*321.5*	389
	61.8	+	***13***	-	
P3	119	*200.52*	1037	*955.5*	1156
	20.8	-	***4.4***	+	
Total		268		1277	1545

Tableau 35 : Effectifs réels et théoriques et contribution au khi2 (159) dans *Beroul*

Le khi2 est de 159.26**,** et il correspond à une probabilité de $1.64*10^{-36}$, toujours extrêmement basse. La liaison la plus marquée se situe à nouveau entre P1 et son expression, sous forme d'attraction (contribution de 61.8% au khi2), et c'est ensuite la répulsion entre P3 et l'expression qui contribue le plus au khi2, comme dans *Roland*.

***Chroniques* de Froissart**

	Sp exprimé		Sp non exprimé		Total
P1	92	*38.3*	30	*83.7*	122
	62.1	+	***28.4***		
P3	315	*368.7*	859	*805.3*	1174
	6.4	-	***3***	+	
Total		407		889	1296

Tableau 36 : Effectifs réels et théoriques et contribution au khi2 (121.1) dans *Froissart*

Le khi2 est de 121.1, et il correspond à une probabilité de $3{,}70*10^{-28}$.

La liaison se caractérise prioritairement par une forte attraction entre P1 et l'expression (contribution de 62.1% au khi2), et par une répulsion, moins marquée, entre P1 et la non-expression.

Robert de Clari, *La Conqueste de Constantinople*

	Sp exprimé		Sp non exprimé		Total
P1	37	*9.42*	12	*39.58*	49
	77.5	+	***18.4***	-	
P3	180	*207.58*	900	*872.42*	1080
	3.5	-	***0.83***	+	
Total		217		912	1129

Tableau 37 : Effectifs réels et théoriques et contribution au khi2 (104) dans *Clari*

Le khi2 est de 104.53 ; la probabilité qui lui est associée est de $1.55*10^{-24}$. La configuration est très proche de celle du texte précédent : une forte attraction entre P1 et l'expression et, dans une moindre mesure, une répulsion entre P1 et la non-expression.

***Mémoires ou Vie de Saint Louis* de Joinville**

	Sp exprimé		Sp non exprimé		Total
P1	266	*188.44*	71	*148.56*	337
	30.7	+	***38.9***	-	
P3	353	*430.56*	417	*339.44*	770
	13.4	-	***17***	+	
Total		619		488	1107

Tableau 38 : Effectifs réels et théoriques et contribution au khi2 (104) dans *Joinville*

Le khi2 est de 104.1; la probabilité qui lui est associée est de $1{,}91*10^{-24}$. C'est à nouveau du côté de P1 que les liaisons sont le plus marquées, mais c'est ici la répulsion entre P1 et sa non-expression qui apporte la contribution la plus forte au khi2 (38.9%) devant celle de l'attraction entre P1 et l'expression (comme dans *Quinze Joyes*).

Manieres de langage

	Sp exprimé		Sp non exprimé		Total
P1	336	*293.13*	16	*58.87*	352
	6.4	+	***31.5***	-	
P3	137	*179.87*	79	*36.13*	216
	10.4	-	***51.7***	+	
Total		473			568

Tableau 39 : Effectifs réels et théoriques et contribution au khi2 (98.6) dans *Manieres*

Le khi2 est de 98.6 ; la probabilité qui lui est associée est de $3.1*10^{-23}$. Il s'agit de chiffres assez voisins de ceux de *Joinville*, mais la configuration est néanmoins différente, et inédite : c'est l'attraction entre P3 et la non-expression qui contribue le plus au khi2 (51.7%), suivie de la répulsion entre cette même non-expression et P1.

Ami et Amile

	Sp exprimé		Sp non exprimé		Total
P1	112	*58.6*	200	*253.4*	312
	61	+	***14***	-	
P3	123	*176.4*	816	*762.6*	939
	20.2	-	***4.7***	+	
Total	235		1016		1251

Tableau 40 : Effectifs réels et théoriques et contribution au khi2 (79.8) dans *Amile*

Le khi2 est de 79.8 ; la probabilité qui lui est associée est de $4.17*10^{-19}$. On retrouve une configuration plus « classique » : une forte attraction entre P1 et l'expression (contribution de 61% au khi2), et une répulsion entre l'expression et P3.

Aucassin et Nicolete

	Sp exprimé		Sp non exprimé		Total
P1	65	*31.5*	39	*72.5*	104
	57	+	***24.7***	-	
P3	107	*140.5*	357	*323.5*	464
	12.8	-	***5.5***	+	
Total	172		396		568

Tableau 41 : Effectifs réels et théoriques et contribution au khi2 (62.6) dans *Aucassin*

Le khi2 est de 6.6 ; la probabilité qui lui est associée est de $2,54*10^{-15}$. A nouveau l'attraction entre expression et P1 apporte la plus forte contribution au khi2 (57%), suivie de la répulsion entre non-expression et P1.

Enéas

	Sp exprimé		Sp non exprimé		Total
P1	95	*57.6*	241	*278.4*	336
	69.2	+	***14.3***	-	
P3	254	*291.4*	1446	*1408.6*	1700
	13.7	-	***2.8***	+	
Total	349		1687		2036

Tableau 42: Effectifs réels et théoriques et contribution au khi2 (35.1) dans *Eneas*

Le khi2 est de 35.1 ; la probabilité qui lui est associée reste largement inférieure à 0.001 ($3.11*10^{-9}$). La configuration est très proche de *Aucassin*, tant en termes d'attraction (P1 et l'expression) que de répulsion (P1 et la non-expression), et les contributions respectives au khi2 sont assez voisines (69.2% et 14.3%).

***Miracles* de Gautier de Coinci**

	Sp exprimé		Sp non exprimé		Total
P1	46	*24.9*	110	*131.1*	156
	67.6	+	***12.9***	-	
P3	82	*103.1*	563	*541.9*	645
	16.4	-	***3.1***	+	
	128		673		801
Total					

Tableau 43: Effectifs réels et théoriques et contribution au khi2 (26.3) dans *Miracles*

Le khi2 est de 26.3 ; la probabilité qui lui est associée est de $2.89*10^{-7}$.
On retrouve une situation proche de celle d'*Amile*, entre autres, à savoir une forte attraction entre P1 et l'expression (contribution de 67.6% au khi2), la seconde contribution étant fournie assez loin derrière, par la répulsion entre P3 et l'expression.

Estoire de Griseldis en rimes et par personnages

	Sp exprimé		Sp non exprimé		Total
P1	131	*114.1*	244	*260.9*	375
	17.7	+	***7.7***	-	
P3	22	*38.9*	106	*89.1*	128
	51.9	-	***22.7***	+	
Total	153		350		503

Tableau 44 : Effectifs réels et théoriques et contribution au khi2 (14.2) dans *Griseldis*

Le khi2 est de 14.2, plus de 20 fois moins élevé que celui de *Queste* ; la probabilité qui lui est associée est de 0,00016 : la répartition observée a donc 0.016% de chances d'être le résultat d'une distribution aléatoire.

C'est le seul texte dans lequel les contributions les plus fortes au khi2 n'impliquent pas une liaison entre P1 et la variable expression/non-expression. C'est ici la répulsion entre P3 et son expression qui apporte la plus forte contribution au khi2 (51.9%) suivie de l'attraction entre P3 et la non-expression.

De ces différents calculs on retiendra, d'une part le caractère toujours très surprenant (en termes statistiques) de la répartition de l'expression (ou non) de Sp et des personnes P1 et P3 ; on retiendra d'autre part que c'est la liaison entre P1 et la variable expression qui contribue presque toujours le plus fortement au khi2, le plus souvent par l'attraction entre P1 et son expression.
Plus précisément, on peut dégager deux configurations régulières. La plus fréquente consiste en la conjugaison d'une attraction entre P1 et son expression, et d'une répulsion entre P1 et sa non-expression (*Graal*, *Quinze Joyes*, *Froissart*, *Clari*, *Joinville*, *Aucassin* et *Eneas*). La seconde associe une forte attraction entre P1 et son expression, et une répulsion, généralement moindre, entre P3 et cette même expression (*Roland*, *Beroul*, *Amile* et *Miracles*). Deux textes enfin présentent une configuration différente : dans *Manieres*, c'est l'attraction entre P3 et sa non-expression qui contribue le plus au khi2, suivie de la répulsion entre P1 et cette même non-expression. Dans *Griseldis* enfin, c'est la répulsion entre P3 et l'expression qui est la plus forte, suivie de l'attraction entre P3 et la non-expression : P1 n'est impliqué que dans une moindre mesure.
Il est difficile de dégager des caractéristiques communes entre les textes qui présentent des configurations voisines, si ce n'est que la configuration la plus fréquente se retrouve dans les trois textes historiques (*Clari*, *Joinville*, et *Froissart*).

A bien des égards, on est loin de ce que nous avons pu observer pour la position de P1 et P3. Non seulement c'était une répartition homogène qui prévalait, mais les quelques cas de distribution marquée correspondaient en outre à des configurations assez différentes (attraction/répulsion de P1 ou de P3 pour la position postverbale).
On notera aussi que, pour l'évaluation de la répartition personne/expression, la valeur de khi2 la plus faible est tout juste inférieure à la valeur la plus élevée du khi2 rendant compte de la distribution personne/position : 14.2 (*Griseldis*) contre 21.3 (*Queste*).

Voici ci-dessous, pour l'ensemble des textes, une récapitulation des valeurs de khi2 et des probabilités qui leur sont associées. Les valeurs étant toutes significatives, hormis celles de *Coustumes*, aucune n'est surlignée.

Comme pour l'inversion, les données sont ordonnées, dans des tableaux successifs, en fonction des différents critères : chronologie, dialecte, forme et domaine.

Il ne ressort apparemment pas d'affinités saillantes entre les textes présentant des valeurs de khi2 proches, si ce n'est pour les trois textes historiques (voir tableau 45.4, khi2 autour de 100), qui présentent en outre une configuration voisine (voir ci-dessus).

Texte + date	**Dialecte**	**Forme**	**Domaine**	**khi2**	**Probabilité**
Roland (1100)	anglo-normand	vers	littéraire	245.4	$2.54*10^{-55}$
Eneas (1155)	normand	vers	littéraire	35.1	$3.11*10^{-9}$
Beroul (1165-1200)	traits normands	vers	littéraire	159.3	$1.64*10^{-36}$
Ami Amile (1200)	non marqué	vers	littéraire	79.8	$4.17*10^{-19}$
Clari (1205)	picard	prose	historique	104.5	$1.54*10^{-24}$
Aucassin (début 13ème)	traits picards	mixte	littéraire	62.6	$2.54*10^{-15}$
Miracles (1220)	non marqué	vers	religieux	26.3	$2.88*10^{-7}$
Queste (1230)	non marqué	prose	littéraire	296	$2.4*10^{-66}$
Coustumes (1283)	traits picards	prose	juridique	0.001	0.97
Joinville (1307)	non marqué	prose	historique	104.1	$1.91*10^{-24}$
Froissart (1369-1400)	franco-picard	prose	historique	121.1	$3.69*10^{-28}$
Griseldis (1395)	traits picards	vers	littéraire	14.2	0.00016
Manieres (1396-1399)	non marqué	mixte	didactique	98.6	$3.1*10^{-23}$
Quinze Joyes (1400)	non marqué	prose	littéraire	193.8	$4.6*10^{-44}$

Tableau 45.1. : Distribution de l'expression et de la non-expression de P1 et P3 : valeur du Khi2 et probabilité associée. Présentation selon le critère chronologique.

Texte + date	**Dialecte**	**Forme**	**Domaine**	**khi2**	**Probabilité**
Roland (1100)	**anglo-normand**	vers	littéraire	245.4	$2.54*10^{-55}$
Eneas (1155)	**normand**	vers	littéraire	35.1	$3.11*10^{-9}$
Beroul (1165-1200)	**traits normands**	vers	littéraire	159.3	$1.64*10^{-36}$
Clari (1205)	**picard**	prose	historique	104.5	$1.54*10^{-24}$
Aucassin (début 13ème)	**traits picards**	mixte	littéraire	62.6	$2.54*10^{-15}$
Coustumes (1283)	**traits picards**	prose	juridique	0.001	0.97
Froissart (1369-1400)	**franco-picard**	prose	historique	121.11	$3.69*10^{-28}$
Griseldis (1395)	**traits picards**	vers	littéraire	14.2	0.00016
Ami Amile (1200)	**non marqué**	vers	littéraire	79.8	$4.17*10^{-19}$
Miracles (1220)	**non marqué**	vers	religieux	26.3	$2.88*10^{-7}$
Queste (1230)	**non marqué**	prose	littéraire	296	$2.4*10^{-66}$
Joinville (1307)	**non marqué**	prose	historique	104.1	$1.91*10^{-24}$
Manieres (1396-1399)	**non marqué**	mixte	didactique	98.6	$3.1*10^{-23}$
Quinze Joyes (1400)	**non marqué**	prose	littéraire	193.8	$4.6*10^{-44}$

Tableau 45.2. Distribution de l'expression et de la non-expression de P1 et P3 : valeur du Khi2 et probabilité associée. Présentation selon le critère dialectal.

Texte+date	**Dialecte**	**Forme**	**Domaine**	**khi2**	**Probabilité**
Roland (1100)	anglo-normand	**vers**	littéraire	245.4	$2.54*10^{-55}$
Eneas (1155)	normand	**vers**	littéraire	35.1	$3.11*10^{-9}$
Beroul (1165-1200)	traits normands	**vers**	littéraire	159.3	$1.64*10^{-36}$
Ami Amile (1200)	non marqué	**vers**	littéraire	79.8	$4.17*10^{-19}$
Miracles (1220)	non marqué	**vers**	religieux	26.3	$2.88*10^{-7}$
Griseldis (1395)	traits picards	**vers**	littéraire	14.2	0.00016
Aucassin (début 13ème)	traits picards	**mixte**	littéraire	62.6	$2.54*10^{-15}$
Manieres (1396-1399)	non marqué	**mixte**	didactique	98.6	$3.1*10^{-23}$
Clari (1205)	picard	**prose**	historique	104.5	$1.54*10^{-24}$
Queste (1230)	non marqué	**prose**	littéraire	296	$2.4*10^{-66}$
Coustumes (1283)	traits picards	**prose**	juridique	0.001	0.97
Joinville (1307)	non marqué	**prose**	historique	104.1	$1.91*10^{-24}$
Froissart (1369-1400)	franco-picard	**prose**	historique	121.1	$3.69*10^{-28}$
Quinze Joyes (1400)	non marqué	**prose**	littéraire	193.8	$4.6*10^{-44}$

Tableau 45.3. : Distribution de l'expression et de la non-expression de P1 et P3 : valeur du Khi2 et probabilité associée. Présentation selon le critère de la forme.

Texte + date	Dialecte	Forme	Domaine	khi2	Probabilité
Roland (1100)	anglo-normand	vers	**littéraire**	245.4	$2.54*10^{-55}$
Eneas (1155)	normand	vers	**littéraire**	35.1	$3.11*10^{-9}$
Beroul (1165-1200)	traits normands	vers	**littéraire**	159.3	$1.64*10^{-36}$
Ami Amile (1200)	non marqué	vers	**littéraire**	79.8	$4.17*10^{-19}$
Aucassin (début 13ème)	traits picards	mixte	**littéraire**	62.6	$2.54*10^{-15}$
Queste (1230)	non marqué	prose	**littéraire**	296	$2.4*10^{-66}$
Griseldis (1395)	traits picards	vers	**littéraire**	14.2	0.00016
Quinze Joyes (1400)	non marqué	prose	**littéraire**	193.8	$4,6*10^{-44}$
Clari (1205)	picard	prose	**historique**	104.5	$1.54*10^{-24}$
Joinville (1307)	non marqué	prose	**historique**	104.1	$1,91*10^{-24}$
Froissart (1369-1400)	franco-picard	prose	**historique**	121.1	$3,69*10^{-28}$
Miracles (1220)	non marqué	vers	**religieux**	26.3	$2.88*10^{-7}$
Coustumes (1283)	traits picards	prose	**juridique**	0.001	0.97
Manieres (1396-1399)	non marqué	mixte	**didactique**	98.6	$3.1*10^{-23}$

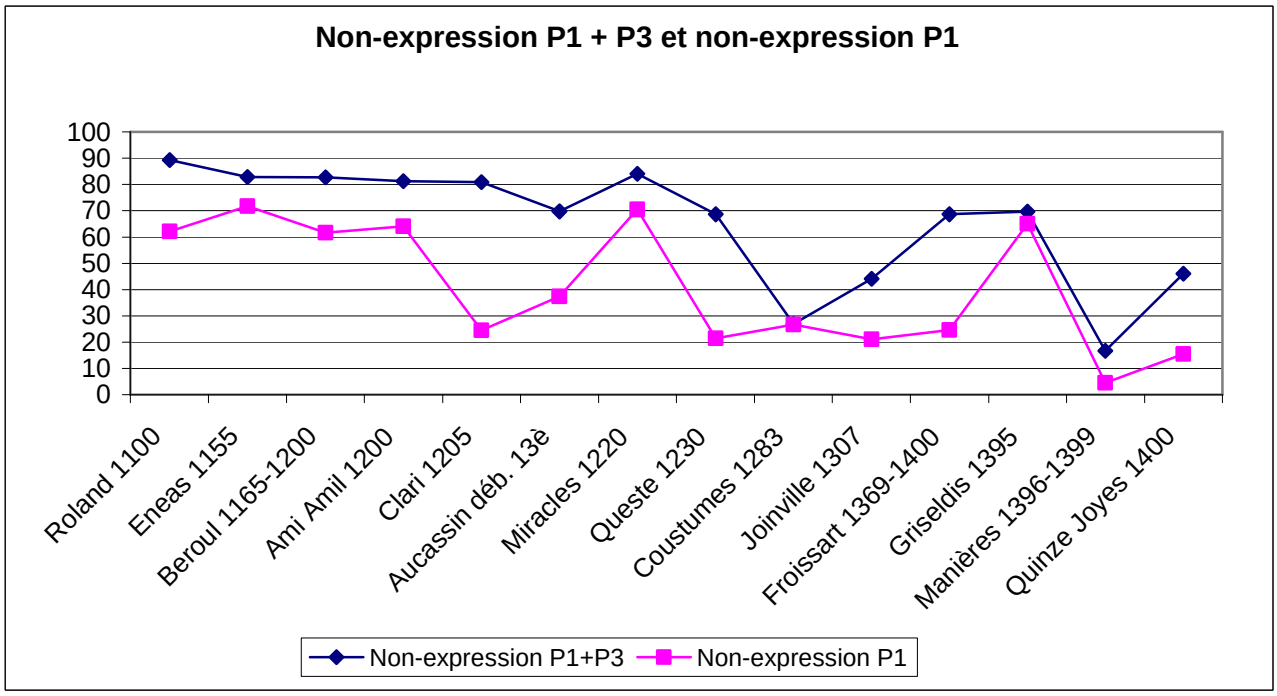
Tableau 45.4. : Distribution de l'expression et de la non-expression de P1 et P3 : valeur du Khi2 et probabilité associée. Présentation selon le critère du domaine.

4.4.3. Représentations graphiques : lignes et barres

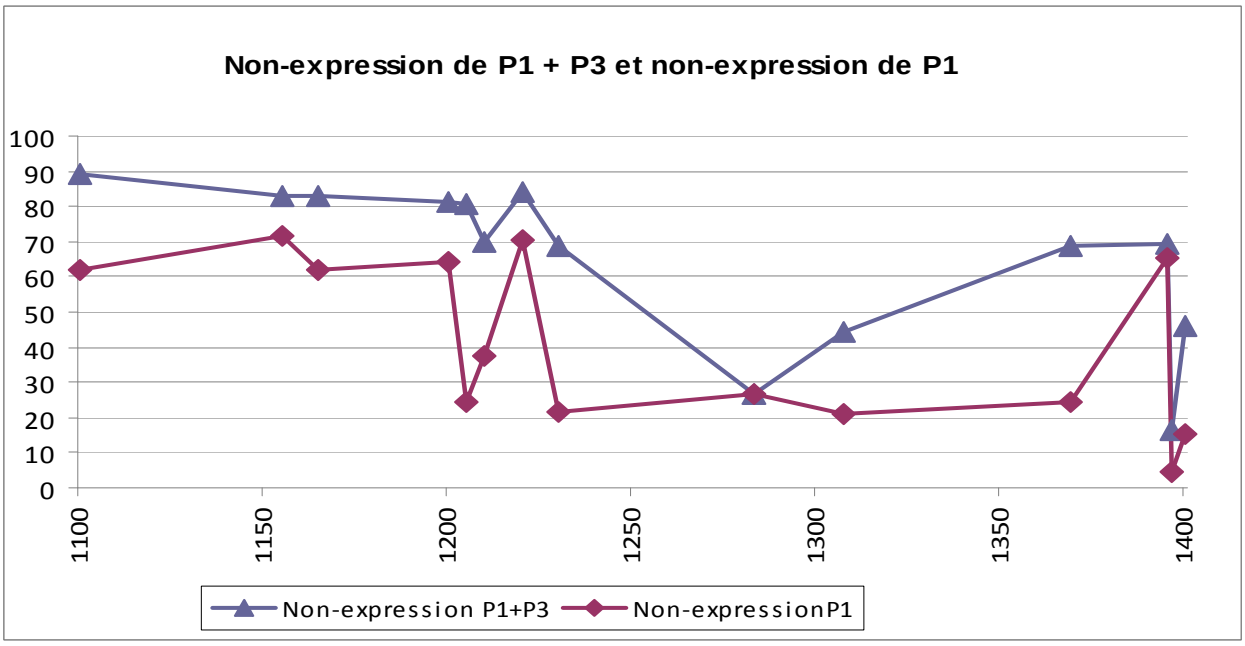
Comme pour l'inversion, voici ci-dessous différents modes de visualisation des fréquences de non-expression de Sp dans les textes. Les fréquences de non-expression de P1 et de P3 sont successivement mises en perspective avec celle de P1 + P3 (graphiques 5 et 6), puis elles sont réunies dans le graphique 7. Pour chacun de ces

trois graphiques il est proposé une double représentation, la première sans axe chronologique proportionné, la seconde avec un tel axe.

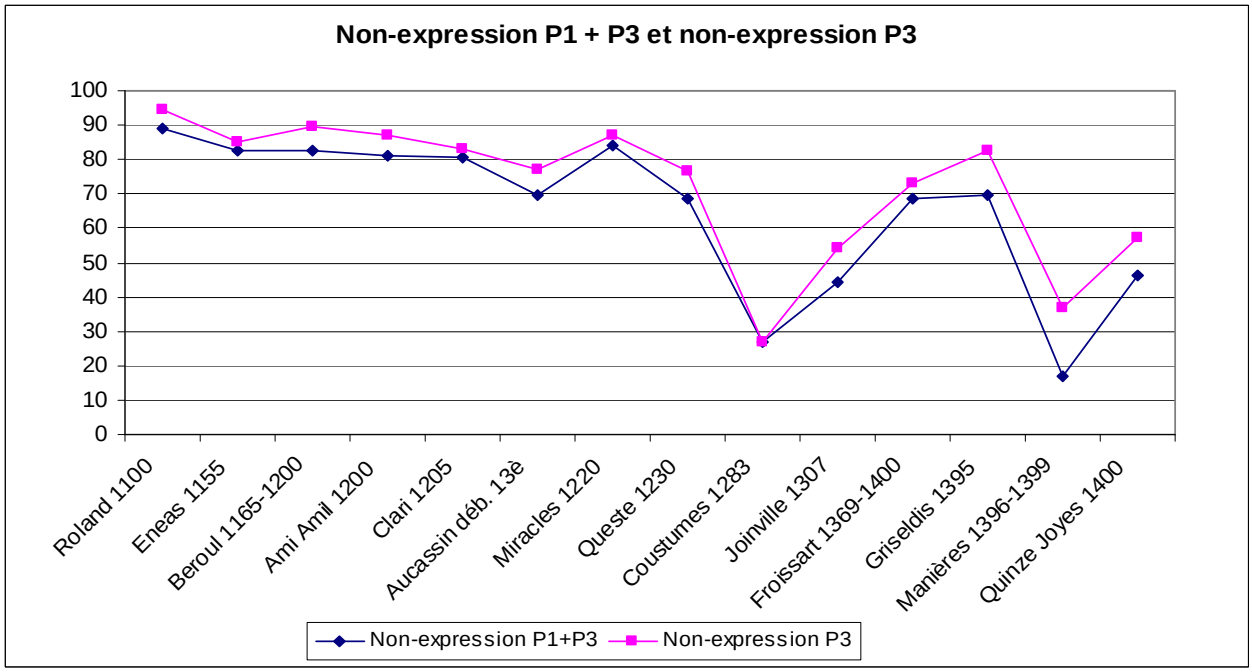
Le graphique 8 propose une représentation en barres des fréquences de non-expression de P3+P1, P1 et P3.



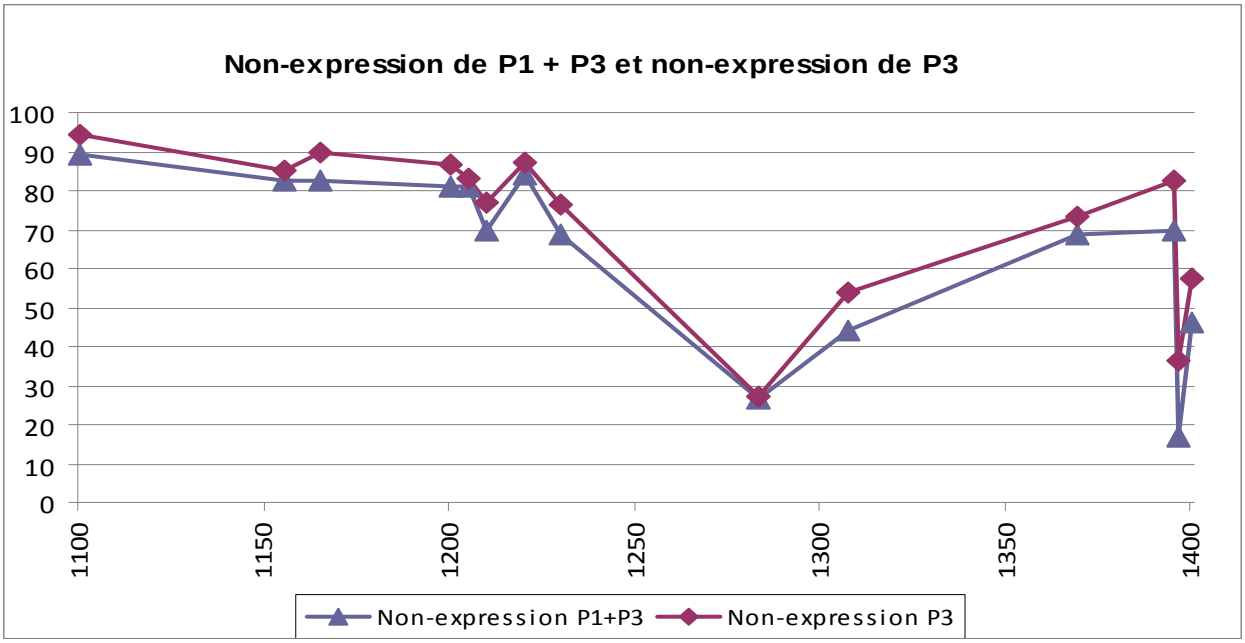
Graphique 5a : non-expression de P1 + P3 et non-expression de P1 (axe des ordonnées : pourcentages)



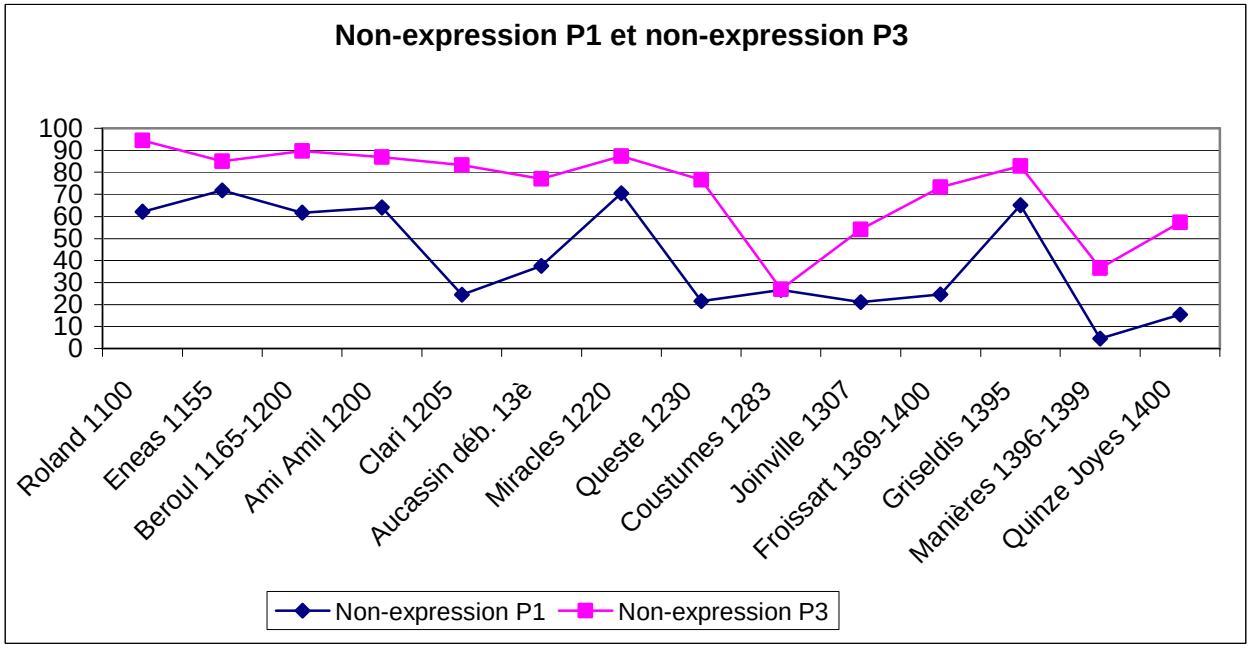
Graphique 5b : non-expression de P1 + P3 et non-expression de P1. Chronologie stricte.



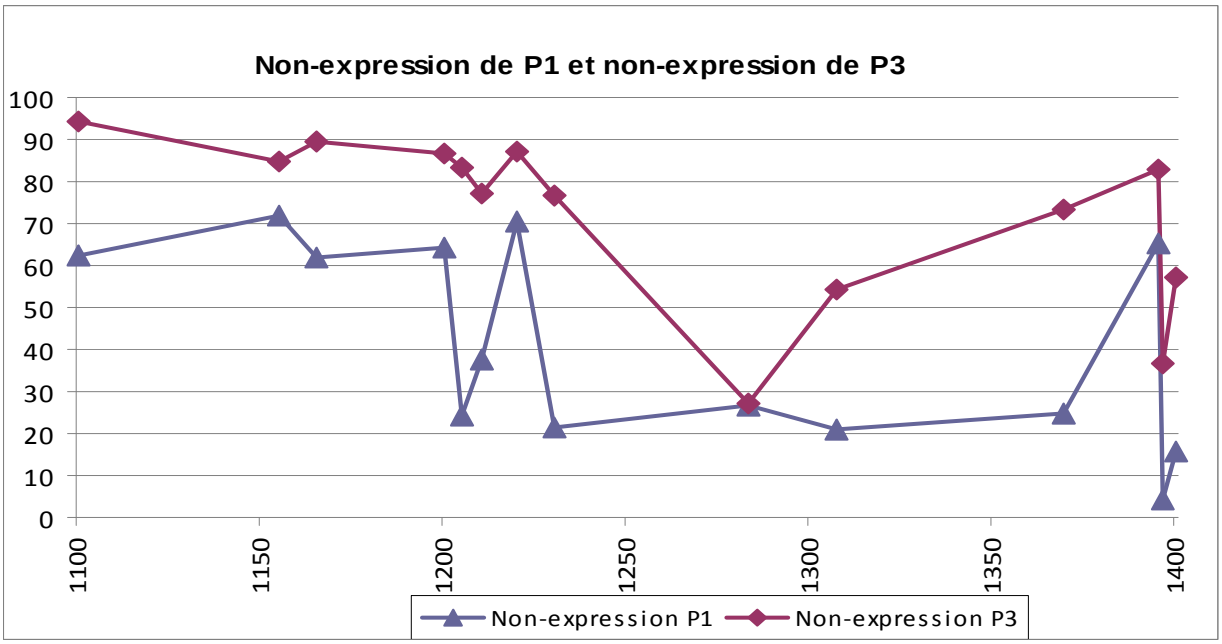
Graphique 6a : non-expression de P1 + P3 et non-expression de P1



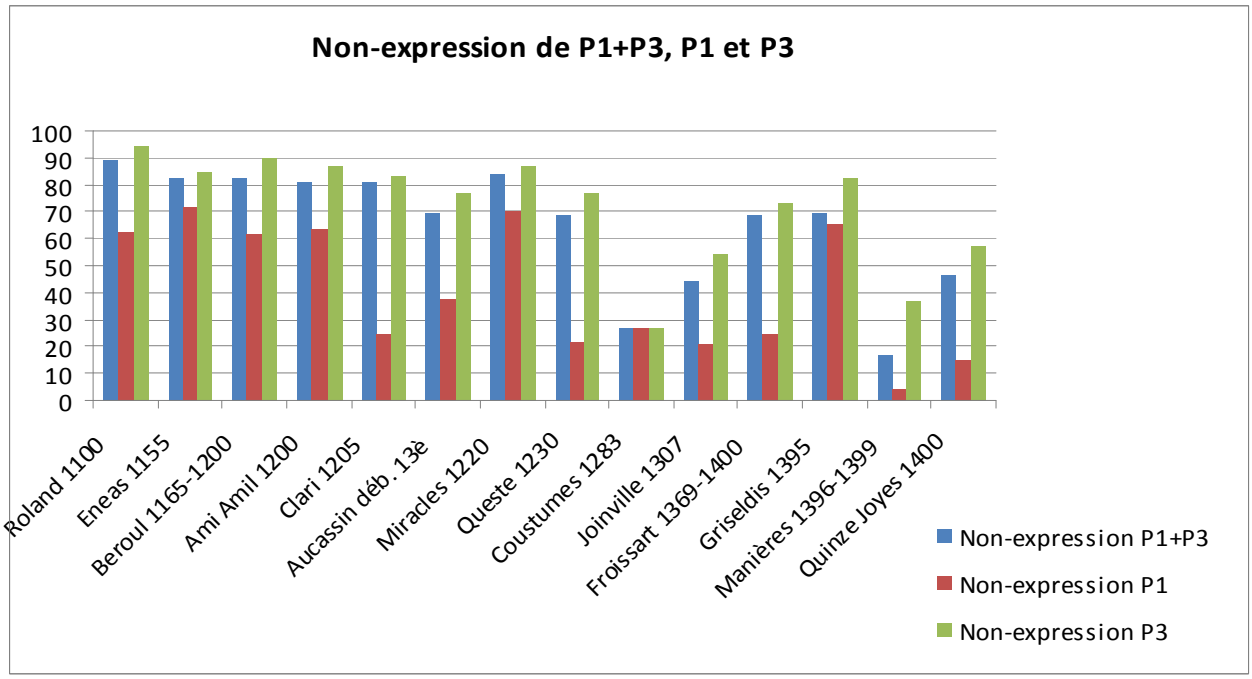
Graphique 6b: non-expression de P1 + P3 et non-expression de P1 . Chronologie stricte



Graphique 7a : non-expression de P1 et non-expression de P3



Graphique 7b : non-expression de P1 et non-expression de P3. Chronologie stricte.



Graphique 8 : Fréquences de non-expression de P1+P3, P1 et P3

La comparaison des graphiques 5 (P1+P3 et P1) et 6 (P1 + P3 et P3) met nettement en relief la grande proximité de la courbe de P3 avec celle de la moyenne P1 + P3, phénomène que nous avions déjà noté en considérant le tableau des fréquences, et qui s'explique par la forte proportion de P3 sur l'ensemble des Sp non-exprimés. Par ailleurs, corollaire du fait que la fréquence de P1 est toujours inférieure à celle de P3, il n'y a pas de croisement des courbes tels que ceux que l'on a pu observer pour l'inversion. La courbe de P1 apparaît cependant légèrement plus accidentée que celle de P3 (voir graphiques 7).

Bien qu'il ne se dégage pas d'évolution chronologique continue, on observe néanmoins des tendances plus nettes que pour l'inversion.

4.4.4. Essai d'interprétation des tableaux et des représentations graphiques

Si l'on considère tout d'abord le **paramètre chronologique**, on constate que, contrairement à ce que l'on a pu observer pour l'inversion, le début de la période présente une relative stabilité, au moins en ce qui concerne la courbe de P3 (et celle de P1 + P3). De *Roland* à *Queste*, c'est à dire du début du 12ème siècle au 1er tiers du 13ème, la non-expression de P3 dépasse 70%. C'est en revanche moins net pour P1. Relativement stable (entre 60% et 70%) pour *Roland*, *Eneas* et *Beroul*, c'est-à-dire les

3 textes les plus anciens, mais aussi pour *Amile* (1200), la fréquence de la non-expression de P1 connaît une chute vertigineuse dans *Clari*, texte qui s'est révélé très atypique en ce qui concerne l'inversion (très élevée pour P3 et nulle pour P1). La non-expression de P1 reste faible dans *Aucassin*, remonte dans *Miracles*, mais connaît une nouvelle chute dans *Queste*. On notera que *Aucassin* et *Queste*, dans lesquels la non-expression de P1 est donc nettement en deçà de celle de P3, sont deux textes qui présentent à l'inverse une fréquence d'inversion de P1 supérieure à celle de P3[138].

Les choses sont globalement plus confuses après *Queste*. *Coustumes* accuse une chute généralisée des fréquences, avec en outre une parfaite concordance (quantitative au moins) entre la non-expression de P1 et celle de P3. La non-expression repart à la hausse dans les 3 textes suivants, *Joinville* au début du 14ème siècle, *Froissart* dans la seconde moitié, et *Griseldis* à la fin du siècle, ce texte renouant avec des fréquences proches de celles des textes les plus anciens (plus de 80% pour la non-expression de P3, et 65% pour celle de P1). Les fréquences connaissent une nouvelle chute dans *Manieres*, particulièrement en ce qui concerne P1 (4.5%), suivie d'une légère remontée dans *Quinze Joyes*, qui renoue avec des fréquences proches de celles de *Joinville*. Alors que *Manieres* et *Quinze Joyes* présentent de fortes affinités en ce qui concerne l'inversion de Sp, les deux textes divergent assez nettement au regard de sa non-expression.

Si l'on considère maintenant les textes dans lesquels les fréquences de non-expression de P1 et de P3 sont les plus divergentes, il ne se dessine rien de net non plus du point de vue chronologique : ce sont en effet dans *Clari*, *Aucassin*, *Queste*, *Froissart* et *Quinze Joyes* que les écarts sont les plus marqués, textes qui s'étalent du début du 13ème au début du 15ème siècle. A l'inverse, les textes qui présentent les fréquences de non-expression de P1 et de P3 les plus proches sont *Eneas*, *Miracles*, *Coustumes* et *Griseldis*, pareillement distants sur l'axe temporel.

A défaut de la mise au jour de tendances évolutives bien nettes, on peut tenter d'opérer des regroupements entre textes, comme nous l'avons fait pour l'inversion, en nous fondant à la fois sur la valeur des fréquences (plus ou moins élevées) et sur le rapport entre les fréquences entre P1 et P3 (fréquences proches ou éloignées). De ce double point de vue, on peut rapprocher six textes. Il s'agit de *Roland*, *Eneas*, *Beroul*

[138] Mais, à l'inverse, elle est nulle dans *Clari*.

et *Amile*, mais aussi *Miracles* et *Griseldis*, qui ont tous des fréquences de non-expression qui oscillent entre 65% et plus de 80%. Les quatre premiers, s'ils ne sont pas véritablement concentrés dans le temps (*Roland* est séparé d'*Amile* par un siècle) appartiennent en tout cas à la période la plus ancienne de ce corpus. *Miracles* n'est guère distant d'*Amile*, mais *Griseldis* est en revanche bien plus tardif. Les quatre textes les plus anciens ont en commun de présenter des traits normands ou anglo-normands (ou d'être non-marqués du point de vue dialectal : *Amile*), d'être en vers et de relever du domaine littéraire (genre romanesque ou épique). Mais il est difficile, en l'absence d'autres textes chronologiquement voisins mais présentant des caractéristiques externes différentes, de déterminer si c'est leur proximité temporelle qui joue, ou bien les autres facteurs (si tant est que les uns ou les autres jouent un rôle...). Quant aux deux autres textes, *Miracles* et *Griseldis*, rien ne les rapproche entre eux, ni des textes précédents, hormis le fait d'être en vers, ce qui, il est vrai, les distingue des autres textes des 13ème et 14ème siècle, tous en prose (hormis *Aucassin* et *Manieres* qui sont mixtes). *Griseldis* par ailleurs relève du domaine littéraire, mais le genre dramatique auquel il appartient est si différent des genres romanesques et épiques qu'il serait hasardeux d'opérer un rapprochement sur ce critère.

Quatre autres textes présentent par ailleurs des fréquences assez voisines, aussi bien pour P1 que pour P3 : il s'agit de *Clari*, *Aucassin*, *Queste*, et de *Froissart*, caractérisés par le fait que la non-expression de P1 est nettement inférieure à celle de P3 (elle est néanmoins un peu plus élevée dans *Aucassin* que dans les autres textes). Les trois premiers textes sont proches dans le temps (étalés sur une trentaine d'années), mais tout distingue *Clari* et *Queste*, hormis d'être en prose. Quant à *Aucassin*, il partage avec *Clari* son caractère picardisant (*Clari* est nettement picard, *Aucassin* l'est moins), et avec *Queste* le fait d'appartenir au domaine littéraire. Mais leurs genres (récits brefs et roman) sont bien différents. *Froissart*, enfin, nettement plus tardif, présente en revanche de nombreux traits communs avec *Clari* : le caractère picard (qu'il partage aussi avec *Aucassin*) et l'appartenance au domaine historique.

Caractérisés par des fréquences plus basses que les textes précédents, *Quinze Joyes* et *Joinville* (dont la non-expression de P1 est légèrement plus fréquente) ont presque un siècle d'écart. Ils ne présentent pas de traits dialectaux marqués, et appartiennent à des domaines différents. Peu de points communs entre eux, donc.

Il reste enfin deux textes isolés. Il s'agit, à la fin du 14ème siècle, de *Manieres*, caractérisé par des fréquences de non-expression globalement très faibles (en

particulier pour P1 : 4.5%), et de *Coustumes*, qui se distingue par son très faible pourcentage de non-expression de P3 (27% : le plus faible du corpus) ainsi que par la coïncidence entre les fréquences de non-expression de P1 et de P3.
Redisons-le : hormis pour plusieurs textes du début de la période (mais pas tous) qui ont des configurations de non-expression de Sp assez proches, le critère chronologique n'apparaît pas, dans ce corpus, comme nettement déterminant. Les deux derniers textes de la fin de la période (*Manieres* et *Quinze Joyes*) ont des fréquences moindres que bon nombre des textes précédents : peut-être cela préfigure-t-il le début d'une baisse généralisée, mais il faudrait, pour le confirmer, considérer d'autres textes, contemporains et postérieurs.

Considérons maintenant de façon plus sytématique les autres critères. Le **critère dialectal** ne s'est pas révélé discriminant pour l'inversion de Sp. Il semble l'être un peu plus en ce qui concerne la non-expression. C'est cependant un critère qui doit être manié avec précaution, dans la mesure où ses différentes valeurs recouvrent en partie les valeurs des autres critères. En particulier, il apparaît que les textes présentant des traits anglo-normands ou normands se trouvent au 12ème, tandis que ceux à caractère plus ou moins picardisants se trouvent au 13ème, et encore au 14ème siècle[139]. A cela s'ajoute que deux des textes picardisants sont historiques, et que les deux textes « normands », non seulement sont assez proches chronologiquement, mais sont aussi tous deux des romans (*Eneas* et *Beroul*). C'est donc avec précaution qu'il faut considérer les affinités entre *Eneas* et *Beroul* d'une part, et *Clari* et *Froissart* d'autre part (pour les deux derniers, les effets de l'affinité dialectale n'est cependant pas parasitée par la proximité temporelle qui existe entre *Beroul* et *Eneas*).

Les mêmes précautions sont de mise en ce qui concerne **le domaine et le genre**, dans la mesure où les affinités observées jusqu'ici recoupent des affinités sur le plan dialectal et/ou chronologique. Pour le reste, nous avons vu que les rapprochements entre textes ne correspondaient guère à des affinités en termes de domaine et de genre (ainsi de *Joinville* et *Quinze Joyes*, par exemple). Rappelons cependant que nous avons mis au jour, dans le cadre du calcul du khi2, des affinités entre les trois textes historiques.

[139] Cette question a été évoquée en 3.1.2. (présentation du corpus). Rappelons que ces coïncidences dépassent ce corpus : au 12ème siècle, les textes anglo-normands sont assez nombreux, et la littérature picarde est florissante au 13ème siècle.

Reste enfin **le critère de la forme**. S'il est vrai que les textes les plus anciens sont en vers, et que, dans ce corpus, les textes des 13ème et 14ème siècles sont majoritairement en prose, il s'en trouve cependant deux qui sont en vers (*Miracles* et *Griseldis*), et deux qui sont mixtes (*Aucassin* et *Manieres*, la prose étant cependant prévalente dans le second). Aucune spécificité au regard de ce critère ne s'est dégagée pour l'inversion. On constate en revanche que les deux textes du début du 13ème et de la fin du 14ème qui présentent des affinités avec les textes plus anciens sont précisément ceux qui sont en vers : *Miracles* et *Griseldis*, que tout sépare par ailleurs. Deux textes ne permettent certes pas de poser une corrélation entre non-expression élevée et forme versifiée : il faudra explorer d'autres textes pour le vérifier[140].

En conclusion, on retiendra que, plus que pour l'inversion, les configurations de la non-expression de Sp dans les différents textes font apparaître quelques tendances, sur les plans chronologique, dialectal, du domaine et du genre, et de la forme. Tendances qui, par définition, n'ont rien de systématique. Disons que le désordre est moindre que pour l'inversion de Sp.

Avant de proposer une esquisse de mise en perspective de l'inversion et de la non-expression de Sp, j'aimerais ajouter quelques mots sur les descripteurs des textes. Il en est un qui mériterait d'être ajouté, mais qui se révèle lourd à mettre en place. Il s'agit de la proportion de discours (par opposition au récit) qui se trouve dans les textes. C'est un critère qui est rarement pris en compte, parce que coûteux à quantifier. De fait, à ma connaissance, l'information n'est jamais fournie pour les textes, qu'ils soient sur support numérique ou papier. Or si l'on admet l'idée que, dans un texte, le discours présente certaines affinités avec l'oral de l'époque (en dépit des tendances avérées à la correction ou au contraire à la caricature), et que l'oral et l'écrit tendent à avoir des codes en partie divergents, il n'est pas inutile de connaître la part de l'un et de l'autre dans un texte. Au-delà de la mise au jour de la proportion de discours dans un texte, il est surtout intéressant de pouvoir déterminer si les occurrences collectées appartiennent au discours ou au récit. Dans le cas de la présente étude, il est aisé de statuer sur le contexte d'occurrences de P1 (même s'il n'est pas sûr que la prise de parole du narrateur présente les mêmes caractéristiques que celle d'un personnage du récit). C'est en revanche beaucoup plus difficile pour P3, la taille du contexte de

[140] Dans les deux textes mixtes, il ne semble pas y avoir de préférence marquée de la non-expression pour les passages versifiés.

l'occurrence collectée ne permettant pas toujours de déterminer s'il s'agit de discours ou de récit. Or il n'est pas incongru d'envisager que la syntaxe de P3 n'est pas la même en discours et en récit. Bien sûr, en retournant au texte intégral, on trouve l'information, mais c'est une démarche longue et coûteuse qui ne peut être réalisée pour un large corpus. Par conséquent, l'encodage de cette information dans les textes numérisés serait fort précieuse[141].

4.5. Inversion et non expression de Sp : peut-on dégager des relations ?

Nous avons vu dans les essais de synthèse des fréquences d'inversion et de non-expression de VSp qu'il est difficile de déterminer des tendances, en particulier pour ce qui est de l'inversion. Bien sûr entre le texte le plus ancien du corpus (*Roland*) et le plus tardif (*Quinze joyes*), les fréquences d'inversion et de non-expression de Sp ont largement baissé, personnes P1 et P3 confondues[142], mais selon des pentes très accidentées, parsemées de remontées.

Voici ci-dessous un tableau qui synthétise les différentes fréquences, les textes étant présentés selon un ordre chronologique.

[141] A ma connaissance, c'est précisément un projet en cours dans le cadre du projet BFM/Corptef (ENS Lyon).
[142] Et la connaissance que nous avons des textes ultérieurs confirme cette décroissance (voir Prévost 2010 pour quelques chiffres).

Texte + Date	Dialecte	Forme	Domain	Invers. P1+P3	Non-expr. P1+P3	Invers. P1	Non-expr. P1	Invers. P3	Non-expr. P3
Roland (1100)	anglo-normand	v.	litt.	21.1	89.3	17.8	62.2	25.3	94.4
Eneas (1155)	normand	v.	litt.	9.7	82.8	12.6	71.7	8.7	85
Béroul (1165-1200)	traits normands	v.	litt.	17.5	82.7	16.1	61.7	19.3	89.7
Ami Amile (1200)	non marqué	v.	litt.	23.8	81.5	25.9	64.1	21.8	87.3
Clari (1205)	picard	p.	histo.	42.6	80.8	0	25	51.1	83.3
Aucassin (début 13ème)	traits picards	m.	litt.	9.9	69.7	16.9	37.5	5.6	77
Miracles (1220)	non marqué	v.	relig.	21.9	84	19.6	70.5	23.2	87.3
Queste (1230)	non marqué	p.	litt.	22.1	68.6	33.1	21.5	15.9	76.5
Coustumes (1283)	traits picards	p.	jurid.	15.8	26.9	18.2	26.7	15.4	27
Joinville (1307)	non marqué	p.	histo.	12.3	44.1	15	21.1	10.2	54.1
Froissart (1369-1400)	franco-picard	p.	histo.	8.8	68.6	5.4	24.6	9.9	73.2
Griseldis (1395)	traits picards	v.	litt.	13.7	69.6	11.4	65.1	27.3	82.8
Manières (1396-1399)	non marqué	m.	didact.	4.2	16.7	3	4.5	7.3	36.6
Quinze Joyes (1400)	non marqué	p.	litt.	5.7	46.1	3.5	15.5	7.3	57.3

Tableau 46 : présentation synthétiquedes fréquences d'inversion et de non-expresion de P1+P3, P1 et P3.

Le tableau 46 ne met pas au jour de convergences régulières entre les fréquences d'inversion et de non-expression.

Ainsi les quatre textes (*Beroul, Eneas, Amile et Clari*) qui ont une fréquence d'inversion P1+P3 qui avoisine 80% présentent des fréquences d'inversion très différentes : 9.7% pour *Eneas*, 17.7% pour *Beroul*, 23.8% pour *Amile* et 42.6% pour *Clari*. De même les quatre textes dont la moyenne de non-expression de P1 + P3 tournent autour de 69% varient quant à la fréquence d'inversion de P1+P3 : 9.9% pour *Aucassin*, 22.1% pour *Queste*, 8.8% pour *Froissart*, et 13.7% pour *Griseldis*. On note

cependant des affinités entre *Froissart* et *Aucassin*, deux textes que tout sépare (date, domaine, genre et forme) sauf la présence de traits picards.

Les fréquences liées à P3 présentent des affinités un peu plus marquées : *Amile* et *Miracles*, dont la fréquence de non-expression de P3 est de 87.3%, tendent à inverser P3 dans des proportions similaires, respectivement 21.8% et 23.2%. De ces deux textes on peut rapprocher *Beroul*, dont les fréquences de non-expression et d'inversion sont de 89.7% et de 19.3%. Il s'agit de trois textes qui se situent au tournant des 12ème et 13ème siècles, mais qui n'ont sinon que peu de points communs. Les autres textes présentant une fréquence de non-expression de P3 élevée (entre 83% et 85%) ont en revanche des tendances à inverser P3 très variables : 8.7% pour *Eneas*, 51.1% pour *Clari* et 27.3% pour *Griseldis*. Si l'on considère à l'inverse les deux textes dans lesquels la non-expression de P3 est la plus faible (*Coustumes* : 27% et *Manieres* : 36.6%), on constate de mêmes divergences en matière d'inversion (respectivement 15.4% et 7.3%).

C'est cependant du côté de P1 que l'absence de convergences entre les deux fréquences est la plus frappante. Parmi les quatre textes qui offrent une fréquence de non-expression oscillant entre 61 % et 65%, deux ont des fréquences d'inversion assez proches (*Roland* : 17.8%, *Beroul* : 16.1%), mais dans les deux autres la proportion d'inversion de P1 est au contraire très différente : 25.9% dans *Amile*, et 11.4% dans *Griseldis*. Notons que le texte dans lequel la non-expression est la plus fréquente (*Eneas* : 71.7%) est aussi parmi ceux qui offrent une fréquence d'inversion peu élevée (12.6% dans *Eneas*). Si l'on considère les fréquences d'inversion, on constate que les quatre textes les plus tardifs, qui offrent les chiffres les plus bas (de 3% à 11.4%), présentent en revanche des fréquences d'expression excessivement variables : 24.6% pour *Froissart*, 65.1% pour *Griseldis*, 4.5% pour *Manieres*, et 15.5% pour *Quinze Joyes*. On pourrait poursuivre ce type de rapprochements et souligner la difficulté à déceler des convergences autres que ponctuelles. Signalons pour finir que le texte qui présente la fréquence d'inversion de P1 la plus forte, et largement en tête des autres (33.1%), est aussi celui dont la fréquence de non-expression de P1 est relativement basse : 21.5%. *Queste* est d'autant plus spécifique à cet égard que c'est le seul texte du corpus dans lequel la fréquence d'inversion de P1 est supérieure à celle de sa non-expression. Dans les autres textes, l'inversion de Sp, quelle que soit la personne, est toujours plus rare, souvent de loin, à sa non-expression. Rappelons que *Queste* est le

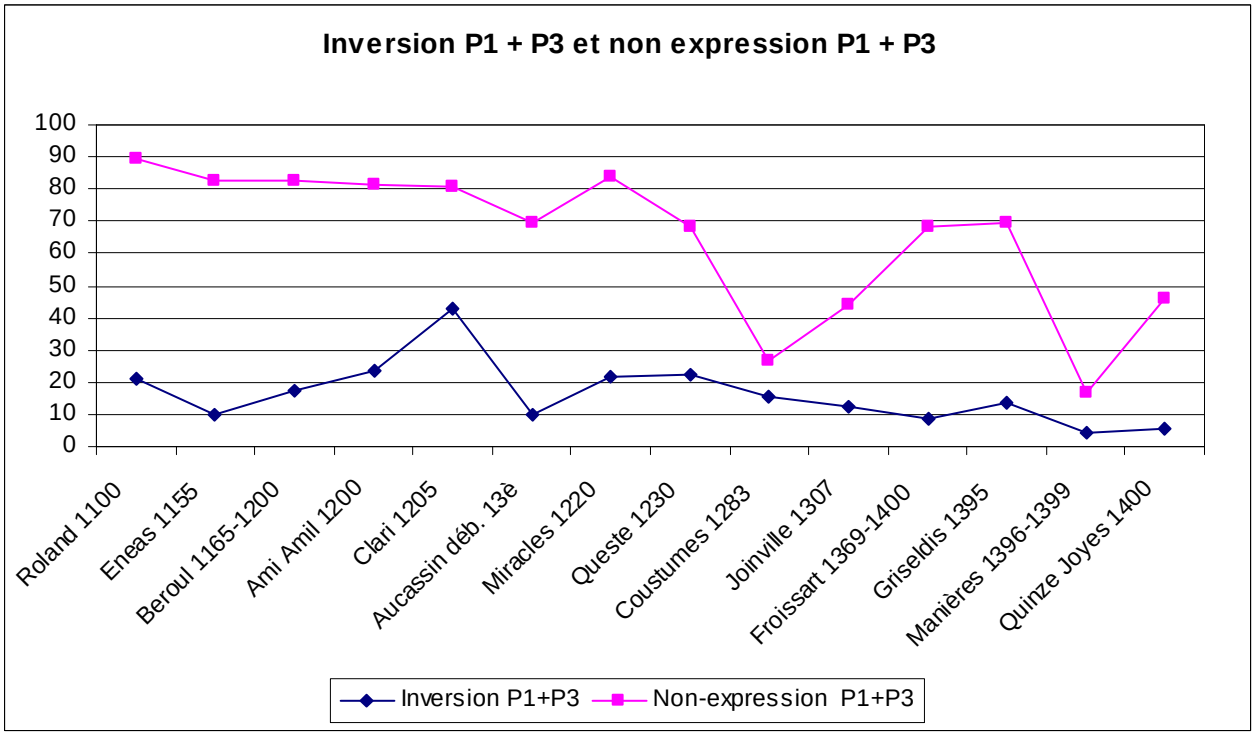
texte qui offre les khi2 liés à la position et à l'expression de Sp les plus élevés : les calculs ont montré une attraction très forte entre P1 et la position postverbale d'une part, et entre P1 et l'expression d'autre part.

Les graphiques présentés ci-dessous mettent en perspective inversion et non-expression pour chacune des personnes, puis pour P1+P3. Les modalités de présentation sont les mêmes que celles adoptées en 4.3. et 4.4., à savoir un graphique en lignes sans axe des abscisses strictement proportionné (9a, 10a et 11a), puis le même graphique avec une représentation chronologique stricte (9b, 10b et 11b), et enfin un graphique en barres (9c, 10c et 11c).

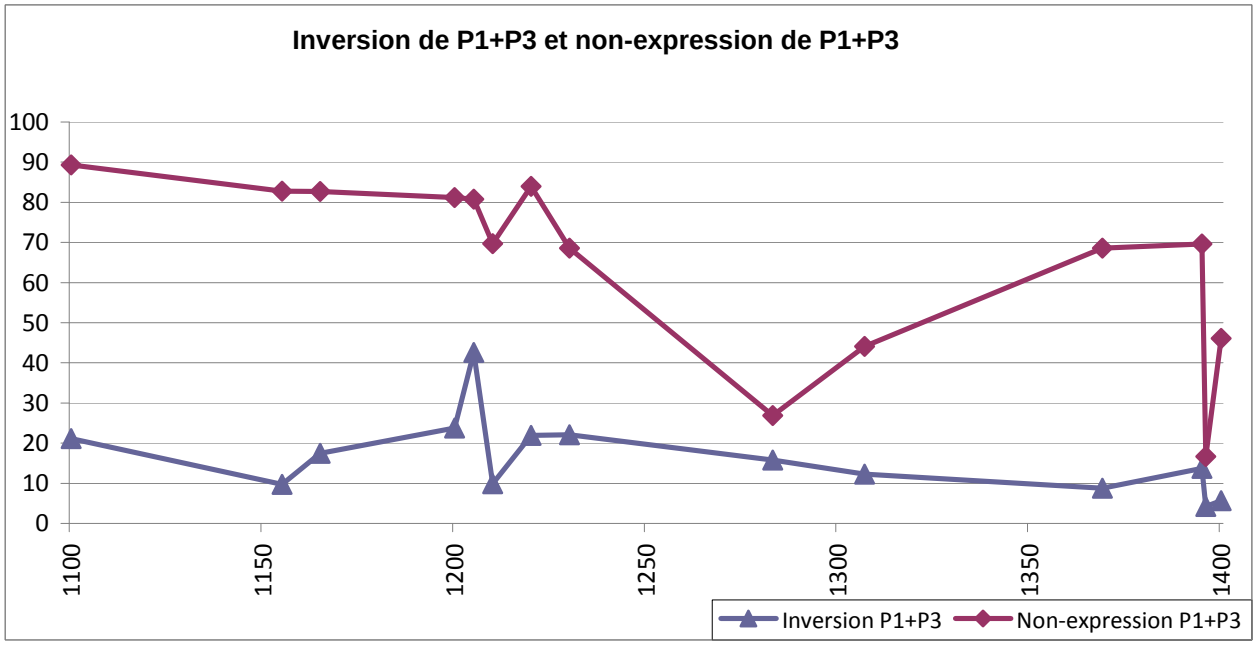
Inversion et non-expression de P1+P3 :

Comme nous l'avons vu dans le cadre des synthèses sur l'inversion et la non-expression, la progression des fréquences d'inversion et de non-expression de P1+P3 est très voisine de celle de P3, ce qui s'explique par la fréquence généralement beaucoup plus importante des effectifs de P3.

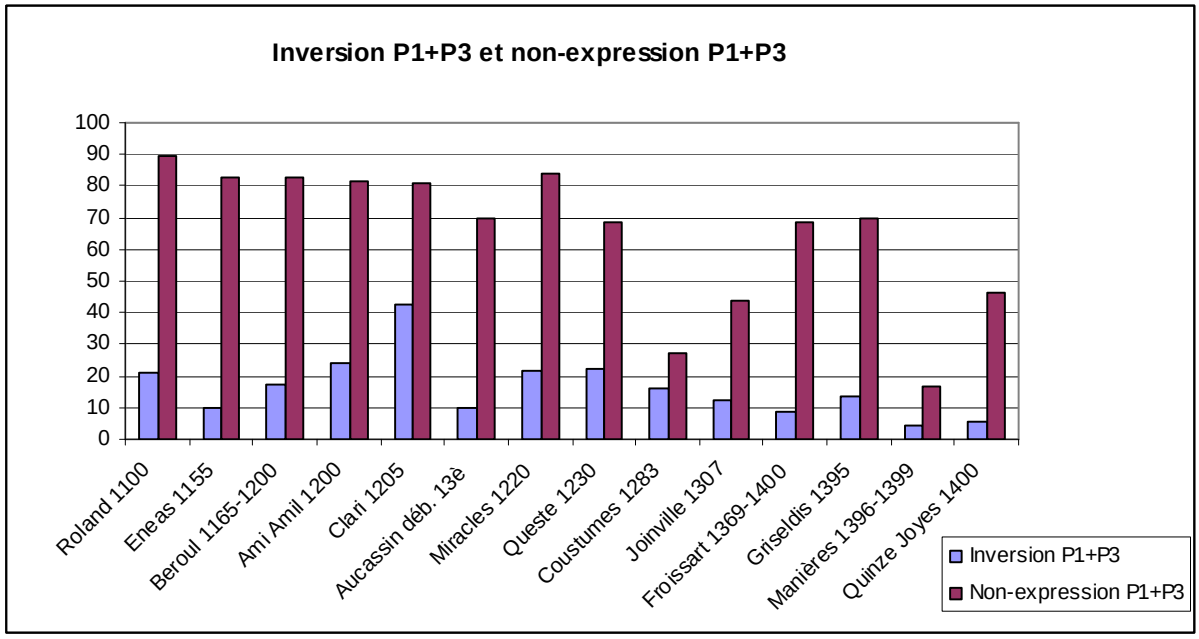
C'est donc plus spécifiquement aux graphiques de P1 et de P3 que je m'attacherai. On notera simplement, en ce qui concerne P1 et P3, un relatif « parallélisme » entre les deux lignes, jusque la fin du 12ème siècle, c'est-à-dire en ce qui concerne les quatre premiers textes du corpus, *Eneas* provoquant malgré tout une petit chute du côté de l'inversion. Signalons aussi deux « pics » : d'une part une *poussée* de l'inversion au niveau de *Clari*, et d'autre part une chute brutale de la non-expression dans *Coustumes.* Ces deux cas illustrent, de manière saillante, la difficulté à traiter les textes singuliers dans la perspective d'une évolution générale.



Graphique 9a : Inversion et non-expression de P1 + P3

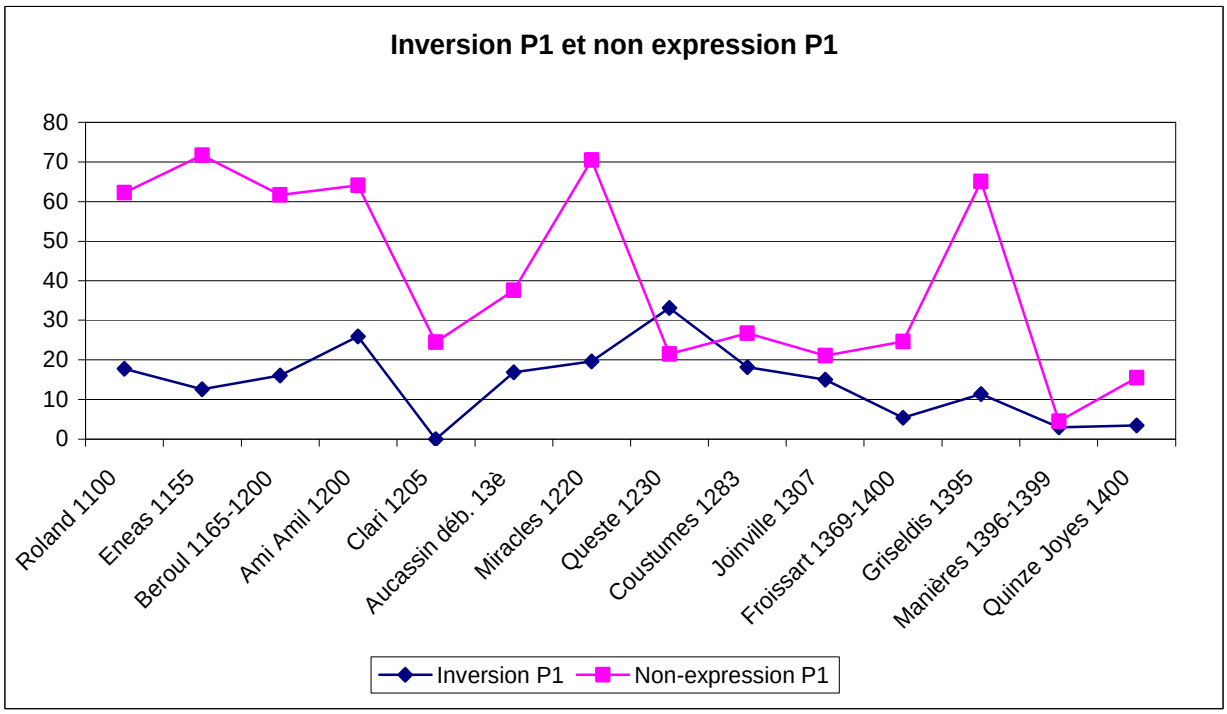


Graphique 9b : Inversion et non-expression de P1 + P3. Chronologie stricte.

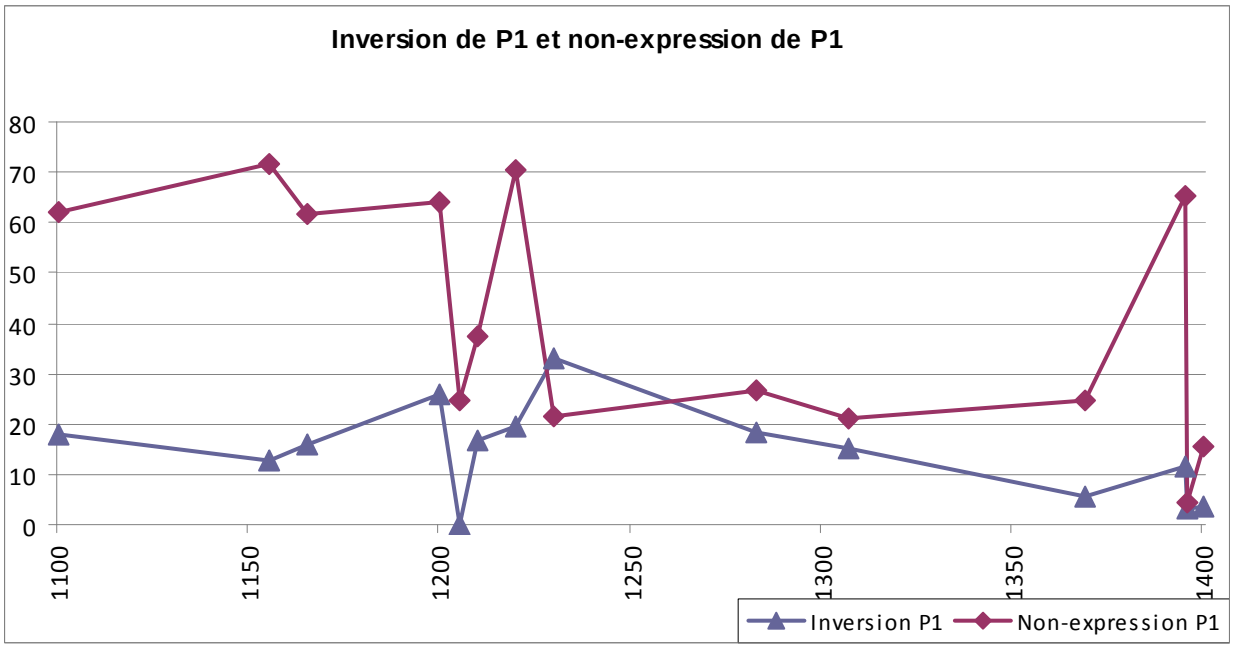


Graphique 9c : Inversion et non-expression de P1 + P3

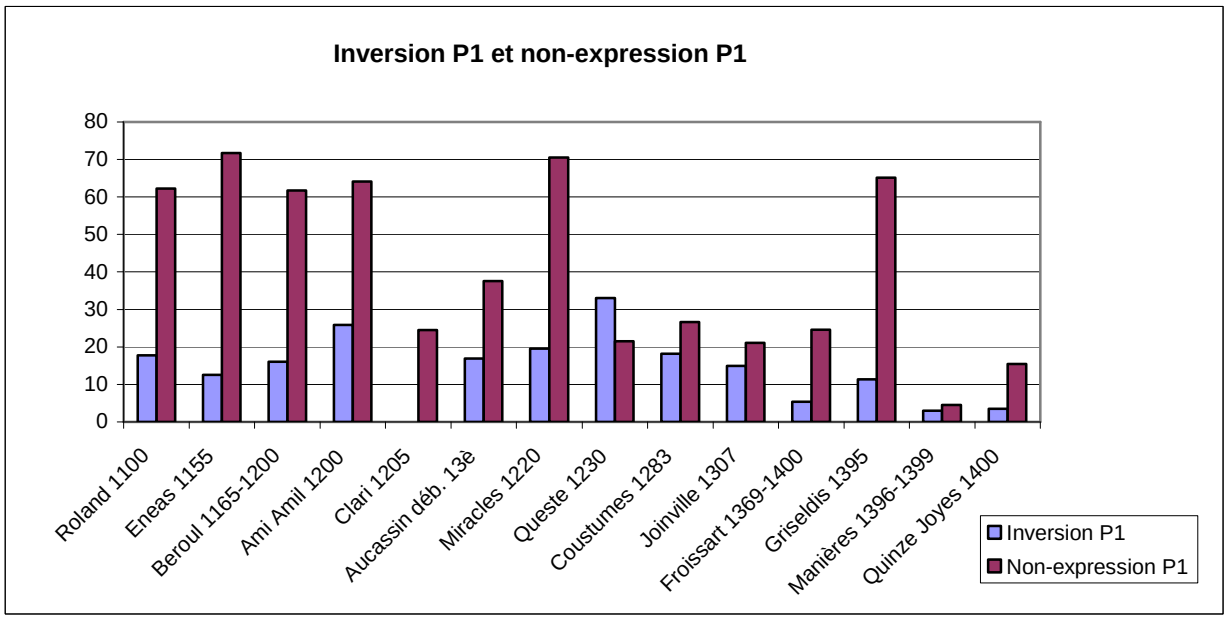
Inversion et non-expression de P1 :



Graphique 10a : Inversion et non-expression de P1..



Graphique 10b : Inversion et non-expression de P1. Chronologie stricte.

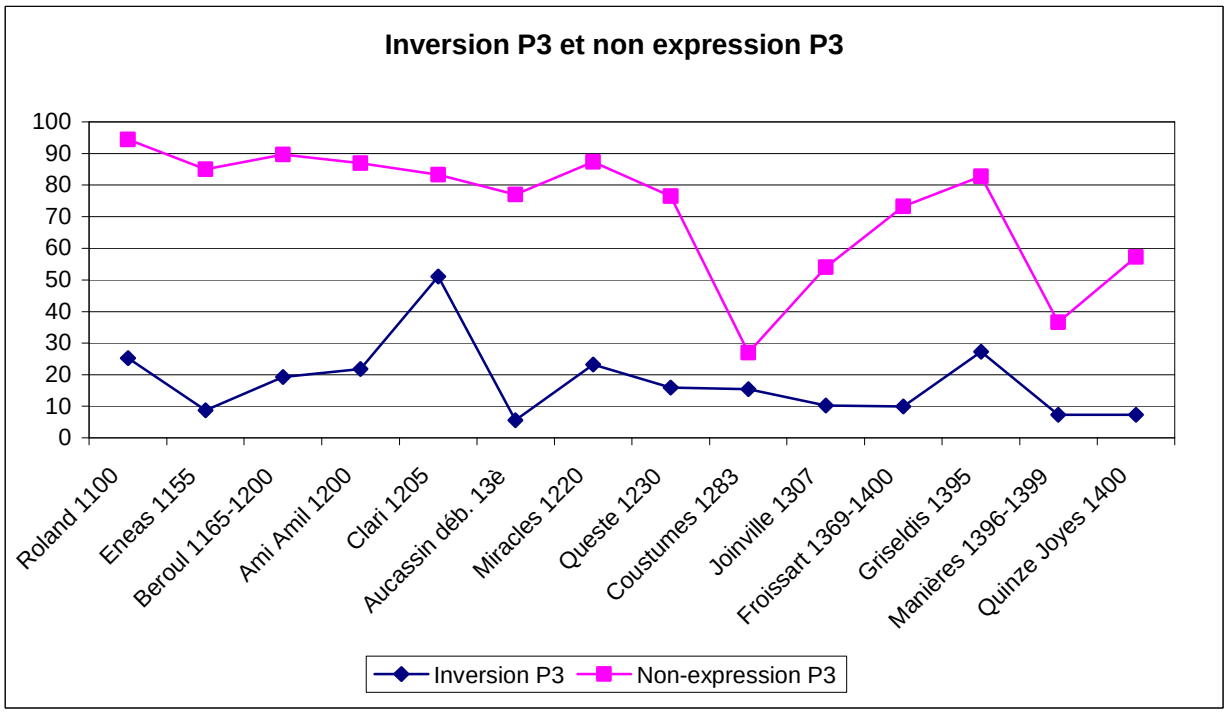


Graphique 10c : Inversion et non-expression de P1.

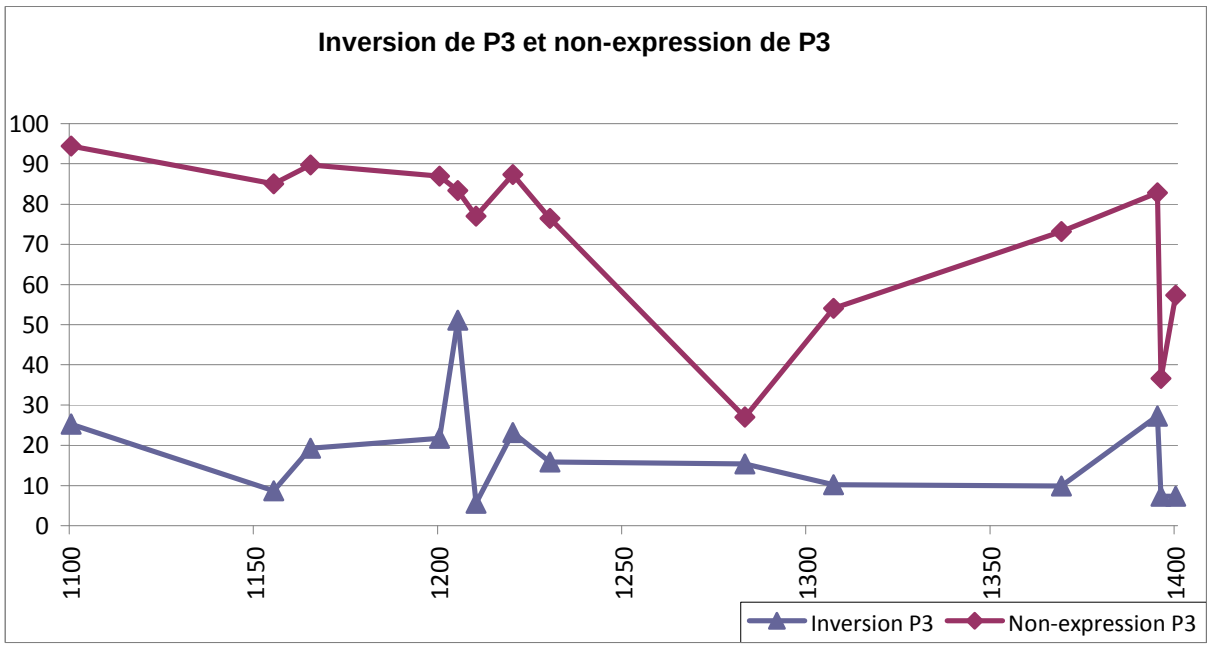
L'observation des graphiques 10a-c montre que les deux lignes (ou les deux crêtes dessinées par le sommet des barres en 10c) suivent un mouvement relativement parallèle jusque *Miracles* (1220), qui présente un point culminant remarquable au niveau de la non-expression. A partir du texte suivant (*Queste*, 1230), les lignes divergent nettement (hormis de *Coustumes* à *Joinville*). Signalons le croisement des lignes au niveau de *Queste*, qui de manière très atypique, présente une fréquence d'inversion de P1 supérieure à sa non-expression. J'ai déjà eu l'occasion de souligner

la spécificité de ce texte au regard tant de la position que de l'expression de P1. Signalons aussi, à la fin du 14ème siècle, le caractère pareillement singulier de *Griseldis*, qui accuse une forte poussée de la non-expression de P1.

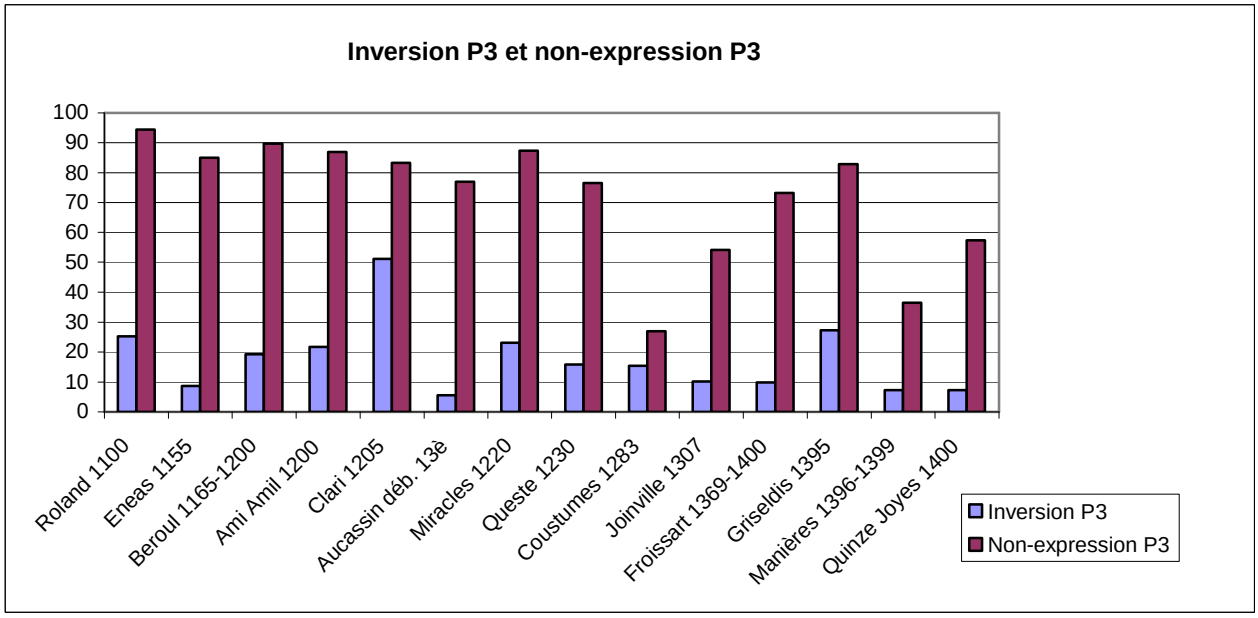
Inversion et non-expression de P3



Graphique 11a : Inversion et non-expression de P3



Graphique 11b : Inversion et non-expression de P3. Chronologie stricte.



Graphique 11c : Inversion et non-expression de P3.

On retrouve pour P3 des tracés proches de ceux observés pour P1+P3, avec néanmoins des mouvements plus nettement convergents entre les deux lignes de fréquence. Certes il ne s'agit pas d'un véritable parallélisme, mais hormis les deux pointes que constituent *Clari* pour l'inversion (la plus fréquente du corpus) et *Coustumes* pour la non-expression (la plus rare du corpus), les deux lignes dessinent une progression assez voisine, bien plus en tout cas que ce que nous avons pu observer pour P1.

Cette plus grande convergence entre les fréquences d'inversion et de non-expression de P3 qu'entre celles de P1 est confirmée par le calcul du coefficient de corrélation.
Celui-ci a été établi pour évaluer la liaison entre l'inversion et la non-expression de P1 d'une part, et entre celles de P3 d'autre part. Le coefficient a été calculé selon les modalités présentées en 3.5.2.3., avec cependant une légère dérogation à ce qui se fait le plus couramment : il a été établi à partir des fréquences relatives et non pas des fréquences absolues. Ce choix se justifie par la différence de taille entre les textes, qui contribue largement (avec les fréquences relatives) aux écarts que l'on peut observer entre effectifs absolus d'un texte à l'autre. Le calcul de la moyenne, sur lequel se fonde celui des écarts réduits, aurait été biaisé par la forte variabilité de la taille des textes.

Sans surprise, le coefficent de corrélation est légèrement plus élevé pour P3 (0.43) que pour P1 (0.37). L'écart entre les deux n'est cependant pas spectaculaire. Surtout, 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 : il n'y a pas de dépendance statistique entre les deux séries de variables (inversion et non-expression) et cela quelle que soit la personne. S'il est apparu des affinités plus grandes entre inversion et non-expression de P3, on ne peut malgré tout parler de corrélation au sens statistique du terme.

A l'issue de cette étude, on voit combien il est difficile, sur le seul plan quantitatif, de dégager des tendances sans écraser les spécificités. Nous avons pu l'observer en ce qui concerne les fréquences d'inversion et de non-expression des personnes 1 et 3, dont les écarts de l'une à l'autre sont souvent masqués par le recours à une moyenne[143]. De même, il apparaît que des textes proches dans le temps peuvent présenter des écarts de fréquence assez importants, phénomène qu'occulte un même recours au calcul de la moyenne.

C'est donc bien la variabilité des fréquences qui domine, entre personnes au sein d'un même texte, et d'un texte à l'autre.

Nous avons vu par ailleurs que quatre textes ont un comportement particulièrement singulier : *Clari*, dans lequel l'inversion en général est atypique, de même que la non-expression de P3 (rare) ; *Coutusmes*, qui offre des fréquences de non-expression très basses et proches les unes des autres; *Griseldis*, dont la non-expression en général et l'inversion de P3 sont au contraire très élevées ; et *Queste* enfin, qui présente une fréquence d'inversion de P1 supérieure à celle de sa non-expression.

Que faire de ces textes singuliers ? Faut-il les traiter à part ?

Si la pertinence du critère chronologique est avéré sur le long cours (non-expression et inversion ont très largement régressé), il ne s'impose pas nécessairement comme caractère déterminant au sein de la période considérée : il n'y a pas de baisse régulière. Les autres critères envisagés (dialecte, domaine et genre, forme) ne jouent de leur côté que ponctuellement un rôle. J'ai par ailleurs insisté sur le problème de

[143] Moyenne qui inclut généralement aussi les personnes 2, 4 et 5.

'parasitage' entre les paramètres : par exemple, lorsque deux textes présentent des fréquences voisines et qu'ils appartiennent au même domaine et au même dialecte, il n'est pas facile d'établir lequel de ces facteurs est déterminant, ou bien s'ils le sont tous les deux (il se peut aussi qu'aucun ne le soit).

Assurément le corpus mérite d'être élargi, afin de décliner de manière plus sytématique les différents paramètres, et afin aussi de mieux représenter le milieu des 13ème et 14ème siècles. J'ai justifié le nombre moins important de textes pour ces deux périodes par le fait que j'avais souhaité bien documenter le tournant des 12ème-13ème siècles, ainsi que la fin du 14ème siècle. Ce choix était motivé méthodologiquement : ne pouvant multiplier le nombre de textes pour l'ensemble de la période couverte pour des raisons de coût de traitement, j'ai pris le parti de le faire pour deux créneaux temporels. Mais ce choix répondait aussi à des motivations linguistiques : j'ai fait l'hypothèse qu'il s'agissait des deux bornes entre lesquelles les changements s'opéraient. De fait, il apparaît une relative homogénéité parmi les textes de la fin du 12ème siècle, tant en ce qui concerne l'expression que la position du sujet. Dans une moindre mesure, on observe aussi des affinités fréquentielles entre les textes de la fin du 14ème siècle, au moins en ce qui concerne l'inversion de Sp. Entre ces deux bornes, la situation est bien plus désordonnée et complexe.

Il semble donc – mais cela méritera d'être confirmé par la prise en compte d'un nombre plus important de textes – que la phase de transition se caractérise, non pas par une baisse régulière et généralisée, mais par des pratiques très diversifiées d'un texte à l'autre. On peut faire l'hypothèse que ce moment de 'passage', et donc de désorganisation partielle du système, ouvre la voie à une variété d'usages plus grande que pendant les phases de relative stabilité. Dès lors que les 'règles' passées, ou du moins les régularités, s'effritent sans que les nouvelles soient encore bien établies, plusieurs comportements sont possibles : celui qui tend à conserver l'usage passé, celui qui, au contraire, s'inscrit déjà dans la nouveauté, ou bien encore celui qui, simplement, exploite cet espace de liberté langagier.

Les données observées ont par ailleurs montré que l'évolution de la non-expression de P1 est plus chaotique que celle de P3. Or, globalement, la non-expression de P1 est bien moins fréquente que celle de P3 (hormis dans *Coustumes*). Il se peut qu'il y ait un lien entre la fréquence d'une construction et la régularité plus ou moins grande de son évolution. Cette hypothèse doit cependant être envisagée avec précautions, dans la

mesure où elle se fonde sur un phénomène qui implique la première personne. Nous savons en effet que la personne du locuteur est le lieu privilégié de l'expressivité et de la subjectivité : que ses modalités d'expression soient soumises à une forte variabilité d'un texte à l'autre n'est pas surprenant. On notera cependant qu'il n'y a pas une telle différence P1/P3 au niveau de l'inversion, or il se trouve que la position postverbale n'est pas préférentiellement associée, dans ce corpus, à l'une ou l'autre personne, comme c'est le cas pour l'expression, qui 'attire' P1 et rejette P3 (la non-expression de P1 restant néanmoins assez élevée, surtout en comparaison avec l'inversion).

Les remarques formulées ci-dessus constituent bien des hypothèses, qu'il conviendra de confirmer (ou non) par l'exploration de données supplémentaires.
La période intermédiaire du corpus (milieu 13ème-milieu 14ème) exige donc d'être enrichie, mais il convient aussi d'élargir la chronologie retenue, en particulier en remontant au-delà de la borne initiale actuelle. On pourra ainsi essayer de déterminer si la limite actuelle (*Roland*) coïncide temporellement avec le début du recul de l'inversion et de la non-expression, ou si la baisse est déjà amorcée. Cela permettra par ailleurs d'évaluer à quand remonte la préférence de P1 pour l'expression de Sp. Il faut toutefois garder à l'esprit que, au fur et à mesure que l'on recule dans le temps, les tendances « générales » deviennent particulièrement difficiles à établir en raison du nombre restreint de textes auxquels nous avons accès.

Si la poursuite de cette étude passe par un élargissement et un enrichissement du corpus, elle passe tout autant par une approche qualitative, aspect laissé de côté ici pour des raisons exposées au début de ce travail.

Chapitre 5. Vers une approche qualitative

L'étude qualitative se propose de décrire les caractéristiques des constructions qui régressent – énoncés à sujet postverbal et énoncés sans sujet exprimé – et, conjointement, celles des énoncés qui progressent et s'imposent peu à peu. En effet, dès lors que certains contextes n'autorisent plus les séquences à sujet inversé ou non exprimé, il est intéressant d'observer si ces contextes sont compatibles avec les séquences qui deviennent majoritaires. Ainsi, alors que se développent les séquences XSpV, on constate que certains éléments qui se rencontraient dans des séquences à sujet postverbal ou sans sujet exprimé ne s'accommodent pas d'un sujet préverbal. C'est le cas de l'adverbe *si*.

L'observation de l'évolution des contextes se fera sur le plan syntaxique, en considérant en particulier les éléments initiaux qui précèdent le sujet et/ou le verbe, la présence d'une négation, le type de verbe (transitif ou non, modal ou non). Pour les constructions sans sujet exprimé, on distinguera en outre celles qui se trouvent dans un contexte de coordination (82), et les autres (83) :

(82) « Sire, fet il, je sui de la meson le roi Artus, et compainz de la table reonde, **et si ai non Yvains** li Avoltres et filz le roi Urien. (*Queste*)

(83) « - De par Dieu, fet il, et g'irai volentiers. » **Lors dist a un escuier** qu'il mete la sele en son cheval, et li aport ses armes, et cil si fet tout maintenant. (*Queste*)

L'analyse des données portera aussi sur la dimension sémantico-pragmatique des constructions, et des contextes dans lesquelles elles apparaissent. On s'attachera en particulier à la reconnaissance des contextes de continuité ou au contraire de discontinuité thématique, ainsi qu'aux effets de rupture logico-pragmatiques (dont il s'agira d'évaluer la pertinence pour les constructions à sujet postverbal).

Ces pistes seront explorées en préservant la distinction qui a prévalu tout au long de la présente étude entre les personnes 1 et 3. Il s'agira ainsi d'évaluer si les divergences fréquemment observées entre P1 et P3 sur le plan quantitatif trouvent un écho sur le plan qualitatif.

D'une manière générale on adoptera une approche analogue à celle qui a été suivie ici, en cherchant à croiser différents critères.

En particulier il conviendra de déterminer si les textes dont les fréquences d'inversion ou de non-expression de Sp sont voisines présentent des constructions similaires. A l'inverse, on pourra observer si des textes distants en matière de fréquences offrent des affinités quant aux constructions et à leur contextes d'occurrences. Le cas échéant, partagent-ils une appartenance à un même dialecte, à un même domaine ?
Un autre aspect à explorer est celui du rapport entre les fréquences d'inversion et de non-expression et la diversité des constructions et des contextes auxquelles elles correspondent : une fréquence élevée d'inversion de P3 est-elle synonyme de constructions variées ? ou bien se caractérise-t-elle à l'inverse par des schémas réguliers ? Observe-t-on des tendances de ce point de vue, ou au contraire l'absence de régularités ?
De même, les textes qui ont montré une liaison forte entre position et personne (calcul du khi2) offrent-ils des caractéristiques de l'inversion très spécifiques ?

J'illustrerai très brièvement la pertinence des questions soulevées ci-dessus.

Beroul et *Amile* sont deux textes qui se situent au tournant des 11ème et 12ème siècles. Ils sont tous deux en vers et appartiennent au domaine littéraire. Le premier présente des traits normands, le second n'est pas marqué du point de vue dialectal. Leurs fréquences d'inversion de P3 sont proches, respectivement 19.3% et 21.8%, chiffres qui correspondent en outre à des effectifs absolus voisins : 23 pour le premier, 26 pour le second. Cette grande proximité entre les deux textes recouvre néanmoins des modalités d'inversion fort différentes. Ainsi, dans *Amile*, l'objet nominal précède le verbe dans 10 cas sur 26 (ex. 84). On ne trouve par contre aucune occurrence d'objet nominal dans *Beroul*, texte dans lequel on rencontre en revanche 4 occurrences de l'adverbe *si* (85), totalement absent des énoncés VSp d'*Amile*.

(84) Li serf l'entendent, grant joie en ont mené.
***Le droit chemin* ont il bien demandé**,
Toute jor vont tant qu'il fu avespré.
(*Amile*)

(85) Li plus coverz est Guenelons :
Gel connois bien, ***si* fait il moi.**
(*Beroul*)

Clari est le texte du corpus dans lequel l'inversion de P3 est la plus élevée (51%), largement en tête des autres textes. On aurait pu s'attendre à ce que ce chiffre soit synonyme d'une relative diversité des contextes d'inversion. Or il se passe exactement l'inverse : de tous les textes, *Clari* est celui dans lequel l'inversion de P3 est la plus *figée*, au sens où elle se produit derrière l'adverbe *si* dans 71% des cas (66 sur 92), que *si* débute la proposition (86) ou qu'il soit précédé d'une subordonnée, temporelle ou hypothétique (87) :

(86) Quant li marchis oï ches nouveles, si ne fu mie a aise. ***Si* vient il le nuit meesme**, si fait il atorner ses galies, si se met il en mer, anchois qu'il fust jours, si s'en va il ; ainc ne cessa, si vint a Sur. (*Clari*)

(87) Quant chil virrent que leur sires fu mors, si secommenchent a desconfire, ***si* tornent il en fuies**. (*Clari*)

On ne retrouve une fréquence élevée de *si* intial que dans *Quinze joyes* (avec une fréquence cependant moindre : 9 occurrences des 32 énoncés à P1 inversé, soit 28%), texte fort éloigné de *Clari* à tous égards (date, forme, domaine et dialecte) :

(88) - Ave Maria, fait el, je amasse mieulx qu'elles fussent a leurs mesons, **et si feissent elles** si elles savoient bien le plesir (*Quinze joyes*)

On observe des divergences analogues en ce qui concerne P1. Ainsi, alors que *Queste* est le texte qui présente la fréquence d'inversion la plus élevée, les verbes sont relativement peu diversifiés, au sens où quatre verbes (*dire, voloir, prier et voir*) regroupent plus d'un tiers des 64 occurrences :

(89) Et que savez vos ? fet li rois. - Sire, fait il, je le sai bien, et **encor vos dirai je autre chose**, car je voil que vos sachiez qu en cest jor d ui comenceront les granz aventures et les granz merveilles dou saint Graal. » (*Queste*)

Il est cependant vrai que la diversité des verbes est souvent moindre avec P1 qu'avec P3. C'est un point, parmi d'autres, qu'il faudra approfondir.

Voici une dernière comparaison : *Miracles* et *Aucassin* sont deux textes quasiment contemporains, mais qui diffèrent quant aux autres caractéristiques. Leurs fréquences d'inversion de P1 sont assez voisines : 19.6% pour le premier, 16.9% pour le second, chiffres qui recouvrent des effectifs assez voisins (respectivement 9 et 11 occurrences). Or dans *Miracles* 5 des 9 occurences impliquent un verbe nié :

(90) La n'a nïent, n'en lairai mie.
En cele ancïenne abeïe
Ne vuel je que plus soit enclose

(*Miracles*).

alors que dans *Aucassin* une petite majorité des structures VSp est précédée de l'adverbe *encor* :

(91) Et puis j'arai la teste caupee, ja mais ne parlerai a Nicolete me douce amie que je tant aim. **Encor ai je ci une bone espee** et siec sor bon destrir sejorné (*Aucassin*)

Je ne poursuivrai pas ici l'exploration des aspects qualitatifs de l'inversion et de la non-expression de Sp et de leur évolution. Les quelques exemples pointés ci-dessus ont pu montrer combien les pistes sont nombreuses, et combien la dimension qualitative, comme la dimension quantitative, semble ne pas vouloir se laisser réduire à des tendances facilement décelables. L'enjeu de la suite de ce travail consistera à caractériser les structures à sujet inversé et non exprimé, en essayant, malgré tout, de dégager certaines tendances, des affinités, et en les reliant aux caractéristiques quantitatives mises au jour dans le présent travail. L'objectif final est de parvenir à mieux appréhender les limites de l'espace de variation dans lequel évoluent l'expression et la position du sujet pronominal.

Références bibliographiques:

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

Badia, J., Bastida, R. et Haït, JR. 1997. *Statistiques sans mathématique*, Paris : Ellipses

Biber, D. 1988. *Variation across speech and writing*. Cambridge : Cambridge Univ. Press.

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

Brunot, F. (1905-1938, puis 1979-2001. *Histoire de la langue française.* Paris : A. Colin & CNRS Editions, 17 vol.

Champely, S. 2004.*Statistique vraiment appliquée au sport*, Bruxelles : De Boeck

Dees, A., 1979. « Variations temporelles er spatiales de l'ordre des mots en ancien et moyen français », in *Sémantique lexicale et sémantique grammaticale*, M. Wilmet (éd), Bruxelles, VUB Centrum voor Taal- Literatuuurwetenschap, 292-303

De Bakker, C. 1997. *Germanic and Romance Inversion in French, a diachronic study*, Leiden : Holland Institute of Generative Linguistics

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

Carton, F., 2009. « Etude prosodique d'un cas de détachement. Les pronoms personnels pseudo-disjoints dans un corpus de presse parlée en français.» in D. Apothéloz, B. Combettes et F. Neveu (eds) *Les linguistiques du détachement*. *Actes du colloque international de Nancy*, Bern, Peter Lang, 173-187

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).

Combettes, B. 2006. *Grammaticalisation et parties du discours : la différentiation des pronoms et des déterminants en français…*in Guillot C., Heiden S. et Prévost S**.** (eds), *A la quête du sens. Études littéraires, historiques et linguistiques en hommage à C. Marchello-Nizia*, Lyon, ENS Editions

Dufresne, M. 1995. « Etude diachronique de la cliticisation des pronoms sujets à partir du français médiéval ». *Revue Québécoise de Linguistique* 24, 84-109.

Dees, A., 1979. « Variations temporelles er spatiales de l'ordre des mots en ancien et moyen français », in *Sémantique lexicale et sémantique grammaticale*, M. Wilmet (éd), Bruxelles, VUB Centrum voor Taal- Literatuuurwetenschap, 292-303

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.

Dupuis, F. 1989. *L'expression du sujet dans les subordonnées en ancient français.* PhD dissertation, Université de Montréal.

Evrard, E. et Mellet S., 1998. «Méthodes quantitatives en langues anciennes », *LALIES* 18: Paris: Presses de l'Ecole Normale Supérieure, 111-155.

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

Foulet, L. 1935. « L'extension de la forme oblique du pronom personnel en ancien français », *Romania* 61.
Fournier, N. 1998. *Grammaire du français classique*. Paris : Belin
Franzén, T. 1939. *Étude sur la syntaxe des pronoms personnels sujets en ancien français*, Uppsala : Almqvist et Wiksells
Fuchs, C. (eds) 1997. *La place du sujet en français contemporain*, Louvain, Duculot (Coll. Champs linguistiques).10
Fuchs, C., 2006a. La place du sujet nominal en français, *Enonciation et syntaxe*, (F. Hrubaru & A. Velicu, eds.), Cluj : Echinox, 9-25.
Fuchs, C. 2006b. « Locatif spatial initial et position du sujet nominal : pour une approche topologique de la construction de l'énoncé », *Linguisticae Investigationes*, 29 : 1, 61-74.
Fuchs, C. et Fournier, N. 2003. « Du rôle cadratif des adverbiaux initiaux selon la position du sujet nominal », *Travaux de linguistique* 47, 79-109.
Givón, T. 1971. « Historical syntax and synchronic morphology: An archaeologist's field trip », *CLS #7*, Chicago: University of Chicago, Chicago Linguistics Society
Givón, T. 1979. *On Understanding Grammar*, NY: Academic Press
Guimier, C. 1997. « La place du sujet clitique dans les énoncés avec adverbe initial ». in C. Fuchs (ed.), 43-96
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'*, 11-58. Perpignan : Presses universitaires de Perpignan.
Habert, B. et Zweigenbaum, P. 2002. « Régler les règles », *TAL*, 43 :3, 83-105
Herman, J. 1954. « Recherches sur l'ordre des mots dans les plus anciens textes français en prose ». *Acta linguistica Academiae Hungaricae* IV, p. 69-93 et 351-379.
Hopper, P. J. 1998 « Emergent Grammar ». In Tomasello M. (éd), *New psychology of language*, 155-175.
Hopper, P. J. et Traugott, E. C. 2003/1993. *Grammaticalization*, Cambridge: Cambridge University Press.
Kayne, R. 1980. « Extensions of Binding and Case-Marking», *Linguistic Inquiry*, vol. 11, no 1, 75-96.
Lahousse, K. 2003. « La complexité de la notion de topique et l'inversion du sujet nominal », *Travaux de Linguistique*, 47, 111-136.
Lebart L. et Salem A., 1994. *Statistique textuelle*, Paris: Dunod.
Leon, J. 2008. « Aux sources de la Corpus linguistics : Firth et la London school », *Langages* 171 :12-33.
Marandin, J.-M. 2003. « Inversion du sujet et structure de l'information dans les langues romanes », in Danièle Godard (ed), *Langues romanes. Problèmes de la phrase simple*, Paris : éditions du CNRS.
Marchello-Nizia, C. 1979, rééd. 1997. *Histoire de la langue française au XIV^{ème} et XV^{ème} siècles*. Paris : Nathan.
Marchello-Nizia, C. 1985. *Dire le vrai : l'adverbe « si » en français médiéval. Essai de linguistique historique*, Genève : Droz.
Marchello-Nizia, C. 1995. *L'évolution du français : ordre des mots, démonstratifs, accent tonique*. Paris : Armand Colin.

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

Marchello-Nizia, C. 2003. « La "contrainte de contiguïté ordonnée" dans l'évolution du latin au français et aux autres langues romanes », *Presencia y renovacion de la lingïistica francesa*, Actes du Colloque des langues romanes de Salamanque, I. Uzcangar Vivar, E. Llamas Pombo et J.-M. Pérez Velasco éd., Salamanque, Ediciones Universidad, 231-244.

Marchello-Nizia, C. 2006a. *Grammaticalisation et changement linguistique*, Bruxelles : De Boeck.

Marchello-Nizia, C. 2006b. « From personal deixis to spatial deixis: the semantic evolution of demonstratives from Latin to French », *Space in Languages: Linguistic Systems and Cognitive Categories,* M. Hickman et S. Robert éd., Amsterdam, Philadelphie, John Benjamins, 103–120.

Martin, R., 1979. « L'ordre des mots dans le *Jehan de Saintré* » in *Sémantique lexicale et sémantique grammaticale*, M. Wilmet (éd), Bruxelles, VUB Centrum voor Taal- Literatuurwetenschap, 305-337.

Meillet, A., 1982, *Linguistique générale et linguistique française.* Paris-Genève, Champion-Slatkine. « Comment les mots changent de sens » (1906/1982 : 230-271). « L'évolution des formes grammaticales » (1912/1982 : 131-148). « Le renouvellement des conjonctions » (1915/1982 : 159-174). « Convergence des développements linguistiques » (1918/1982 : 61-75).

Moignet, G., 1973. *Grammaire de l'ancien français.* Paris, Klincksieck.

Muller C., 1992 (1ère ed. 1973). *Initiation aux méthodes de la statistique linguistique*, Paris : Champion.

Ollier, M-L., 1995. « *Or*, opérateur de rupture ». *Linx* 32, 13-31.

Paufler, H.D., 1983. « La sociolinguistique et les facteurs internes (Quelques aspects du developpement des pronoms personnels sujets du français) » *Revista de filología románica*, nº 1, 23-34

Perlmutter, D. 1971. *Deep and Surface Structure Constraint in Syntax*, New-York: Holt, Rinehart & Winston.

Prévost, S. 2001. *La postposition du sujet en français aux 15ème et 16ème siècles : une approche sémantico-pragmatique*, Paris, Editions du CNRS

Prévost, S. 2002. « Evolution de la syntaxe du pronom personnel sujet depuis le français médiéval : la disparition d'alternances signifiantes », in D. Lagorgette et P. Larrivée (éds) *Représentations du sens linguistique*, Munich : Lincom, *Studies in Theoretical Linguistics*, 22, 309-329.

Prévost, S. 2003. « La grammaticalisation : unidirectionnalité et statut », *Le Français Moderne*, tome LXXI (2), 144-166.

Prévost S. 2007. « *à propos de, à ce propos, à propos* : évolution du 14ème au 16ème siècle», *Langue Française* 156, 108-126.

Prévost, S. 2010. « 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. [En ligne] http://dx.doi.org/10.1051/cmlf/2010106

Prévost, S. 2011a. « Expression et position du sujet pronominal en français », *Mémoires de la Société de Linguistique de Paris*, Tome XIX, J. François et S. Prévost (éds) 'L'Evolution grammaticale à travers les langues romanes', 13-33.

Price, G. 1966. « Contribution à l'étude de la syntaxe des pronoms personnels en sujets en ancien français », *Romania* n° 87, 476-504.

Romaine, S. 1982. *Socio-historical Linguistics*, Cambridge : Cambridge University Press.

Rouveret, A. 2004 *Les clitiques pronominaux et la périphérie gauche en ancien français*

Schøsler, L., 1991 « les causes externes et internes des changements morpho-syntaxiques », *Acta lignusitica Hafniensia*, 23, 83-112.

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.

Traugott, E. C. 2003. « Constructions in grammaticalization ». In B. D. Joseph and J. Janda (eds.) *The Handbook of Historical Linguistics.* Oxford: Blackwell. 624-647.

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

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

Zink, G. 1997. *Morpho-syntaxe du pronom personnel (non réfléchi) en moyen français (14ème-15ème siècles)*, Genève : Droz.

Zwanenburg, W. 1974. « Perte de la flexion nominale et fixation de l'ordre des mots en français médiéval» dans *XIV Congres so Internazionale di Lingüistica e Filologia Romanze Atti*,III

Zwanenburg, W. 1978. « L'ordre des mots en français médiéval », in *Etudes de syntaxe du moyen français*, R. Martin (éd.), Paris, Klincksieck, 153-171.