Stylistique et Statistique textuelle :

À partir de l'article de C. Muller sur les "pronoms de dialogue"[1]

Denise Malrieu

UMR 7114 MoDyCo – Equipe Linguistique des textes – CNRS – Université Paris X –
Département de Linguistique – 200 Av. de la République – 92001 Nanterre Cedex – France

Abstract

In the frame of the interpretative semantics, the interpretation of a linguistic unit involves to take into account its insertion in contexts of variable spans (text genre, textual sequence and her enunciative status, narration, direct or indirect speech…). Starting from the Muller's statistical analysis of the frequences of personnal pronouns vs possessive adjectives and pronouns in the french theatre (17^{th}), we will discuss the stylistic indices of a familiar interlocution, the problem of the enunciative ambiguity of the personal pronouns, the usage of syntactic parsers, and what might be textual statistics, able to take into account the hierarchical contraints of the text parts.

Résumé

Partant de la démarche adoptée par C. Muller pour construire des indices stylistiques à partir de fréquences d'index, nous discuterons les problèmes méthodologiques liés à la conception de ces indices : nature des corpus de travail et de référence; nature des variables prises en compte; problème de l'ambiguité sémantique des grammèmes selon les genres textuels et les séquences textuelles, en particulier le statut dialogique des pronoms de personnes. L'enjeu étant de passer d'une statistique lexicale à une statistique textuelle, nous développerons cette discussion à travers l'exemple concret des indices proposés par C. Muller en proposant d'autres indices calculables à partir des sorties de CORDIAL Analyseur et d'HYPERBASE d'E. Brunet et en proposant des orientations de développements prenant en compte les contraintes de localités et portées variables.

Mots-clés : statistique textuelle, sémantique interprétative, genres textuels, pronoms et adjectifs de personne, théâtre classique.

1. La place de notre démarche dans les approches statistiques des textes

Constatant que la relecture des articles de C. Muller est toujours aussi revigorante, l'envie nous a prise de discuter et tester avec d'autres outils les propositions qu'il a émises en 1962 sur le repérage de différences stylistiques à l'aide de l'observation des fréquences des pronoms personnels et pronoms et adjectifs possessifs des $1^{ère}$ et $2^{ème}$ personnes. Après avoir situé notre démarche parmi les approches statistiques du texte, nous discuterons le décompte des formes et l'ambiguïté sémantique des grammèmes ainsi que le choix des indices. Nous présenterons ensuite les résultats statistiques sur un corpus de théâtre classique et discuterons enfin les limites des outils actuels et les propositions de développement.

Comme le soulignait V. Beaudouin, les ambitions des analyses quantitatives des textes chez les statiticiens comme Benzécri étaient de nature linguistique : leur but était de tester le bien fondé ou d'apporter des éléments aux analyses des grammairiens par une analyse quantitative des distributions dans les corpus observés. Après une longue période où l'approche statistique des textes était essentiellement lexicale, les développements récents voient naître un courant de

[1] Article repris dans : C. Muller, *Langue Française et Linguistique Quantitative*, Genève, Slatkine, 1979, p. 116-124.

linguistiques de corpus (Biber, 2000, Barlow et Kemmer, 2000), qui rompt avec le présupposé de l'unicité de la norme. Les méthodologies quantitatives ont eu l'heureux effet d'introduire en linguistique le souci de l'observation systématique de la langue en discours qui dépasse la construction introspective de la norme chez le linguiste. Les développements récents des analyseurs sont en train de changer le paysage des possibles et permettront, on peut l'espérer avec leur amélioration, des études quantitatives prenant en compte les distributions syntaxiques de plus ou moins grande portée[2]. Il est donc possible que puisse s'établir le lien rêvé au départ par Benzécri entre approches quantitatives et linguistique. Progressivement, on passe en effet de l'analyse des formes (chaînes de caractères) à l'analyse des lexèmes et grammèmes(unités linguistiques) en contexte, on passe de l'analyse de cooccurrences des concordanciers à l'analyse syntaxique des contextes.

À ce propos je dirai deux mots sur les apports et limites des statistiques sur les étiquettes des catégoriseurs. Les études statistiques des catégories morphosyntaxiques mettent en évidence des distributions contrastées des catégories selon les discours, les champs génériques et parfois selon les genres, lorsqu'ils sont très typés[3]. Cela peut avoir des retombées intéressantes pour la classification automatique des textes (Kessler et al., 1997, Karlgren et Cutting, 1994), pour les recherches d'attribution, ou dans la définition d'heuristiques adaptatives de désambiguïsation différentielles selon les genres textuels. L'analyse syntaxique que les analyseurs actuels effectuent essentiellement au niveau de la phrase constitue une première étape de la désambiguïsation sémantique, mais l'interprétation sémantique d'une unité lexicale (lexème ou grammème) nécessite souvent la prise en compte de contextes plus larges que la phrase, qui comportent des contraintes interprétatives de différents niveaux de localité (prise en compte du genre, de l'inclusion dans une séquence de discours direct ou indirect…). Le genre étant caractérisé par des structures à la fois graphiques, typographiques, des agencements de séquences textuelles, des modalités de propagation d'isotopies, des modalités énonciatives, qui fixent sur des portées variables des contraintes interprétatives, la démarche qui paraît la mieux adaptée consiste dans l'homogénéisation sémiotique et une normalisation relative sous forme de balisage des attributs des séquences textuelles. Le balisage a l'avantage d'autoriser à la fois le commentaire multiple en langage naturel et la définition des portées de ces commentaires.

J'opposerai donc statistique textuelle à statistique lexicale, dans le sens où la première ne prend pas forcément le mot comme unité d'analyse et où elle souhaite prendre en compte de façon conjointe dans le processus interprétatif des unités de contexte de tailles diverses véhiculant des contraintes interprétatives hétérogènes.

Dans la suite, je développerai cette problématique à partir de l'étude stylistique amorcée par C. Muller dans une approche de type statistique lexicale.

2. Notions de norme et de corpus

Dans ses multiples écrits C. Muller insiste de façon récurrente sur le fait que les fréquences sont toujours relatives à des genres textuels et qu'il paraît plus intéressant de comparer des sous-corpus homogènes ou un texte à un corpus de son genre qu'à un corpus, forcément aléatoire, qui prétendrait être représentatif de la langue française[4].

[2] Je ne sous-estime pas pour autant le problème de la nature des modèles sous-jacents aux analyseurs et de leur fiabilité plus ou moins grande selon les corpus utilisés au départ pour construire leurs heuristiques.

[3] Je renvoie à ce propos à l'article de D. Malrieu et F. Rastier à paraître dans *t.a.l.* "Genres et variations morpho-syntaxiques" où nous montrons de fortes variations statistiques des catégories grammaticales en fonction des champs génériques.

[4] On peut citer, p. 118, à propos de son indice : "On pourrait parler dans ce domaine de moyennes pour un genre bien délimité ou pour un groupe de textes très homogène; mais non d'une norme pour la langue, même bornée à une époque définie. Nous renoncerons donc d'emblée à comparer les chiffres fournis par un texte à ceux d'un

Le corpus de travail de C. Muller comporte du théâtre classique, de la poésie symboliste, l'index Södergräd, le corpus Henmon de textes littéraires de la fin du 19ème et début du 20ème (400000 mots), le corpus Gougenheim de textes parlés (312000 mots), Corneille (400000 mots). C. Muller qualifie lui-même son corpus de 2 millions de mots d'"ensemble disparate".

De fait on peut noter que tout au long de l'article, C. Muller s'en tient constamment à un raisonnement comparatif soit entre genres dans un même champ générique[5] et une même période (comédie vs tragédie classiques), soit entre champs génériques (théâtre vs poésie lyrique, soit entre périodes à l'intérieur d'un genre et d'un auteur, soit entre genres à l'intérieur d'un texte lorsque plusieurs genres s'enchâssent).

Il paraît tout à fait pertinent d'ériger ce comportement spontané en principe méthodologique dans l'approche statistique des textes, même si cela n'est pas toujours facile à concrétiser.

3. Les objectifs poursuivis

C. Muller insiste à juste titre sur le fait que l'analyse stylistique repose sur une analyse comparative des configurations des mots grammaticaux et non sur l'étude du lexique[6]. Je reformulerai sa proposition en disant que le style s'analyse à travers les usages différentiels des catégories grammaticales, de structures syntaxiques, prosodiques et thématiques, en ajoutant que le problème qui se pose à nous consiste à départager dans les variations entre textes ce qui tient au genre textuel (qui contient lui-même des contraintes de ce type) et ce qui tient au style.

L'article de C. Muller est quelque peu déroutant en première lecture car il ne définit pas clairement ce qu'il entend par style, il n'explicite pas les raisons qui le font choisir les catégories pronoms personnels et possessifs de 1ère et 2ème personnes pour cette étude du style. Ce n'est qu'en fin de lecture que l'on induit sa définition du style, à partir des résultats retenus (différences de valeur de l'indice entre genres et entre personnages d'une même pièce), ou des remarques sur le remplacement des pronoms personnels par des syntagmes nominaux à déterminant possessif dans le style galant (*je, vous* remplacés par *mon cœur, vos yeux*)[7].

Après lecture on pourrait reformuler l'objectif de C. Muller de la façon suivante : le remplacement du *je* et du *tu* ou du *vous* par des syntagmes nominaux à déterminants possessifs des personnes correspondantes serait à la fois l'expression d'un style précieux ou galant, qui s'exprime soit par des synecdotes corporelles euphémisantes soit dans la sphère des sentiments par une préférence pour l'expression indirecte ou voilée du vécu qui s'effectue à travers une nominalisation du subjectif et de la relation à l'interlocuteur[8]; cette nominalisation agit à la fois comme effacement apparent du locuteur et comme objectivation de son dire. La forme nominale sonne effectivement différemment de la forme verbale et évoque un mode d'interlocution plus relevé: effacement de l'adresse personnelle "inconvenante" au profit d'un

---

ensemble qui prétendrait représenter la langue, ou s'en rapprocher; notre seule intention est de comparer entre eux les chiffres obtenus dans différents textes."

[5] Nous distinguons dans l'arborescence les discours liés à des domaine de pratiques sociales (juridique vs littéraire), les champs génériques (dans le discours littéraire : les genres narratifs, la poésie, le théâtre), les genres (par exemple l'opposition comédie / tragédie au 17ème siècle, le roman policier), les sous-genres (le roman par lettres dans le "roman sérieux" ou le polar dans le roman policier).

[6] p. 123 : "la caractérisation d'un style doit tenir compte des mots grammaticaux et de leur distribution. Ceux-ci forment dit-on à peu près 50% du vocabulaire d'un texte quelconque. Mais que se passe-t-il à l'intérieur de cette masse incolore qui forme comme le tissu conjonctif du discours et dans laquelle viennent s'enchasser les mots de signification?"

[7] Ces objectifs étant plus explicités dans l'article précédent du recueil : « Sur quelques scènes de Molière, Essai d'un indice du style familier. »

[8] Quelques exemples de tournures :
Mon amour s'est rendu mille fois odieux;
Il n'a fait qu'outrager vos attraits glorieux;(….)
(Dom Garcie de Navarre, Acte V, Scène V).

discours pseudo-dégagé du *je* et du *tu*; elle évite en même temps la prise en charge des évaluations implicites et elle présente le locuteur comme capable d' objectiver en permanence son propre vécu, d'en faire une représentation pour l'interlocuteur, d'en faire un jeu social.

4. Choix des variables et des indices

4.1. Grammèmes et ambiguïté sémantique selon les genres

C. Muller part d'une liste des "pronoms de dialogue" : pronoms personnels de 1ère et 2ème personnes singulier et pluriel, ainsi que pronoms et adjectifs possessifs des mêmes personnes.

On est ici obligé de poser la question du statut dialogique des pronoms personnels selon les genres et séquences textuelles. Le *je* et le *tu* sont prototypiquement les pronoms de personne, de l'interlocution proximale, du discours direct. Mais le statut du *je* n'est pas le même dans le roman homodiégétique et dans le théâtre, suivant les types de verbes et les temps des verbes dont il est le sujet (discours vs récit)[9]. On peut donc dire que selon la situation énonciative, le couple cadre *je-tu* lié à cette dernière est différent : dans la situation d'interlocution distale[10], le couple cadre du récit du roman homodiégétique est le couple narrateur / lecteur, les *je* et *tu* locaux référant à l'interlocution du discours direct enchâssée dans le récit. Dans le théâtre, le couple par défaut est les *je -tu* locaux du discours direct, avec des *je* de narrateur dans les séquences de récit enchâssées dans le discours direct, récit qui a souvent pour destinataire non l'allocutaire mais le spectateur. Dans ces séquences de récit, dans la tragédie par exemple, le *je* se retire partiellement de l'interlocution directe, ce qui peut expliquer la présence occasionnelle du passé simple[11].

On peut donc objecter que la 1ère personne n'est pas forcément un pronom de dialogue, qu'elle ne l'est que dans les séquences de discours direct[12]. De même l'impératif a valeur de forme de l'interlocution dans les séquences de dialogue mais non dans les genres à interlocution distale comme les notices techniques ou les proverbes. On peut aussi avoir des doutes sur la valeur interlocutive du "nous" selon les genres et séquences textuelles . Comme le *je*, le "nous" peut avoir des valeurs différentes selon qu'il est dans le discours direct ou dans le récit, il peut désigner un sous-ensemble qui inclut ou pas l'interlocuteur, sans compter le "nous" de majesté. On peut noter que dans *Le Rivage des Syrtes*, il y a davantage de 1PP sujets dans le récit que dans les dialogues.

On se heurte donc là, dès la définition de la variable à l'ambiguïté sémantique de grammèmes, variable selon les genres et les séquences à l'intérieur d'un texte. Les calculs effectués par C. Muller valent donc pour les genres entièrement d'interlocution (le théâtre, et c'est d'ailleurs sur ce champ générique que C. Muller développe son argumentation) mais pas pour l'ensemble des genres qui font partie de son corpus.

Face à ce problème, diverses solutions sont possibles, en partie suggérées par C. Muller :

---

[9] Je renvoie ici aux trois articles de Benveniste : « Structure des relations de personne dans les verbes », « Les relations de temps dans les verbes français », « La nature des pronoms », in *Problèmes de linguistique générale*, 1.

[10] Quoiqu'en dise Benveniste sur la netteté de la distinction discours/ récit à ne pas confondre avec oral. écrit un travail de systématisation reste à faire concernant les contraintes interprétatives liées aux genres, contraintes qui doivent tenir compte à la fois de la situation physique de co-énonciation (avec tout les dégradés de la co-présence multimodale: théâtre=interlocution directe, zone proximale vs roman) et des contrats de l'énonciation représentée (roman hétéro ou homodiégétique par exemple). Autrement dit, ne pas prendre pour argent comptant le "ici personne ne parle", mais analyser comment selon les genres, les contrats de co-énonciation se construisent.

[11] Je ne discuterai pas ici les intéressantes propositions de J.M. Adam et al. qui veulent briser la dichotomie récit/ discours, car ce n'est pas le lieu. Je pose cependant la question de savoir s'il s'agit simplement d'alternance de séquences ou s'il existe des contraintes liées à l'enchâssement (ou hiérarchie) textuel des séquences, lié au genre .

[12] par exemple dans *le Rivage des Syrtes*, les parties "récit" (opposé à "dialogues") contiennent plus de *je* que les parties dialogues

- À l'intérieur des genres narratifs ou du théâtre reprendre les calculs des indices de C. Muller en se limitant aux 2èmes personnes dont on suppose qu'elles ne souffrent pas de l'ambiguité reconnue pour les 1ères personnes[13].
- Reprendre les indices de C. Muller en se cantonnant aux genres exclusivement interlocutifs.

4.2. Le choix des indices et types de calculs

C. Muller propose le calcul de trois indices :

- P/N : Nb de pronoms personnels des 1ère et 2ème personnes (singulier et pluriel) / Nb de mots du texte.
- p/N : Nb de pronoms et adjectifs possessifs des 1ère et 2ème personnes (singulier et pluriel, masculin et féminin) / Nb de mots du texte.
- P/p : Nb de pronoms personnels des 1ère et 2ème personnes / Nb de pronoms et adjectifs possessifs des 1ère et 2ème personnes.

Les deux premiers indices sont une fréquence, indice du poids des Pronoms personnels 1P et 2P ou des pronoms et adjectifs possessifs par rapport à l'ensemble des mots du texte. On peut voir que ces deux indices ont un dénominateur beaucoup plus grand que le numérateur et que donc les variations entre genres ou textes risquent de ce fait d'être rendues très peu visibles. Nous avons vu qu'il paraît plus raisonnable d'écarter le "nous" de l'analyse dans un premier temps. On définira donc P comme Nb de pronoms personnels de 1PS, 2PS, 2PP; et "p" comme Nb de pronoms et adjectifs possessifs de 1PS, 2PS, 2PP. On peut aussi utiliser l'information donnée par les personnes des verbes conjugués.

Quant à "p", la question se pose de son homogénéité, car il cumule pronoms et adjectifs possessifs. On admettra avec C. Muller que le pronom possessif est une anaphore d'un syntagme nominal avec déterminant possessif et qu'il est donc pertinent de cumuler pronoms et adjectifs possessifs.

Mais le rapport p/N est-il pertinent? Il présente de façon renforcée l'inconvénient du très faible numérateur sur très fort dénominateur; on proposera donc aussi d'établir le rapport p/substantifs qui donne un indice du degré de personnalisation en *je* et *tu* des déterminants du syntagme nominal. Pour compléter cette analyse on peut aussi calculer le % d'adjectifs possessifs des trois personnes / déterminants.

Enfin qu'en est-il du rapport P/p dont C. Muller souligne l'intérêt du fait qu'il est indépendant de N?

La valeur de P cumule les pronoms personnels des différentes personnes 1PS et 2PS, 2PP et différentes fonctions sujets et compléments. On pourrait distinguer un indice de familiarité dans l'adresse : % de Pronoms personnels et de pronoms et adjectifs possessifs de 2PS vs 2PP / pronoms personnels + pronoms et adjectifs possessifs[14] et un indice de circonlocution dans l'adresse qui serait le P/p de C. Muller revu, que l'on pourrait décomposer en P1PS/p1PS, P2PS/p2PS et P2PP/p2PP, pour voir si cet indice est homogène selon les personnes.

On peut aussi différencier les indices selon les fonctions syntaxiques : pour chaque texte on dispose du % de pronoms personnels 1P et 2P sujets / nb total de pronoms personnels sujets, ce qui donne un indice d'interlocution directe dans l'usage des pronoms personnels.

[13] les usages non interlocutifs du « vous » restant très minoritaires dans le théâtre.

[14] Bien qu'il ne soit pas évident que le *tu* ait même valeur de familiarité dans la tragédie et dans la comédie (cf plus bas).

5. Les traitements sur un corpus réel

5.1. Le corpus

Nous utilisons les données statistiques fournies par Synapse[15] en restreignant les traitements au corpus de théâtre, pour être certaine de travailler sur des contextes d'interlocution et au théâtre classique pour compléter les statistiques disponibles par CORDIAL(43 comédie, 12 drames) avec les statistiques fournies par la base Théâtre classique (16 comédies, 10 drames) et la base Molière d'E. Brunet.

5.2. Les spécificités du théâtre et les différences comédie / drame sur l'ensemble du corpus (toutes périodes confondues).

Variables en déficit pour le théâtre	**Variables en excès pour le théâtre**	**Comparaison drame / comédie**
	% de noms propres humains (++)	D > C
	% noms propres prénoms (++)	C > D
% de déterminants / total des mots		D > C
% d'adj poss/ mots (0,28)		
	% d'adj poss / déterm (++)	D (24%) > C (20%)
	% d'adj poss 1PS/ adj poss (++) (<Mémoires)	
% d'adj poss 2PS Comédie		D > C
% d'adj poss 1PP		D > C
	% d'adj poss 2PP (++)/ adj poss	C > D
	% de pron poss/mots (0,06)	
	% pron poss 1PS/ pron poss (++)	C > D
	% pron poss 2PS/ pron poss (< poésie)	D > C
	% pron poss 2PP (++)	C > D
	% de pron pers / Pronoms	
	% pron pers sujets / pron	
	% pron parmi sujets(++)	C > D
	% pron pers parmi sujets++	C > D
	% Pron pers 1PS parmi les sujets (++)	
	% Pron pers 2PS parmi les sujets (++)	D > C
	% Pron pers 2PP parmi les sujets (++)	C > D
	% Pron Pers 1PS/ pron pers (++)	
	% Pron Pers 2PS (<Poésie)	D > C
% Pro Pers 1PP< Mémoires		
	% Pron Pers 2PP (++)	C > D

Spécificités du Théâtre / Corpus d'ensemble et comparaison Drame / Comédie

Un premier examen des spécificités du théâtre par rapport aux autres champs génériques pour l'ensemble du corpus de Synapse donne déjà des indications intéressantes qui peuvent se résumer ainsi[16] :

[15] Je remercie D. Laurent pour l'ensemble des données statistiques fournies sur environ 250 catégories morphosyntaxiques à partir des sorties de l'analyseur CORDIAL sur un corpus de 2567 ouvrages intégralement étiquetés, constitué de 81% d'ouvrages littéraires, dont 67% de genres narratifs, 8,7% de théâtre et 4,9% de poésie. Ces variables consistent essentiellement en % qui permettent des comparaisons directes entre textes ou corpus.

[16] Les spécificités sont ici calculées sur chaque variable à l'aide des tests statistiques disponibles dans SAS (entre autres le test de Scheffe), les variables étant exprimées en %.

- Déficit de la 3ème personne (singulier et pluriel) au profit de la 1PS pour les verbes et pronoms personnels, et de la 2PP pour les pronoms personnels et possessifs; du Nb de substantifs par proposition et du % d'adjectifs /mots signifiants.

- Excès de pronoms personnels / pronoms; d'adjectifs possessifs / déterminants; verbes à l'impératif (et au présent de l'indicatif au détriment de l'imparfait et du passé simple).

- Paragraphes et phrases brefs.

Le déficit constaté des déterminants dans le théâtre est dû à deux facteurs essentiels : non pas un déficit des substantifs mais à un excès des noms propres et des pronoms sujets. Du coup le % d'adjectifs possessifs / mots est en déficit, mais il est en excès par rapport aux déterminants surtout dans le drame (sauf pour les 2PS et 1PP, en déficit). On voit donc là l'intérêt de ne pas prendre le Nb de mots comme dénominateur mais la catégorie qui subsume directement : le Nb de déterminants, lui-même dépendant du % de noms propres / noms[17].

Sur l'usage des personnes : pour la 2PS on constate que pour les adjectifs et pronoms possessifs et personnels le drame est supérieur à la comédie (on reviendra plus tard sur ce phénomène), qui préfère la 2PP et que c'est la poésie qui utilise le plus la 2ème personne du singulier. La 1PP est en déficit pour l'adjectif possessif et pour le pronom personnel, (elle est par contre très utilisée dans les mémoires et récits de voyages), ce qui confirme que la 1PP n'est pas une personne de l'interlocution. On peut donc déjà souligner que les paradigmes pronominaux sont contrastés selon les champs génériques.

5.3. Les variables disponibles et variables à inférer

Nous considérons dans un premier temps les variables fournies par Synapse et non les variables qui seraient issues des traitements par programmation des sorties des textes étiquetés car nous ne disposons pas de suffisamment de textes pour ce faire.

Le calcul des indices définis plus haut suppose de disposer des chiffres absolus sur un certain nombre de catégories, chiffres dont nous ne disposons pas et que nous ne pouvons pas inférer des chiffres disponibles et une suggestion qui ressort est que pour un texte donné il serait utile de disposer des chiffres absolus des étiquettes du catégoriseur, de façon à autoriser une définition plus souple des variables inférables. C'est ce à quoi vise le module d'interrogation des codes de Cordial dans la nouvelle version du logiciel Hyperbase.

Certains de ces chiffres n'étant pas inférables, nous utiliserons pour certaines variables les chiffres obtenus avec la base de théâtre classique d'Hyperbase (version non catégorisée), et avec la base Molière (où l'on peut obtenir les statistiques sur certaines catégories) combinant ainsi l'usage des deux logiciels Hyperbase et Cordial Analyseur.

5.4. Le calcul des indices décrits en 4.2.

5.4.1. Les différences entre drame et comédie au 17ème.

Le calcul des indices P et p/ Nb de mots et / Nb de Noms, de P/p confirment bien les résultats de C. Muller, comme l'indique le tableau ci-dessous.

On note donc une plus forte présence de pronoms personnels dans la comédie et une plus forte présence d'adjectifs et pronoms possessifs dans le drame, ce qui signifie un mode d'interlocution plus direct dans la comédie que dans le drame. La différence concernant la fréquence des adjectifs possessifs / déterminants est très nette. On constate que ce phénomène est observé pour chacune des trois personnes.

[17] On voit ici les retombées directes du non balisage des textes sur l'interprétation des statistiques.

Indice	Comédie	Drame	Total
P/Nb mots*100	7,93	7,02	7,85
P/Nb noms*100	35,01	30,86	33,41
p/Nb mots*100	2,03	3,05	2,42
p/Nb noms*100	9,04	13,65	10,81
P/p	3,87	2,34	3,14
P1PS/p1PS	3,80	2,47	3,17
P2PS/p2PS	3,71	1,72	2,50
P2PP/p2PP	4,06	2,39	3,32
% Adj poss/ détermin	23,41	34,42	28,9

Les différences Drame / Comédie sur les indices de C. Muller revus[18]
(où P = Nb de pronoms personnels 1PS, 2PS, 2PP; p = Nb d'adjectifs et pronoms possesifs 1PS, 2PS, 2PP; détermin = déterminants; P1PS/p1PS désigne le rapport P/p à la 1PS)

Pour aller plus loin dans l'analyse, on peut se demander si ces différences seraient dues à une proportion différente de substantifs / mots signifiants : y en aurait-il plus dans le drame? On constate qu'il n'en est rien : légèrement plus de noms dans la comédie et de verbes dans le drame / mots signifiants. Mais un doute subsiste sur ces chiffres car les textes ne sont pas balisés et comprennent les noms des personnages (très fort % de noms propres prénoms dans la comédie, et lié à ceci, plus fort % de déterminants / mots dans le drame).

Si maintenant nous comparons les proportions occupées par chaque personne pour les verbes, les adjectifs et pronoms possessifs, on obtient ceci :

	1PS		2PS		3PS		1PP		2PP		3PP	
	C	D	C	D	C	D	C	D	C	D	C	D
Verbe	24	21,5	5,31	6,7	46,8	50	3,46	2,58	15,36	13	6,17	7,39
Adj poss	35,6	44,5	6,4	15,2	14,4	12,12	5,5	3,14	36,2	23,8	1,7	1,17
Pron poss	40,3	33,3	4,3	7,4	22,8	32,1	6,5	5,4	21,4	16,3	4,5	5,4

Répartition des personnes sur les 3 catégories (CORDIAL): % en ligne

On peut noter que : i) pour les verbes, le drame connaît un déficit de 1PS, 1PP et 2PP au profit des 2PS et 3P, ce qui est en concordance avec le possible remplacement des pronoms de personne par le syntagme nominal ; ii) la répartition des adjectifs possessifs se distingue nettement de celle du verbe : la 1PS et la 2PP y sont beaucoup plus dominantes, au détriment de la non personne, ce qui semble indiquer que l'adjectif possessif est fortement lié soit à l'interlocution soit au monologue, avec un excédent pour le drame en 1PS et 2PS et pour la comédie en 2PP.

Si l'on revient aux % issus d'Hyperbase sur les 3 personnes 1PS, 2PS et 2PP, les seules différences sensibles entre drame et comédie portent sur l'excédent de Pronoms Personnels et adjectifs possessifs 2PP pour la comédie et d'adjectifs possessifs 2PS dans le drame (cf. le tableau sur la répartition des 3 personnes)[19].

On peut préciser les choses concernant l'interlocution directe ou indirecte, en tenant compte de la fonction syntaxique. On constate que la comédie comporte un plus fort % de pronoms parmi les sujets (76,7 contre 64,04) et de pronoms personnels parmi les sujets (64,12 contre 55,4)

[18] Les chiffres obtenus ne correspondent pas avec ceux de C. Muller car je n'ai pas inclus la 1PP.
[19] Dans le théâtre classique d'Hyperbase l'analyse factorielle des pronoms et adjectifspossessifs pour chaque personne 1PS, 2PS, 2PP montre sur le premier axe (plus de 54% de la variance) d'un côté les pronoms et les comédies et à l'opposé les adjectifs possessifs et les tragédies.

avec un léger excédent de pronoms sujets 2PS pour le drame[20] et un excédent de 2PP pour la comédie (19,2 contre 15,7), sans différence pour la 1PS.

On peut se poser la question de savoir si le mode d'expression précieux ou indirect n'est pas partiellement lié au choix de la modalité vers / prose et si la forte fréquence des adjectifs et pronoms possessifs 2PP dans la comédie n'est pas explicable par le fait qu'un bon nombre des comédies du corpus sont en vers. Il est en effet possible que le syntagme nominal à déterminant possessif soit plus compatible avec le rythme de l'alexandrin que le pronom personnel sujet d'un verbe. Il serait donc intéressant d'explorer les positions réciproques des deux types de patrons syntaxiques à l'intérieur de l'alexandrin. En ce qui concerne l'influence de l'alexandrin, le travail de V. Beaudouin (1999) donne des pistes :

p. 296 : "Outre le fait que les sixième et douzième positions [positions dans l'alexandrin, D.M.] sont quasi-exclusivement occupées par des mots pleins et que sur les positions 1 et 7 les mots-outils dominent largement, les détails par catégorie permettent de voir que le nom l'emporte très largement en positions 6 et 12 et que parmi les mots pleins, seuls les verbes n'occupent pas spécifiquement les positions finales du segment métrique : ils apparaissent surtout sur les positions centrales".

p. 312 : "Les allures syntaxiques des deux hémistiches sont proches, même s'il existe quelques différences significatives : les noms et les adjectifs sont plus fréquents sur le second hémistiche, tandis que verbes, adverbes et noms propres le sont davantage sur le premier. Les prépositions sont très représentées en position 7, les conjonctions en position 1."

En position 12, on observe 54% de noms et 20% de verbes.

Ces résultats montrent l'usage préférentiel du syntagme nominal en fin de vers[21]. Si, de plus, l'on tient compte du fait que les thématiques du drame comme V. Beaudouin les a catégorisées (honneur et gloire, passion amoureuse, victoire et défaite) sont les lieux de l'expression du conflit et de la subjectivité, la forte fréquence des pronoms et adjectifs de personne n'est pas étonnante. Nous allons donc contraster les comédies en vers et en prose de Molière.

5.4.2. Les différences significatives entre comédies de Molière en vers et en prose

(8 pièces de chaque dans le corpus *Théâtre classique* d'Hyperbase, 19 pièces en prose et 14 en vers dans le corpus Synapse comme dans la base *Molière* d'Hyperbase).

Indice	Com P H	Com V H	Com P M	Com V M	Com H	Drame H
P/Nb mots*100	8,16	7,70	7,91	7,43	7,93	7,02
P/Nb noms*100	35,53	35,20	33,80	34,52	35,01	30,86
p/Nb mots*100	1,75	2,30	1,72	2,30	2,03	3,05
p/Nb noms*100	7,58	10,5	7,33	10,64	9,04	13,65
P/p	4,76	3,47	4,85	3,32	3,87	2,34
P1PS/p1PS	4,48	3,57	4,70	3,42	3,80	2,47
P2PS/p2PS	6,46	2,83	6,58	4,22	3,71	1,72
P2PP/p2PP	5,26	3,60	5,14	3,06	4,06	2,39
% Adj poss/ déterm	22,71	24,57	21,24	23,03	23,41	34,42

Comparaison des comédies en prose et en vers de ***Molière*** *(H signifie baseThéâtre classique, M signifie base Molière)*

[20] Le plus fort % de 2PS dans le drame par rapport à la comédie n'exprime pas une familiarité plus grande, mais le poids très fort dans le drame du dialogue du héros et de son confident. Cf V. Beaudouin : "Ces figures du double nous disent que héros et confident sont les deux faces d'une même unité. Le dialogue entre héros et confident permet de rendre explicite, sous forme rhétorique, ce qui ne serait autrement qu'un monologue intérieur du héros."

[21] Voici quelques exemples ;

> Approu**vez** ma fai**bless**e, et souf**frez** ma dou**leur ;** (Corneille, *Horace*, vers 1.)
> Insultez, inhumaine, encore à mon malheur.
> Allez, il vous sied mal de railler ma douleur,
> Et d'abuser, ingrate, à maltraiter ma flamme,
> Du faible que pour vous vous savez qu'a mon âme. (Molière, *Les Fâcheux*, vers 233-236.)

On vérifie bien que les indices différencient les deux types de comédie et que ceux de la comédie en vers sont intermédiaires entre ceux de la comédie en prose et ceux du drame. La CV ne comporte pas moins de pronoms personnels que la CP mais elle comporte beaucoup plus d'adjectifs possessifs des 3 personnes (souvent le double). La différence est la plus forte pour la 2PS (davantage de Pronoms personnels 2PS en CP et d'adjectifs et pronoms possessifs en CV).

On peut noter de plus que les % d'adjectifs possessifs de l'interlocution (2PS et 2PP)/ adjectifs possessifs sont plus forts dans la comédie en prose que dans la comédie en vers. On observe davantage de pronoms et de pronoms personnels parmi les sujets dans la CP ainsi que davantage de Pronoms personnels 2PP sujets, ce qui semble bien confirmer que l'interlocution est plus directe dans la comédie en prose.

Un examen de la répartition des personnes des pronoms personnels d'un côté et des adjectifs et pronoms possessifs de l'autre indique les chiffres suivants :

	Pronoms personnels			Adj et pronoms possessifs		
	%1PS/P	%2PS/P	%2PP/P	%1PS/p	%2PS/p	%2PP/p
M Drame	60,5	9,7	29,8	57,18	13,85	28,16
M Comédie	58,78	7,37	33,85	59,26	6,44	34,30
Comédie V	62,09	7,38	30,53	60,41	6,74	32,85
Comédie P	56,71	7,40	36	60	5,34	34,62

Répartition des 3 personnes selon les genres, les sous-genres et P vs p

On constate que quelque soit le genre la 1PS domine largement; elle connaît un maximum pour les Pronoms Personnels dans la CV et un minimum dans la CP au profit du *vous*. On pourrait faire l'hypothèse que la CP est davantage dans l'interlocution et la CV plus dans le monologue. Les données dont nous disposons ne permettent pas de savoir si la 2PS participe réellement du phénomène de remplacement du Pronom Personnels par un syntagme possessif (le décompte de *ton* est surestimé de 17%). Si l'on observe les écarts-réduits dans Hyperbase, on vérifie que pour la 1PS, les pièces où *je* est en déficit **et** *mon* en excès sont des comédies en vers alors que le cas inverse correspond à des comédies en prose. Pour la 2PP, *vous* en excès **et** *votre* en déficit correspond à des CP et le cas inverse à des CV, dont l'exemple le plus typique est Dom Garcie. Par contre, pour la 2PS, les fréquences des pronoms personnels et adjectifs possessifs semblent être corrélées positivement, ce qui semblerait indiquer que la 2PS ne participe pas comme les deux autres personnes au style "relevé" dont parle C. Muller. Il serait évidemment intéressant de différencier les patrons de fonctions syntaxiques en fonction des personnes, des types de verbes 1PS[22] et de la place dans l'alexandrin[23], sans parler des différences thématiques entre comédies en prose et en vers. L'examen complémentaire sur 3 comédies en vers (*Amphitryon, Le Misanthrope* et *Dom Garcie*) et trois comédies en prose (*L'Avare, Les Fourberies* et *le Bourgeois Gentilhomme*) des concordances des adjectifs possessifs de 1PS et 2PP et des mots spécifiques des contextes grâce aux fonctionnalités d'Hyperbase donne des informations intéressantes : dans les comédies en vers, les pronoms personnels de 1PS et 2PP ne sont pas sélectionnés par l'écart-réduit dans les contextes de *mon, ma, mes*, par contre *vos* et *votre* le sont, ainsi que *mon* et *mes* dans les contextes de *votre*. A contrario, dans les comédies en prose, les contextes de *mon, ma, mes* sélectionnent par l'écart-réduit les pronoms personnels *je, me, m'* et *vous*, et les contextes de *votre* sélectionnent *vous* et des verbes à la 2PP. Cela

[22] Il est possible que les 1PS de la comédie en prose et de la comédie en vers ne correspondent pas aux mêmes types de verbes (en particulier des verbes "indicateurs de subjectivité" dont parle Benveniste (in *De la subjectivité dans le langage*).

[23] Il faudrait d'ailleurs étudier si les variations de la longueur du substantif à déterminant possessif 1PS ou 2PP en nombre de syllabes sont les mêmes dans la comédie en prose et en vers .

confirme que dans un cas on a substitution des pronoms personnels et dans l'autre coexistence des adjectifs possessifs et pronoms personnels. Sur la composante thématique, un examen superficiel des contextes droits des concordances des adjectifs possessifs 1PS et 2PP met en évidence dans la CV un vocabulaire des sentiments (*âme, cœur, yeux, amour, courage, désespoir, feux, désirs, sentiments, serments*…) et un vocabulaire familial et domestique dans la CP (*père, mère, fille, fils, frère, soeur, argent, chevaux, habit, carrosse*…)

Malgré les limites de cette exploration, liées à la non disponibilité du corpus numérisé et au faible nombre de textes, on peut déjà tirer quelques conclusions et lignes de recherche. Nos études antérieures avaient montré de fortes différences entre discours et entre champs génériques sur des catégories énonciatives (personnes, temps par ex.). Les résultats de C. Muller sur les différences de « familiarité » de l'interlocution entre le drame et la comédie se confirment: interlocution plus directe dans la comédie, plus indirecte par euphémisation et nominalisation des sentiments et de la relation à l'autre dans le drame. Mais il faut aussi noter que, à genre et auteur constants, le choix du mode d'expression en prose ou en alexandrins s'accompagne aussi de ces différences (moins nettes il est vrai) à l'intérieur de la comédie, ce qui confirme bien que ce choix formel met obligatoirement en jeu tout à la fois les composantes dialogique et thématique, qui donnent le ton. Le calcul d'indices fondés sur les statistiques de catégories met en évidence des différences entre genres et entre sous-genres (prose vs vers) mais une articulation reste à établir entre genres et modalités normées de l'énonciation. De plus l'élucidation du statut dialogique des *pronoms de personnes* exige de prendre en compte non seulement le genre mais les séquences textuelles.

En ce qui concerne la définition des indices notre analyse montre que le rapport entre une catégorie et le total des mots d'un texte n'est pas forcément pertinent dans la mesure où certaines variables intermédiaires la contraignent (on a vu par exemple que le rapport adjectifs possessifs/ mots n'est pas pertinent car dépendant du % de noms propres et de pronoms personnels dans le texte); on pourrait d'ailleurs préférer de façon systématique pour le calcul des taux des catégories à paradigme restreint, prendre comme dénominateur non le nombre de mots mais la catégorie qui subsume directement (ici les déterminants).

Pour conclure je dirai deux mots sur les outils à développer. On a vu au cours de l'analyse que le non balisage des textes introduit dans les statistiques un biais non maîtrisable[24]. Il paraît nécessaire de développer des outils de balisage automatique des parties du texte en priorité celles qui sont graphiquement marquées, mais aussi celles qui ne le sont pas (dialogues, récit enchâssé, discours indirect….) pour ne plus confondre des lexies qui peuvent être identiques quant à leur forme et leur fonction syntaxique mais non homogènes sur la composante dialogique. En particulier, si l'on suit les suggestions de Benveniste, il n'est pas pertinent de faire des comptages indépendants des personnes, temps verbaux, et types de verbes, ce qui suppose de travailler sur des patrons distributionnels et non sur des statistiques séparées de ce type de variables et donc de travailler sur un texte étiqueté et les relations syntaxiques et pas seulement sur les catégories et des indices de proximité topologique. On a pu entrevoir en particulier que les fonctions syntaxiques des pronoms étaient un indice utile. Il ne s'agit plus alors de l'étude de corrélations entre variables à l'intérieur d'un type de textes mais de l'analyse contrastive de fréquences de distributions respectant les relations syntaxiques à l'intérieur d'unités typées de longueurs variables.

L'évaluation et l'amélioration des analyseurs représente une autre dimension de l'effort à fournir pour des statistiques textuelles fiables et ceci demande un effort collectif coordonné. De plus l'exploitation des sorties d'analyseurs suppose des outils suffisamment sophistiqués pour retrouver des patrons syntaxiques complexes (cf l'effort en ce sens de L. Audibert avec LoX).

[24] L'exploitation du corpus balisé de théâtre classique établi par C. Bernet constituerait en ce sens un terrain privilégié d'expérimentation.

Une méthodologie d'exploitation des balises du texte[25] reste à développer qui suppose l'élaboration de modèles conceptuels des balises selon les genres textuels pour être en mesure de développer progressivement une modélisation des interactions des contraintes de localités et portées variables.

Références

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

Adam J.-M. et al. (1998). Pour en finir avec le couple récit/discours, *Pratiques*, n° 100 : 81-98.

Barlow M. & Kemmer S., editors (2000). *Usage-based models of language*, CSLI Publications, Standford.

Beaudouin V. *Mètre et rythmes de l'alexandrin classique : Corneille et Racine*, Paris, Champion, coll. Lettres numériques, à paraître.

Benveniste E. (1966). *Problèmes de linguistique générale, 1*, Paris. Gallimard.

Benzécri J.-P. et coll. (1981). *Pratique de l'analyse des données, Linguistique et lexicologie.* Paris. Dunod.

Bernet C. (1983). *Le vocabulaire des tragédies de Racine*. Paris-Genève. Slatkine-Champion : 385 p.

Biber D. (2000). Investigating language use through corpus-based analyses of association patterns , in M. Barlow & S. Kemmer, editors : *Usage-based models of language*, CSLI Publications, Stanford.

Biber D., Johanson S., Leech G., Conrad S.and Finegan E.(2000) *The Longman grammar of spoken and written English*. Londres, Longman.

Brunet E. (1981). Evolution du vocabulaire français de 1789 à nos jours.Slatkine-Champion.

Brunet E. (1986). Méthodes quantitatives et informatiques dans l'étude des textes. En hommage à C. Muller. Slatkine-Champion.

Habert B. and Salem A. (1995). L'utilisation de catégorisations multiples pour l'analyse quantitative de données textuelles. *T. A. L.*, 36, n°1-2, pp. 249-275.

Karlgren J. and Cutting D. (1994) Recognizing text genres with simple metrics using discriminant analysis. *Proceedings of COLING 94*, Kyoto.

Kessler B., Nunberg G.and Schütze H. (1997) *Automatic detection of genre*, Palo Alto Research Center.

Labbé D. (2000). La France, chez de Gaulle et Mitterand, in *Des mots en liberté — Mélanges Maurice Tournier*, Saint-Cloud, ENS Editions. 183-193.

Lebart L., Salem A. (1994). *Statistique textuelle*, Paris, Dunod, 342 p.

Malrieu D. and Rastier F. (2001). Genres et variations morphosyntaxiques, à paraître dans *T.A.L., 43.*

Malrieu D. (2001). Genres et variations morpho-syntaxiques. Quelles variables pertinentes? Journée ATALA sur les genres, Avril 2001.

Muller C. (1977, 1992). *Principes et méthodes de statistique lexicale*, Larousse, 1977, réimpression Champion-Slatkine, 1992, 211 p.

Rastier F. (1989). *Sens et textualité*. Paris. Hachette.

Rastier F. et Coll.(1994). *Sémantique pour l'analyse*. Paris. Masson

Rastier F. (2001). *Arts et sciences du texte*. Paris. PUF.

Oeuvres étudiées et outils de référence :

Audibert L. Librairie LoX, version 3.2. http://laurent.audibert.free.fr/lox.htm

Brunet E. Hyperbase, Etienne.Brunet@unice.fr

Salem A. Lexico3.

CORDIAL Analyseur, Toulouse, Synapse-Développement.Site Web http://www.synapse-fr.com/

Œuvres citées :

Molière : Œuvres complètes, Paris, Gallimard, 1971.

Racine : Théâtre complet, Paris, Garnier, 1980.

[25] L'ergonomie des balises doit autoriser un système évolutif et souple qui n'a d'autre objectif que de tester des hypothèses et de faciliter leur révision.