François Pagès juin 2011
Mini-mémoire de master
Université Paris-Sorbonne

Paternité des œuvres de Molière : Étude critique de l'analyse syntaxique effectuée par E. Rodionova et M. Marusenko.

Introduction.

1 - Mme Rodionova a soutenu en 2008 une thèse intitulée *Méthodes linguistiques d'attribution et de datation (sur l'exemple de l'affaire Corneille-Molière)*. Avec le professeur Marusenko, elle a signé en mars 2010 un article dans le *Journal of quantitative linguistics*. La notice *Paternité des œuvres de Molière* qui figure sur l'encyclopédie en ligne *Wikipedia* indique le lien vers cet article ainsi que vers ce qui est donné comme une version française (en traduction non garantie). Cet article, qui est issu de la thèse de Mme Rodionova, est cosigné par le Pr Marusenko parce que, en Russie, le directeur de thèse est jugé coresponsable de la thèse et se retrouve donc coauteur de tout article procédant de la thèse.

Sauf dans la conclusion, nous nous référerons aux auteurs comme aux « Pétersbourgeois », pour éviter, en employant l'expression « les auteurs », la confusion possible avec Corneille et Quinault. Quand, dans ce travail, nous faisons référence, sans précision, à « l'article », il s'agit de l'article français.

2 - Cet article nous a posé de très significatives difficultés de compréhension.

3 - Certaines de ces difficultés sont dues à des questions de traduction. La principale, mais elle est décisive, est celle-ci : la démonstration repose sur un tableau de 51 paramètres syntaxiques (tableau 2, pages 35-37). Ce tableau est mal traduit, au point d'en être inintelligible. Par exemple, le paramètre X2 est le nombre de phrases simples (=le nombre de phrases comptant une seule proposition) par phrase. On voit que ce nombre ne peut qu'être inférieur ou égal à 1. Or, plus loin dans le texte (dans le tableau 5, pages 41-42), on découvre que ce nombre est, pour Corneille, de 1,8… La lecture de l'article en anglais ne lève pas cette difficulté : la faute de traduction est équivalente (quoique on voie que, en ce passage tout au moins, l'article français n'est pas traduit de l'article anglais, ni l'article anglais traduit de l'article français).

Pour surmonter cette difficulté, nous avons fait appel à une personne russophone (Mme Maria Nikolaïevna Vasilieva), qui a lu avec nous la partie correspondante de la thèse de Mme Rodionova, laquelle thèse est partiellement disponible sur internet. (C'est également à Mme Vasilieva que nous devons l'explication du fait que l'article est cosigné par le Pr Marusenko).

On trouvera ci-dessous le tableau 2 que nous avons traduit avec Mme Vasilieva. Les compétences de Mme Vasilieva en vocabulaire grammatical français étant limitées, notre traduction est certes beaucoup plus intelligible et exacte que les traductions données dans l'article français et dans l'article anglais, mais elle ne possède cependant pas une vraie rigueur scientifique, et il est certain que quelques définitions sont quelque peu approximatives, et il est même possible que quelques-unes soient erronées. On verra que

nous l'avons assortie des exemples qui sont dans la thèse russe, et que nous avons précisé quelles traductions nous semblent peu sûres. Ajoutons cependant que, pour notre travail, la traduction que nous avons donnée est un instrument amplement suffisant.

4 - Certaines de ces difficultés de compréhension sont dues au caractère très mathématique du texte. Pour le comprendre, nous avons eu recours à une personne d'une bonne formation mathématique : M. Christophe Malherbe, titulaire d'un master et d'un Capes en mathématiques, ainsi que d'un master en cryptographie.

5 - La partie mathématique du texte français est quand même substantiellement inintelligible. En effet, on se rend compte que l'article est composé de morceaux de l'argumentation russe mis bout à bout sans que les liens nécessaires entre ces éléments soient présents : ces liens doivent être reconstitués.

6 - L'article anglais nous a également semblé lacunaire. C'est après la fin du tableau 5 (page 42 du texte français, page 46 du texte anglais) que les deux textes se mettent à diverger assez radicalement.

7 - La surcharge de travail en cette fin d'année universitaire nous a empêché d'aller absolument au fond des choses. Nous n'en avons pas moins tenté de comprendre l'essentiel de la démonstration de Mme Rodionova, et nous avons fait relire le présent travail par des spécialistes de mathématiques et de statistiques lexicales pour nous assurer de n'avoir pas commis de contresens.

8 - On trouvera ci-dessous :

I Notre traduction du tableau 2 de l'article.

II Un exposé de la démarche générale des Pétersbourgeois, et une critique de cette démarche.

III Un exposé plus détaillé du début de la démarche, et une critique de ce début.

IV Un exposé plus détaillé du reste de la démarche (telle qu'on peut la reconstituer), exposé divisé en deux parties, chacune accompagnée d'une critique ou d'un commentaire.

Enfin, une conclusion.

I. LE TABLEAU SOCLE.

Le tableau 2 de l'article nous donne cinquante et un paramètres syntaxiques calculables sur une phrase française. La démarche des Pétersbourgeois consistera ensuite à sélectionner les paramètres pour lesquels Corneille et Quinault diffèrent très significativement - et ils en sélectionneront cinq -, et ensuite à tester les œuvres de Molière sur ces paramètres.

Ce tableau est donc le socle de leur argumentation.

Devant chaque définition, il faut sous-entendre « nombre (moyen) de… » ; derrière chaque définition, sauf celle de X1, il faut sous-entendre « …par phrase ». Sauf dans deux cas, la définition est suivie d'un exemple (tiré de la thèse russe), et de la valeur du paramètre dans cet exemple. Quand cela nous a paru ajouter de la clarté, nous avons écrit les phrases en mettant des groupes de mots en italiques, ou en soulignant des mots.

X1 mots par phrase simple. « Je me parle à moi-même » : 5.

X2 propositions. « Mais parce qu'il sent bien le secours qu'il me donne sa familiarité jusque là m'abandonne » : 3

X3 propositions non subordonnées. « Mais où vous a-t-il dit qu'il reçut la clarté ? » : 1.

X4 propositions coordonnées. « C'est là votre vrai nom, et l'autre est emprunté » : 2.

X5 propositions coordonnées dépourvues de verbes conjugués « *Pourquoi cette demande*, et d'où vient ce souci ? » : 1

X6 propositions subordonnées. « Il est bien des endroits où la pleine franchise Deviendroit ridicule et seroit peu permise. » : 1. [on note le parti pris de compter ici une et non deux propositions subordonnées]

X7 propositions subordonnées du premier degré. « Que je ne savois pas, et qui sans doute est belle » : 2.

X8 propositions subordonnées du second degré. « Lucile, dans son âme, rend tout ce que je veux qu'elle rende à ma flamme. » : 1.

X9 propositions subordonnées du troisième degré.

X10 propositions subordonnées du quatrième degré, etc.

X11 propositions sans sujet nominal. « Non, tu feras bien mieux de leur donner avis
Que par mon ordre exprès ils sont de toi suivis. » : 2.

X12 propositions à la fois sans sujet et sans verbe conjugué « La vieillesse devroit ne songer qu'à mourir, et d'assez de laideur n'est pas accompagnée, *sans se tenir encor malpropre et rechignée.* » : 1.

X13 propositions relatives. « Mais son premier amour, que vous avez appris,
Doit de cette contrainte affranchir vos esprits. » : 1.

X14 propositions non relatives en présence d'une relative.

« Dans la confusion que ce grand monde apporte,
Il y vient de tous lieux des gens de toute sorte. » : 1.

[dans une phrase sans relative, x14 = 0]

X15 ''mots pleins'' : Noms, adjectifs, pronoms, verbes, adverbes, déterminants sauf articles [quant aux articles, figurent-ils ou non dans les mots pleins, nous n'avons pas réussi à le déterminer].

« Il est certain Que mon père s' est[1] mis en tête ce dessein » : 11.

X16 ''mots outils''. Complément de X15.

« Il est certain Que mon père s'est mis en tête ce dessein : 2.

X17 noms « Et de ces cotillons appelés hauts-de-chausse » : 2.

X18 adjectifs ou déterminants sauf articles. Exemples donnés : ''ce'', ''mon'', ''quel'', ''quelque''.

« Sa grâce est la plus forte » : 2.

X19 pronoms. « Il s' est fait un grand vol, par qui, l' on n' en sait rien. » : 7. [2]

X20 déterminants numéraux ou pronoms numéraux. « Quiconque de vous deux n'ouvrira pas la porte n'aura point à manger de plus de quatre jours. » : 2.

X21 verbes conjugués. « Il souffre à me voir, ma présence le chasse, Et je ferai bien mieux de lui quitter la place » : 3

« Au reste mon amour, quand je l'ai fait paraître N'a point été mal vu des yeux qui l'ont fait naître» : 3.

X22 formes verbales non conjuguées. « Il souffre à me voir, ma présence le chasse, Et je ferai bien mieux de lui quitter la place » : 2.

« Au reste mon amour, quand je l'ai fait paraître N'a point été mal vu des yeux qui l'ont fait naître » : 4.[3]

X23 adverbes. « Et sans doute bientôt ils viennent en ces lieux » : 2 [« sans doute », et « bientôt »].

X24 prépositions. « C'est à vous que mon coeur a recours aujourd'hui
pour pouvoir s'affranchir de son cuisant ennui » : 3.

X25 conjonctions.

X26 conjonctions de subordination.

[1] On peut remarquer ici que l'auxiliaire est considéré comme un mot plein. De nombreux grammairiens feraient un autre choix.
[2] Signalons que la lettre *l* (euphonique) ne devrait pas être comptabilisée comme un pronom.

« Dis-moi, n'est-il pas vrai, quand tu tiens ton potage,
Que si quelque affamé venoit pour en manger,
tu serois en colère ? » : 3.

X27 conjonctions de coordination : ''et'', ''ou'', ''mais'', ''car'', ''or'', ''donc'', ''ni-ni'', ''soit-soit'', ''tantôt-tantôt''.

X28 attributs.

« Voila de nos maris le procédé commun :
Ce qui leur est permis leur devient importun. » : 1.

X29 compléments d'objet direct.

« Du meilleur de mon coeur je donnerois sur l'heure
Les cent plus beau louis de ce qui me demeure,
et pouvoir, à plaisir, sur ce mufle assener
le plus grand coup de poing qui se puisse donner » : 2 [la thèse russe présente une contradiction : elle souligne les deux Gn, mais donne pour valeur 3].

X30 compléments d'objet indirect.

« Du meilleur de mon coeur je donnerois sur l'heure
Les cent plus beau louis *de ce* qui me demeure,
et pouvoir, à plaisir, sur ce mufle assener
le plus grand coup de poing qui se puisse donner » : 2

X31 sujets

« Au moins, si l'*on* vous voit commettre une sottise,
vous n'imputerez plus l'erreur à la surprise:
votre rôle en ce jeu par coeur doit être su. » : 3.

X32 sujets pronominaux.

« Au moins, si l'on vous voit commettre une sottise,
vous n'imputerez plus l'erreur à la surprise:
votre rôle en ce jeu par coeur doit être su. » : 2.

X33 groupes ayant même fonction (sujet, cod, coi, verbe, cpt circonstanciel, attribut), et dépendant d'un même mot [définition un peu approximative…].

« J'ai souffert qu'elle ait vu *les belles compagnies*,
les divertissements, *les bals*, *les comédies*. » : 4.

X34 mots appartenant à X33.

« J'ai souffert qu'elle ait vu les belles compagnies,
les divertissements, les bals, les comédies. » : 9.

X35 verbes conjugués ayant même sujet, ou, s'ils n'ont pas de sujet, ayant même référent [définition peut-être erronée].

« Va, poursuis ton chemin, cajole tes maîtresses,
adresse-leur tes voeux, et fais-leur des caresses » : 5.

X36 groupes compléments d'un même mot et ayant même fonction [nous donnons la même définition que celle que nous avons donnée pour X33, ce qui signale un problème dans notre compréhension…]

« J'ai souffert qu'elle ait vu *les belles compagnies*,
les divertissements, *les bals*, *les comédies*. » : 4.

X37 groupes gérondifs. « Le ciel ne m'a point fait, *en me donnant le jour*, Une âme compatible avec l'air de la cour. » : 1.

---

[3] ou bien 6. Note complémentaire : on se demande si le participe présent est comptabilisé comme une forme verbale non conjuguée.

X38 mots inclus dans des groupes gérondifs. « Le ciel ne m'a point fait, en me donnant le jour, Une âme compatible avec l'air de la cour. » : 5.

X39 groupes participiaux « Et la crainte *mêlée à mon ressentiment*

jette par tout mon corps un soudain tremblement. » : 1.

X40 mots inclus dans des groupes participiaux. « Et la crainte mêlée à mon ressentiment Jette par tout mon corps un soudain tremblement » : 4.

X41 déterminants - sauf articles -, ou adjectifs épithètes [définition peut-être approximative].

« Quelles *tristes* clartés dissipent mon erreur

Enveloppent mes sens d'une profonde horreur

et ne laissent plus à voir à mon âme abattue

que l'effroyable objet d'un remords qui me tue ! » : 8.

X42 épithètes ou compléments de nom [définition peut-être approximative].

« Il faut, dis-je, pour rompre a toute chose cours,

acheter sourdement l'esclave idolâtrée

et la faire passer en une autre contrée » : 1.

X43 compléments de nom [définition peut-être approximative].

« Je mettrois en ses mains, que je tenois certaines, Quelque bien *de mon père* et le fruit *de mes peines* » : 2.

« C'était le vrai moyen *d'acquérir sa tendresse* » : 1.

X44 sous-cas nominal de X43

« Je mettrois en ses mains, que je tenois certaines, Quelque bien *de mon père* et le fruit *de mes peines* » : 2.

X45 groupes entre virgules, ou entre parenthèses, ou entre tirets, ayant une fonction différente de celles des groupes qui les encadrent [définition approximative…].

« En un lieu, *l'autre jour*, où je faisois visite Je trouvai quelques gens d'un très rare mérite, Qui, parlant des vrais soins d'une âme qui vit bien,

firent tomber sur vous, *Madame*, l'entretien. » : 2.

« Et moi, j'ai de la faim et de l'inquiétude » : 1.

X46 mots compris dans les groupes X45.

En un lieu, l'autre jour, ou je faisois visite,

je trouvai quelque gens d'un très-rare mérite,

qui, parlant des vrais soins d'une âme qui vit bien,

firent tomber sur vous, Madame, l'entretien. » : 4.

X47 noms ''isolés'', en ce sens qu'ils ne sont pas précédés par un déterminant autre qu'un article, et pas complémentés par une épithète ou un complément de nom [définition approximative].

« Vraiment, je ne sais pas si c'est un bruit qui part

De quelque conjecture, ou d'un coup de hasard. » : 2.

X48 groupes nominaux dépourvus de complémentation par une relative adjective. [définition probablement erronée].

« Etre franc et sincère est *mon plus grand talent* » : 1

« Vraiment je ne sais pas si c'est un bruit qui part

De quelque conjecture, ou d'*un coup de hasard*. » : 3.

X49 mots inclus dans X48

« Vraiment je ne sais pas si c'est un bruit qui part

De quelque conjecture, ou d'un coup de hasard. » : 6.

X50 mots pleins (cf X15) inclus dans X48 [le second trait de la définition est peut-être erroné].

« Vraiment je ne sais pas si c'est un bruit qui part
De quelque conjecture, ou d'un coup de hasard. » : 4.
X51 Complément de X50 : X50 ce sont les mots pleins se trouvant dans des groupes caractérisés par un certain trait, X51 ce sont les autres mots de ces groupes.
« Vraiment je ne sais pas si c'est un bruit qui part
De quelque conjecture, ou d'un coup de hasard. » : 2.

II. LA DÉMARCHE GÉNÉRALE, ET SA CRITIQUE.

NB. Dans tout ce qui suit,
C signifie les comédies de Corneille,
Q signifie les comédies de Quinault,
Mn, Mo, etc. signifie toute comédie en vers signée Molière
Xn signifie un des 51 paramètres
Xs signifiera un des 5 paramètres sélectionnés parmi les 51.

On rappelle qu'un écart-type est une mesure de la différence moyenne à la moyenne ; si j'achète une baguette 50 centimes chez un boulanger, et une autre 250 centimes chez un autre boulanger, le coût moyen de la baguette achetée est de 150 centimes ((250+50)/2), et les deux écarts à la moyenne sont de 100 centimes : (100 – 50), et 150 – 100). L'écart-type est fonction de ces deux écarts à la moyenne : racine carrée de la somme des carrés des écarts à la moyenne, divisé par le nombre de cas. Si on décrit un ensemble par une courbe gaussienne (courbe ''en cloche''), la moyenne nous dit où placer cette courbe sur l'axe horizontal, l'écart-type nous donne la forme de cette courbe.

1° La démarche très générale.

Quand on pense qu'un objet Fn doit être rattaché par ressemblance à un ensemble D, ou bien à un ensemble E, ou bien mis dans une classe des « non-attribuables », on utilise une certaine procédure mathématique. Celle-ci permet de distribuer Fn sur D, ou sur E, ou de le rejeter dans l'ensemble des non-attribuables.

Prenons un exemple. On dispose de 100 mauvaises photos de D, de 100 mauvaises photos de E, et de 100 mauvaises photos dites de F. On croit que ces photos dites de F sont, en fait, les unes des photos de D, les autres de photos de E. Comment va-t-on procéder ? On va chercher les traits sur lesquels D et E s'opposent très radicalement. Et on va comparer, sur ces traits-là, chaque photo de F à D, chaque photo de F à E. Cela nous permettra de rattacher les photos dites de F à D ou à E.

Cette procédure n'est essentiellement pas un moyen d'asserter que Fn ressemble beaucoup ou peu à D, que Fn ressemble beaucoup ou peu à E. C'est essentiellement un moyen d'affirmer qu'il ressemble plus à D qu'à E, ou plus à E qu'à D.

Cependant, très secondairement, c'est aussi un moyen de vérifier s'il ne ressemble ni à l'un ni à l'autre. En effet, considérons la classe des non-attribuables : elle comprend des objets dont on ne peut pas affirmer qu'ils ressemblent plus à D qu'à E, mais aussi des objets qui dissemblent et de D et de E. Imaginons qu'une des différences trouvées entre D et E soit que D mesure environ 1,70 m, E environ 1,80

m. Imaginons aussi que la marge d'erreur d'estimation de la taille de F sur les photos dites de F soit estimée à 0,06m. Dans la classe des non-attribuables, on trouvera aussi bien des photos où F semble mesurer 1,74m, 1,75m et 1,76m (il a une valeur intermédiaire entre celle de D et celle de E sur le trait sur lequel D et E diffèrent), qu'on trouvera des photos où F semble mesurer moins d'1,69m ou plus d'1,85m (F diffère de D et de E).

Autrement dit, la procédure est conçue pour distribuer entre D et E, mais elle nous donne aussi une vague indication sur la ressemblance entre les objets de F et ceux de D, et entre les objets de F et ceux de E.

2° On se demandera peut-être : pourquoi ne tester F que sur les paramètres sur lesquels D et E diffèrent très significativement ? Pourquoi ne pas les tester sur tous les paramètres ? Après tout, même un paramètre sur lequel ils diffèrent peu nous donne une indication : D semble avoir 45 ans, à + ou - 20 ans, E semble avoir 35 ans, à + ou – 10 ans. En testant F sur tous les paramètres, on ajouterait en précision…

Nous reviendrons sur ce point (en II, 1°, 5) : répondons d'ores et déjà que, quand on a des paramètres sur lesquels D et E s'opposent très fortement, tester F sur les autres paramètres ne peut pas donner d'indications significatives : c'est donc inutile.

3° On nous dira peut-être : nous avons donné un exemple sur lequel D et E s'opposent nettement sur un seul paramètre. Mais quand ils s'opposent nettement sur plusieurs paramètres, la ressemblance entre Fn et D, ou entre Fn et E, est-ce qu'elle ne nous donne pas un renseignement plus significatif que sur un seul ?

Réponse : c'est effectivement le cas. Cela dit, cela n'altère pas le fait que c'est tout à fait secondairement que cette procédure traite de la ressemblance entre Fn et D, entre Fn et E. Elle ne nous dit pas : en général, Fn ressemble beaucoup à D, ou à E.

4° On se demandera peut-être : pourquoi les auteurs procèdent-ils ainsi ? Ne pourraient-ils pas, pour chaque pièce de Molière, calculer sa ressemblance syntaxique avec C, puis avec Q ? Puis sa ressemblance syntaxique avec Marivaux, et avec Anouilh ? Et nous dire, par exemple, que Mn ressemble beaucoup à C, beaucoup plus qu'il ne ressemble à Q, et qu'il ne ressemble à Marivaux, et qu'il ne ressemble à Anouilh ? La réponse est qu'une ressemblance globale entre deux ensembles est un concept flou, inutilisable pour décider mathématiquement. Pour certains paramètres, Mn va être trouvé ressemblant à C, pour d'autres à Q et à Rostand, pour d'autres à Marivaux et Musset, etc. On pourra peut-être compter qu'il ressemble à C par un peu plus de paramètres qu'à Q ou à Marivaux, mais cela ne nous permettra en rien de prendre une décision. Par contre, si l'on veut décider si Mn ressemble plus à C qu'à Q, ou le contraire, ou pas décisivement plus à l'un qu'à l'autre, alors on dispose d'une procédure mathématique. En d'autres termes, la notion de distance syntaxique moyenne entre deux éléments est une notion floue, molle, dont on ne peut rien tirer. Par contre, la notion de plus grande proximité syntaxique d'un objet avec un ensemble qu'avec un autre ensemble, cette notion est pleine de sens, et est mathématiquement utilisable pour la décision.

5° Si bien que si les Pétersbourgeois n'avaient pas pris pour point de départ que Corneille et Quinault ont écrit nombre des pièces attribuées à Molière, ils n'auraient pas pu utiliser de procédure mathématique décisive. Et l'on en arrive assez naturellement à ce point : si les auteurs n'avaient pas postulé que Corneille et Quinault sont les auteurs de nombre des pièces attribuées à Molière, ils n'auraient rien pu faire mathématiquement. Comme ils disposaient d'une procédure mathématique, ils ont eu la tentation d'adopter le parti selon lequel Corneille et Quinault sont les auteurs des pièces attribuées à Molière…

.6° Mais, si l'on sait que Mn doit être rapproché syntaxiquement de C ou de Q, et si l'on découvre qu'alors effectivement on peut le rapprocher syntaxiquement de l'un des deux, une question se pose encore : est-il sûr que proximité syntaxique vaut identité d'auteur ? Ceci est une question délicate, jamais abordée par les Pétersbourgeois. Nous reviendrons sur ce point capital.

Il constitue le deuxième point de départ ou présupposé des Pétersbourgeois.

7° En résumé.

Que Corneille et Quinault soient l'un et l'autre auteurs de pièces attribuées à Molière n'est pas le résultat de la démarche des Pétersbourgeois, c'en est le point de départ. Ce point de départ posé, toute leur démonstration, ensuite, porte sur le processus de distribution des pièces de Molière entre C et Q, selon la ressemblance syntaxique, distribution qui, pour eux, en application de leur deuxième point de départ ou présupposé, vaut pour attribution respectivement à Corneille et Quinault

.8° Ce que nous venons d'établir, bien évidemment les Pétersbourgeois le savent, mais ils font comme s'ils ne le savaient pas. Ils le savent, puisqu'ils commencent par dire que l'histoire littéraire établit que, très probablement, Corneille et Quinault sont auteurs de nombre des pièces attribuées à Molière. Ils le dissimulent, car, ce que nous venons de dire, à savoir que leur démarche mathématique ne prouve pas l'hypothèse Corneille-Quinault, mais qu'elle la présuppose, on ne le trouve nulle part clairement exposé dans leur article.

III DÉTAIL DU DÉBUT DE LA DÉMARCHE, SUIVI DE SA CRITIQUE.

1° Détail du début de la démarche.

Comment faire, pour mesurer la ressemblance entre un objet et deux autres ensembles (par exemple entre Mn d'une part, et C et Q d'autre part) ? La procédure mathématique à suivre est celle-ci. On mesure C et Q sur tous les paramètres qu'on peut. On cherche ensuite sur quels paramètres ils diffèrent très significativement. On sélectionne ces paramètres et seulement ceux-ci. Après on mesure l'objet (Mn) sur ces paramètres. Puis on cherche si, sur un paramètre sélectionné, Mn ressemble à C, ou à Q, ou ni à l'un ni à l'autre (on ne trouvera jamais qu'il ressemble à la fois à C et à Q : puisque, par construction, C et Q diffèrent sur chaque paramètre sélectionné). Si Mn ressemble à C sur tous les paramètres sélectionnés, on le déclare ressemblant à C, s'il ressemble à Q sur tous les paramètres sélectionnés, on le déclare ressemblant à Q, sinon on le déclare non attribuable à l'un ou à l'autre (on verra plus loin qu'ici nous simplifions un peu, mais enfin, l'idée est celle que nous avons dite ; la simplification tient à ce que, après

avoir fait ceci, on cherche si on ne peut pas réduire la classe des objets déclarés non-attribuables). Ceci est essentiellement ce qui est expliqué en page 10 (avant d'être appliqué une quarantaine de pages plus loin).

Cette page 10 est difficile et cruciale. Commentons-la.

1. On voit que, pour tout paramètre (rappelons au lecteur qu'on désigne un paramètre par la lettre X), on utilise la moyenne et l'écart-type dans n phrases (ici, n=100) d'un ensemble (par exemple, de l'ensemble C), mais que le nombre total de phrases dans C n'est pas pris en compte.

Mais c'est que, quand on a un échantillon suffisant, peu importe la taille de l'ensemble : on le traite comme s'il était infini. Pour faire une comparaison : un sondage de 1 000 personnes permet de connaître les intentions de vote de la population, que celle-ci soit de cent mille, d'un million ou d'un milliard d'individus.

. On a donc, pour Xn,

pour C, mesurés sur 100 phrases, une moyenne et un écart-type.

pour Q, mesurés sur 100 phrases, une moyenne et un écart-type.

2. On calcule alors le ''t de Student'', lequel varie en fonction de ces deux moyennes et de ces deux écart-types (si les moyennes sont proches, t est petit ; si les écart-types sont grands, il est petit) ; t mesurant une dissemblance entre deux ensembles, ça nous donne, pour Xn, la valeur t de la dissemblance entre C et Q.

3. Quand t > 1, 96, alors nous avons un paramètre Xn pour lequel C et Q s'opposent très significativement. Un objet ressemblant à C selon ce paramètre sera dit avoir une ressemblance telle que cet objet sera dit « appartenir » à C, ou « être indiscernable des éléments de C », avec 95% de chances.

Mais comment calculer la ressemblance de cet objet avec C selon Xn ? Eh bien, cette ressemblance se calculera elle-même par un calcul de t. On mesurera, dans cet objet, moyenne et écart-type selon Xn, puis on calculera t pour Mn et C, puis t pour Mn et Q.

En somme : d'abord on retient les paramètres sur lesquels C et Q dissemblent (t > 1,96), ensuite on calcule sur chacun de ces paramètres retenus la ressemblance entre Mn et C, et entre Mn et Q. Si t de Mn et C < 1,96, alors, Mn est dit ressembler à C (selon Xn), sinon il est dit dissembler de lui. Si t de Mn et Q < 1,96, Mn est dit ressembler à Q (selon Xn), sinon il est dit dissembler de lui. (Rappelons que, par construction Mn ne peut pas ressembler à la fois à C et à Q selon Xn, puisque, selon Xn, C et Q sont dissemblants).

4. Ici, nous avons ce qui peut sembler une bizarrerie, mais qui est en fait un prodige des mathématiques, avec lequel tous les statisticiens sont obligés de composer. Il est affirmé ceci : Si t de Mn et Q < 1,96, il y a 95% de chances que Mn puisse être dit indiscernable de Q (selon Xn). Et, si t > 1,96, il est affirmé qu'il y a au moins 95% de chances que Mn puisse être dit non-indiscernable de Q. En somme, 1,96 est une singularité. À 1,96 + epsilon, il y a 95% de chances que telle affirmation soit vraie, et à 1,96 – epsilon, il y a 95% de chances pour que l'affirmation contraire soit vraie. Pourtant, t est continu… Il ne s'agit pas ici cependant pas ici d'une erreur ou d'un coup de force des Pétersbourgeois, comme on pourrait croire ou craindre, mais d'une caractéristique de la procédure mathématique mise en oeuvre. Une image pourra ici peut-être nous être utile. Imaginons un sabre (à deux faces, donc), dont la pointe a pour coordonnée 1,96.

A 1,96 + epsilon, on se trouve sur une face, à 1,96 – epsilon on se trouve sur l'autre face. Une autre façon, moins imagée, de comprendre cette apparente bizarrerie. Les deux énoncés ne sont pas le contraire l'un de l'autre. L'énoncé « Mn n'appartient pas à C » est le contraire de l'énoncé « Mn appartient à C ». Mais l'énoncé ''Mn a 95% de chances d'appartenir à C'' n'est pas le contraire de l'énoncé ''Mn a 95% de chances de ne pas appartenir à C''. Pour en terminer avec ce point, on pourrait le résumer ainsi : Si t>1,96 on admet, avec une chance sur vingt de se tromper, que Mn peut être considéré comme significativement ressemblant à Q. Précisons que le seuil de 1,96 est fréquemment utilisé dans les sciences humaines, mais, selon les exigences des recherches et chercheurs, le choix peut être plus sélectif, de façon à correspondre à un risque d'erreur de 1 sur 50, 1 sur 100, etc.

5. Disons tout de suite que 5 paramètres sont retenus. Mais pourquoi, dira-t-on ne pas prendre en considération les autres paramètres ? C'est que les dissemblances entre C et Q sur ces 5 paramètres sont telles, et les ressemblances entre C et Q sur les autres paramètres sont telles, que prendre d'autres paramètres que les 5 ne modifierait les ressemblances/dissemblances entre Mn et C d'une part, entre Mn et Q d'autre part, qu'infinitésimalement : ces paramètres sont si puissants pour discriminer entre C et Q que, si l'on trouvait, par exemple ressemblance entre Mn et C selon ces 5 paramètres, et dissemblances selon tous les autres, ça ne changerait pas l'affirmation de la ressemblance décisive entre Mn et C.

Précisons que cette notion de ressemblance/dissemblance n'est pas la notion de différence entre les moyennes ; c'est, en fait, t, c'est-à-dire quelque chose qui se calcule en fonction de la différence entre les moyennes et aussi en fonction des écarts-types (et aussi, en fonction de la taille de l'échantillon, ici 100 phrases).

6. Nous avons légèrement anticipé. Pour chaque Xn, on calculera t de Mn et C, puis t de Mn et Q. Si pour les 5 paramètres, on a t de Mn et C < 1,96, on pourra dire qu'il y a ''95%^{5} [95% puissance 5] de chances que Mn soit indiscernable de C (ou appartienne à C, les deux formulations sont équivalentes), ce qui fait, on le voit un très grand nombre.

Récapitulons.

Si nous hésitons entre déclarer qu'un objet (par exemple, une pièce de Molière) ressemble à une certaine collection d'objets (par exemple, l'ensemble des comédies de Corneille) au point qu'il pourrait être classé dans cette collection sans que cela change très significativement les caractères de cette collection, hésiter entre déclarer cela, donc, et entre déclarer que cet objet ressemble à une autre collection d'objets (par exemple, l'ensemble des comédies de Quinault) sans que cela change très significativement les caractères de cette autre collection, nous allons chercher tous les paramètres selon lesquels ces collections (par exemple C et Q) s'opposent très radicalement. Cette recherche se fait selon un calcul, le calcul de t. Nous sélectionnons tous les paramètres pour lesquels t > 1,96.

Ensuite, nous allons prendre une pièce de Molière. Nous allons, pour chaque paramètre sélectionné, la comparer à C. Cette comparaison se fait également par un calcul de t. Si t < 1,96, il y a 95% de chances que la pièce de Molière ressemble radicalement à C, sinon il y a 95% de chances qu'elle ne ressemble pas à C.

Si les critères sélectionnés sont au nombre de cinq, et si chaque fois t < 1,96, on a cinq fois cette vérification à 95% de chances.

On effectue ensuite la comparaison de cette pièce avec Q, de la même manière.

Par construction, il est impossible que, pour un paramètre sélectionné, t < 1,96 sur C, et t < 1,96 sur Q (c'est-à-dire il est impossible que la pièce de Molière, sur un paramètre sélectionné, ressemble à la fois à C et à Q, puisque les paramètres sélectionnés ont précisément été sélectionnés parce que C et Q différaient sur ces paramètres).

Quand la pièce de Molière, sur les cinq paramètres, a été trouvée ressembler radicalement à C, on l'attribue à Corneille. Quand, sur les cinq paramètres, elle a été trouvée ressembler radicalement à Q, on l'attribue à Quinault. Dans les autres cas, on la place dans la classe des non-attribuables, classe qu'on s'efforcera de réduire ensuite.

Ce qui précède, la dernière phrase exceptée, est à peu près un commentaire de la page 10, laquelle est très dense, et étonnamment inintelligible pour un non-mathématicien.

2° Critique.

1. Quels sont-ils, ces paramètres sélectionnés par le fait que C et Q sur eux s'opposent ? Donnons-les, et donnons les valeurs sur C et sur Q. Pour une raison qui apparaîtra rapidement, donnons aussi X7. Dans le tableau ci-dessous, les définitions doivent être précédées de « nombre de… », et suivies de « par phrase ».

TABLEAU 1.

	C		//	Q	
Définition	moyenne	écart-type	//moyenne	écart-type	différence des moyen.
X2 : propositions	1,80	0,89	2,17	1,41	20%
X21 : verbes conjugués	1,76	1,01	2,14	1,52	21%
X31 : sujets	1,47	0,98	1,91	1,46	30%
X32 : sujets pronominaux	1,08	0,95	1,39	1,05	29%
X4 : propositions coordonnées	0,53	0,94	1,22	1,43	130%
X7 : propositions subordonnées	0,50	0,70	0,44	0,78	13%

(Lecture de la première ligne du tableau : Dans C, par phrase, le nombre moyen de propositions est de 1,80, avec un écart-type de 0,89. Dans Q, ce nombre est de 2,17, avec un écart-type de 1,41. L'écart entre les deux moyennes est de 20% : dans une phrase de Q, il y a en moyenne 20% de propositions de plus que dans une phrase de C).

En somme, dans une phrase de Q, il y a plus de propositions que dans une phrase de C, il y a plus de verbes conjugués, il y a plus de sujets, plus de sujets pronominaux, et il y a beaucoup plus de propositions coordonnées. Les quatre premiers traits évoqués, sont, naturellement, si hautement corrélés qu'ils ne sont presque qu'un seul trait : quand, dans un corpus, il y a, par phrase, plus de propositions, il tend à y avoir aussi plus de verbes conjugués, et plus de sujets, et plus de sujets pronominaux. Nous touchons là un point

crucial de la méthode adoptée, qui néglige une règle essentielle en statistique, où l'on recommande de veiller à faire des tests sur des faits indépendants les uns des autres…

Autrement dit, en moyenne, dans une phrase de Quinault,

- . il y a plus de propositions que dans une phrase de Corneille
- . cette augmentation est due à la coordination (et non à la subordination, puisqu'il y a au contraire un tout petit peu moins de subordonnées chez Quinault que chez Corneille, c'est ce qu'on voit à la ligne X7).

Et donc, ce que la réflexion sur ce tableau nous apprend, c'est que dans ce tableau la différence Corneille-Quinault se réduit à un plus grand nombre de coordonnées chez Quinault, couplée à un nombre voisin de subordonnées, et à un nombre probablement voisin de phases simples (paramètre non mesuré) : de cette unique différence et de ces deux ressemblances, découlent les quatre autres différences.

Quand on réfléchit maintenant que c'est seulement sur ces paramètres qu'une différence décisive a été trouvée entre C et Q, on peut tirer une conclusion importante : Quinault et Corneille écrivent leurs comédies de façons très voisines (puisque aucun des 46 autres paramètres n'a fourni de quoi discriminer), mais Quinault, par la coordination de propositions, fait des phrases contenant plus de propositions que Corneille, étant donné qu'ils sont voisins par le nombre de subordonnées, et probablement ressemblants par le nombre de phrases simples.

2. Dans ce qui suit, par ''analyse syntaxique pétersbourgeoise'', nous voulons dire l'analyse syntaxique pratiquée avec les 51 paramètres du tableau donné en Première Partie. Par ''analyse syntaxique'', nous voulons dire ''analyse syntaxique en général''.

- a) Globalement, donc, on a trouvé que C et Q ne diffèrent décisivement que sur un seul trait syntaxique.
- b) C'est dire à peu près que l'analyse syntaxique pétersbourgeoise, dans ce cas C-Q, a prodigieusement échoué : ne parvenir à ne trouver qu'une seule différence décisive entre deux corpus, c'est dérisoire.
- c) Cela ne signifie pas que l'analyse syntaxique pétersbourgeoise échoue toujours ; cela ne signifie pas non plus que l'analyse syntaxique échoue toujours. Mais enfin, cet échec est mauvais signe, pour l'analyse syntaxique pétersbourgeoise en particulier, et pour l'analyse syntaxique en général.
- d) Faisons toucher du doigt cet échec. Supposons que je désire savoir de certains hommes s'ils sont des Auvergnats ou des Normands. Comment ferais-je ? Je commencerais par chercher des traits sur lesquels Auvergnats et Normands diffèrent. Et par exemple, je poserai à un échantillon d'Auvergnats et à un échantillon de Normands 51 questions : 25 questions de géographie locale auvergnate, et 26 questions de géographie locale normande. Si mes questions sont bien faites, il est probable que sur dix ou vingt ou trente questions, j'aurai d'importants contrastes entre Auvergnats et Normands. Pour parvenir à ne trouver qu'une seule question sur lesquels Auvergnats et Normands diffèrent, il faut que je n'aie pu poser que des questions prodigieusement non significatives.
- e). Quelle est la cause de cet échec de l'analyse syntaxique pétersbourgeoise ? On peut penser qu'elle est due au fait que Q et C sont trop proches syntaxiquement. On peut penser aussi qu'ils sont, peut-être, chacun trop variés (pas assez de monotonie syntaxique dans leurs phrases). Ou bien l'on peut penser à des faiblesses plus générales, peut-être à des faiblesses inhérentes à toute analyse syntaxique possible (la

syntaxe serait un très mauvais moyen de différencier les oeuvres, ou les auteurs, ou les oeuvres et les auteurs).

- f) Notre exemple nous permet de voir un moyen éventuel de faire progresser l'analyse syntaxique : si je veux opposer Auvergnats et Normands, je leur pose des questions de géographie locale auvergnate, et de géographie locale normande. Si les questions que je pose sont les mêmes que celles que je poserais à des Picards et à des Limousins, alors ces questions vont donner des résultats extraordinairement pauvres : Auvergnats, Picards, Normands et Limousins seront très difficiles à opposer sur des questions du type ''Quelle est la capitale de la Chine ?'', ou bien ''Aimez-vous La Joconde ?''. Or c'est ce que font les Pétersbourgeois : ils interrogent C et Q en leur posant les mêmes questions qu'ils poseraient s'ils interrogeaient Anouilh ou Ionesco.

3. Critique proprement dite.

- a) Même si l'on adopte le point de départ de nos auteurs, à savoir la théorie Corneille-Quinault, rien ne nous dit que Corneille, quand il écrivait du Molière, ne s'est pas écarté syntaxiquement de C, que Quinault, quand il faisait du Molière, ne s'est pas écarté syntaxiquement de Q. Quand un policier s'adresse à son chef, ou à son subordonné, il ne se comporte pas exactement de la même façon. Les comédies de Molière sont très différentes des comédies de Corneille, et de celles de Quinault. Pourquoi ne seraient-elles pas un peu différentes syntaxiquement aussi ? La syntaxe utilisée dans une comédie, ce n'est certes pas les empreintes digitales, ou le style de l'écriture manuelle, ou le timbre de la voix.

- b) D'autant plus que nulle part (y compris dans la thèse russe, à moins que ce ne soit fait dans un morceau de la thèse russe non accessible sur internet, mais on ne voit pas pourquoi les auteurs n'auraient pas rendu accessible au public cet important travail) les Pétersbourgeois n'examinent la syntaxe de chaque pièce appartenant à C, de chaque pièce appartenant à Q. Rappelons que C et Q ce sont respectivement *l'ensemble* des comédies de Corneille, et l'ensemble de celles de Quinault. Rien ne dit que, si l'on examine pièce à pièce C, on ne trouvera pas pour telle ou telle pièce une ressemblance décisive avec Q ($t < 1,96$ pour les cinq paramètres), rien ne dit que si l'on examine Quinault, de même on ne trouvera pas une ressemblance décisive avec C…

- c) On aurait d'ailleurs pu mener ce travail aussi sur les tragédies de Corneille, sur les tragédies de Quinault. Sur leurs autres écrits. Qui peut croire que le nombre de propositions coordonnées est une constante chez un homme ? On reconnaît X à sa voix dans toutes les langues qu'il parle. Quand il parle français, il utilise en général le mot « auto » quand d'autres diraient « voiture ». Dans toutes les photos qu'on a de lui aux différents âges, l'écart entre les yeux est le même. Mais le nombre de propositions coordonnées, qui peut penser cela comme une constante ?

IV. DÉTAIL DU RESTE DE LA DÉMARCHE, EXPOSE EN DEUX SOUS-PARTIES, CHACUNE ASSORTIE D'UNE CRITIQUE.

1° Première étape de l'attribution dite déterministe.

On se souvient que, pour discriminer C et Q selon Xn, les tests de Student étaient effectués sur 100 phrases, et que nous avions justifié cette pratique en donnant l'exemple d'un sondage d'opinion, qui est valide en testant 1 000 personnes, que ces mille personnes appartiennent à une population de cent mille, d'un million ou d'un milliard d'individus. (Rappelons que, pour discriminer selon X4, on prend 100 autres phrases que les 100 phrases utilisées pour discriminer selon X2. Les phrases sont tirées aléatoirement. Tout ceci vise à augmenter la fiabilité).

Nous avons dit que, quand on s'interroge si l'on doit rattacher Mn à C ou Q selon Xn, on utilise aussi le test de Student. Mais on en fait une utilisation différente de l'utilisation précédente. Dans cette nouvelle utilisation, on ne peut pas prendre 100 phrases. Plus une pièce Mn comporte de phrases, plus l'échantillon sur Mn – qui doit être le même que sur C – doit être important pour un X. En d'autres termes : *Tartuffe* comporte plus de phrases que *La Pastorale comique*, pour comparer *Tartuffe* et C selon X2, il faudra effectuer la comparaison en prenant un nombre de phrases plus grand dans *Tartuffe* que pour comparer *La Pastorale comique* et C selon X2.

Et, pour comparer *Tartuffe* et C selon X4, il faudra prendre un échantillon d'une taille différente de celui qu'il faudra prendre pour comparer *Tartuffe* et C selon X2. Cependant, pour simplifier, les auteurs choisissent, pour chaque pièce de Molière, de prendre, pour X2, X4, X21, X31, X32, des échantillons de même taille, la plus grande.

Exemple : pour *Tartuffe*, l'échantillon devrait être (page 48) :

pour X2, 186 phrases
pour X4, 669 phrases
pour X21, 233 phrases
pour X31, 296 phrases
pour X32, 330 phrases.

On prendra donc à chaque fois 669 phrases.

Le tableau 9, page 50 à 53, donne, pour chaque pièce de Molière, pour chaque paramètre, non seulement moyenne et écart-type, mais aussi taille de l'échantillon (qui est unique pour chaque pièce, donc).

Ce calcul de la taille des échantillons occupe un nombre considérable de pages de l'article.

Nous venons d'évoquer le tableau 9.

Les auteurs devraient ensuite nous donner, pour chaque M, pour chaque Xs, t avec C, et t avec Q (c.a.d., pour le paramètre X, telle pièce de Molière ressemble-t-elle à C, ou à Q, ou ni à l'un ni à l'autre ?). Très bizarrement, ce tableau manque, lacune du traducteur à n'en pas douter. Faisons confiance aux auteurs : il y a des M pour lesquels t de C < 1,96 pour chaque Xs. En cas, les auteurs parlent d'attribution déterministe, par opposition à probabiliste (plus précisément : ils sont dans ce que nous avons appelé la première étape de l'attribution déterministe). Ces M sont quatre : 2M, 5M, 6M et 9M. (Le tableau 9, page 50, nous apprend que 2M est *Le dépit amoureux*, etc.). Il n'y a aucun M à attribuer à Quinault.

2° Critique de la première étape de l'attribution déterministe.

1. En résumé, quatre pièces de Molière ressemblent à C selon les paramètres Xs. Si quelqu'un est de ceux qui croient à la théorie Corneille-Quinault, et si ce quelqu'un en plus croit que Corneille écrivant pour Molière écrit syntaxiquement comme C, qu'apporte ce résultat ?

On se souvient que le paramètre décisif est le paramètre X4. Donnons les moyennes. Les quatre pièces déclarées ressembler à C sont à gauche du vide central.

	C	2M	5M	6M	9M		1M	3M	4M	7M	8M	10M	11M	12M	13M
X4	0,53	0,54	0,49	0,51	0,55		0,79	0,44	0,58	0,47	0,63	0,48	0,41	0,39	0,50

Dès lors, nous dirons que celui qui partage les deux postulats des Pétersbourgeois aurait effectivement tendance à croire que ceux-ci ont distribué correctement quatre pièces de Molière sur l'auteur Corneille.

2 Nous avons, dit, que, secondairement, la procédure donnait des renseignements sur la proximité de Mn et C, de Mn et Q. Qu'en est-il de ces renseignements ?

Considérons le tableau 9, sur X4, et sur X2 (nombre de propositions), auquel nous rajouterons les valeurs pour de C et Q. On verra sous quel ordre nous avons groupé les pièces de Molière. On voit que C, sur X4 et sur X2, est inférieur à Q.

	Sur X4		Sur X2	
	Moyenne	écart-type	moyenne	écart-type
C	0,53	0,94	1,80	0,89
Q	1,22	1,43	2,17	1,41
Groupe 1 : proche de C sur X2 et sur X4 : attribué à C.				
2M	0,54	0,98	1,97	1,31
5M	0,49	0,94	1,91	1,35
6M	0,51	0,94	1,90	1,05
9M	0,55	1,00	1,85	1,23
Groupe 2 : inférieur à C sur X2 et sur X4. Non-attribué : *inférieur à intervalle C,Q*.				
3M	0,44	0,90	1,65	0,93
11M	0,41	0,85	1,74	0,94
12M	0,39	0,91	1,60	1,00
Groupe 3 : sur X2 et X4, intermédiaire entre C et Q. Non-attribué : *intermédiaire*.				
1M	0,79	1,20	2,06	1,52
Groupe 4 : > C sur X2, > Q sur X4 : Non-att. : *intermédiaire/supérieur à intervalle C,Q*.				
4M	0,58	1,00	2,40	1,48
8M	0,63	1,04	2,27	1,45
Groupe 5 : <C sur X2, interméd. sur X4. Non-att. : *inférieur à intervalle/intermédiaire*.				
10M	0,48	1,95	1,94	1,20
Groupe 6 : proche de C sur X02 et X04, mais non attribué.				
7M	0,47	0,96	1,86	1,18
13M	0,50	0,94	1,86	1,12

Comme nous l'avons dit plus haut, par phrase, C est inférieur à Q sur le nombre de propositions coordonnées (X4), ce qui entraîne qu'il lui est également inférieur sur le nombre de propositions (X2), ce qui à son tour entraîne qu'il lui est inférieur sur le nombre de verbes conjugués (X21), le nombre de sujets (X31) et le nombre de sujets pronominaux (X32).

Sachant cela, comment interpréter ce tableau ?

- Dans le Groupe 1, on trouve les quatre pièces de M qui sont proches de C pour les deux paramètres X4 et X2, et donc ressemblantes à C.

Il n'y a pas de pièces ressemblant à Q sur ces deux paramètres, et donc attribuées à Quinault.

Tous les autres groupes rassemblent donc les pièces non-attribuées.

- Dans le Groupe 2, les trois comédies qui, pour les deux paramètres, sont inférieures à C.

Il n'y a pas de pièces, qui pour les deux paramètres, sosnt supérieures à Q.

- Dans le Groupe 3, les pièces non-attribuées car intermédiaires pour les deux paramètres. Ce groupe ne comporte qu'une seule pièce.
- Dans le Groupe 4, les deux pièces intermédiaires pour X2, mais supérieure à Q sur X4 : elles ont dû avoir tellement peu de subordonnées, ou de propositions simples, que leur supériorité à Q en coordonnées n'a pas suffi à les doter de nombreuses propositions.
- Dans le Groupe 5, les pièces intermédiaires pour X4, mais inférieures à C sur X2 : elles ont dû avoir tellement peu de subordonnées, ou de propositions simples, que leur position moyenne en coordonnées n'a pas suffi à les doter de nombreuses propositions. Ce groupe ne comporte qu'une seule pièce.

On voit que ces deux groupes témoignent du même phénomène.

- Le Groupe 6 ne peut pas se comprendre par ce seul tableau.

Il regroupe deux pièces pour lesquelles X31 et X32 y ont des valeurs trop fortes par rapport à C. Donc, excès de sujets et de sujets pronominaux par phrase, malgré le même nombre de propositions par phrase que dans C. Sur X31 et X32, elles sont donc intermédiaires entre C et Q, ce qui les met, avec le Groupe 3, dans un regroupement des intermédiaires.

3. Que tirer, donc, de ce tableau, quant à la proximité entre M, d'une part, Q et C d'autre part ?

Faute d'avoir les valeurs de paramètres pour chaque C, et chaque M, à peu près rien.

4. Et que tirer, quant à la proximité de chaque comédie M par rapport à C et Q ?

On peut dire ceci :

Sur les paramètres sur lesquels C et Q s'opposent, aucune comédie M n'est extérieure à l'intervalle [C, Q] du côté de Q, et aucune pièce n'est proche de Q.

Quatre pièces sont proches de C, trois pièces sont extérieures à l'intervalle, du côté de C.

Trois pièces sont intermédiaires entre C et M.

Trois pièces sont intermédiaires/incohérentes.

3° Etape suivante de l'attribution déterministe, et attribution probabiliste.

1. Etape suivante de l'attribution déterministe.

Les Pétersbourgeois cherchent ensuite à réduire le nombre de pièces non attribuées par la première étape de l'attribution déterministe (appelons qu'elle exige t < 1,96 pour chaque Xs).

Cette nouvelle étape consiste à ajouter à C les quatre pièces de Molière attribuées à Corneille par la première étape, et à recalculer sur ce nouveau C moyennes et écart-types. On refait alors les comparaisons des pièces non-attribuées avec ce nouveau C. On réitère ce procédé, tant qu'il donne quelque chose.

A la première fois, 13M a été trouvé ressembler au nouveau C, et intégré au nouveau nouveau C.

A la deuxième fois, 7M a été trouvé ressembler au nouveau nouveau C.

L'itération suivante n'a rien donné.

2. L'attribution dite probabiliste.

Puis on diminue la rigueur d'attribution (on n'exige plus que t<1,96 pour les cinq paramètres), et on attribue encore.

Ce qui conduit finalement (tableau 18, page 63-64) à attribuer 10 pièces à Corneille, 2 à Quinault, et à en déclarer une non attribuable à l'un ni à l'autre.

4° Commentaire.

La critique à la deuxième étape de l'attribution déterministe et l'attribution probabiliste n'appelle rien de spécifique. Des ressemblances sont indéniablement détectées par ces procédures. Leur assigner des taux de probabilité a un sens (''il y a x% de chances que ceci ressemble à cela'' est un énoncé pourvu de sens). Si l'on croit que Corneille et Quinault ont écrit nombre des pièces de Molière, si l'on a une foi de fer en la constance syntaxique en général de C et Q, en particulier de Corneille et Quinault faisant du Molière, ces exercices ne paraissent pas vains.

On aura noté sans surprise que, par les itérations déterministes, ils intègrent à C notre Groupe 6.

Et note sans surprise que, par les attributions probabilistes, ils y intègrent notre Groupe 2.

Est attribué à Quinault l'unique pièce de notre Groupe 3 (intermédiaire).

Sans surprise, on verra que les deux pièces non-attribuables sont notre groupe 4.

La pièce de notre Groupe 5 est attribuée à C, comme de juste.

En somme, les ressemblances que les auteurs détectent mathématiquement se voyaient assez bien à l'œil nu.

Conclusion.

I. Conclusion proprement dite.

1. Que Corneille et Quinault soient l'un et l'autre auteurs de pièces attribuées à Molière (théorie CQ-M) n'est pas le résultat de la démarche des auteurs, c'en est un des deux points de départ. La procédure mathématique qu'ils emploient est propre à distribuer des objets M sur des ensembles C et Q quand on sait déjà que ces objets M appartiennent à C ou à Q. Certes, si cette procédure échouait, elle infirmerait le point de départ. Mais qu'elle réussisse n'infirme pas la théorie inverse, à savoir celle selon laquelle Molière est bien l'auteur des œuvres qu'il a fait jouer et qu'il a publiées.

Que cette procédure réussisse n'augmente pas la probabilité pour que le premier point de départ (la théorie CQ-M) soit valide, cela empêche simplement qu'elle ait diminué : si je pense qu'il y a 999 chances sur 1 000 pour que ce soit Jacques qui est dans cette boîte, les résultats d'expériences testant la possibilité qu'il s'agisse de Pierre ou de Paul, s'ils étaient négatifs, accroîtraient un peu la probabilité que ma théorie soit juste (étant donné que le nombre des hommes n'est pas infini), les résultats positifs laissent les probabilités en l'état.

Disons que, pour ses opposants, la théorie CQ-M est extraordinairement improbable au point d'être impossible. Ce que les auteurs ont fait, c'est montrer qu'elle n'est pas, en plus, vide ou inconsistante. Enfin, pas vide ou inconsistante, si l'on adopte un deuxième présupposé des auteurs. Si l'on n'adopte pas ce deuxième présupposé, il n'ont rien dit sur CQ-M.

2. Le deuxième présupposé des Pétersbourgeois est en effet que Corneille écrivant du Molière conserve nécessairement la syntaxe moyenne des comédies de Corneille, et Quinault écrivant du Molière de même conserve nécessairement la syntaxe moyenne des comédies de Quinault. Nous avons dit à quel point ceci nous semble extraordinairement implausible.

3. Les auteurs auraient pu essayer d'étayer (un peu) leur deuxième présupposé s'ils avaient montré que chaque comédie de Molière respecte la syntaxe *moyenne* des comédies de Corneille, et que de même chaque comédie de Quinault respecte la syntaxe *moyenne* des comédies de Quinault. S'ils avaient tenté de le faire, et avaient échoué (ce qui nous semble extrêmement probable), cela aurait invalidé leur deuxième présupposé. S'ils avaient tenté de le faire et avaient réussi, cela n'aurait cependant pas prouvé sa validité, cela l'aurait simplement rendu plus plausible, ou moins implausible.

Ils ne l'ont pas fait.

4. Nous dirons donc que même celui qui est adepte de la théorie Corneille-Quinault ne devrait pas trouver significativement probante la démarche des auteurs pour distribuer les pièces de Molière sur les écrivains Corneille et Quinault.

5. Mais qu'en est-il de ceux qui partagent les deux présupposés des auteurs ?

Ceux-là rencontrent une grave difficulté : les auteurs ont en fait trouvé extraordinairement peu de contraste entre la syntaxe moyenne des comédies de Corneille et la syntaxe moyenne des comédies de

Quinault. Ce qu'ils ont trouvé de différence se ramène presque à un seul trait : alors qu'il y a, dans une comédie de Corneille, pour dix phrases, en moyenne à peu près cinq propositions coordonnées, il y en a, dans une comédie de Quinault, à peu près douze. Cette homogénéité syntaxique moyenne de Corneille et Quinault est en soi un phénomène non dépourvu d'intérêt. Faute d'examen comédie par comédie de Corneille et de Quinault, et d'examens d'autres auteurs, c'est cependant un phénomène en général très peu interprétable.

6. Cette absence de contraste a pour conséquence qu'à la procédure mathématique employée il n'a pas été fourni de ''grain à moudre'', qu'elle n'a pas reçu les données qu'il fallait pour qu'elle soit très significativement opérante (rappelons que cette procédure mathématique n'a de sens que pour ceux qui partagent les présupposés des auteurs).

7. Ceux qui partagent les deux présupposés des auteurs, peuvent-ils penser que la procédure suivie, nonobstant cette pauvreté en contrastes des données qui l'ont alimentée, est probante ? En d'autres termes : celui qui croit à la théorie Corneille-Quinault, et qui croit de plus que Corneille écrivant du Molière conserve la syntaxe moyenne des comédies de Corneille, ce qui implique *a fortiori* la croyance que chaque comédie de Corneille respecte la syntaxe moyenne de l'ensemble des comédies de Corneille, et qui croit respectivement la même chose pour Quinault, celui-là sera-t-il amené à considérer comme probante la distribution effectuée par les auteurs des pièces de Molière à Corneille et à Quinault ? C'est un point sur lequel il nous semble que la réponse devrait être positive : la procédure mathématique employée est si puissante, que, même alimentée de façon aussi extraordinairement pauvre, elle donne des résultats significatifs et même plus que significatifs : décisifs.

II. Ouverture.

1. Nous considérons comme si radicalement implausible chacun des deux points de départ des auteurs que faire des recherches supplémentaires pour invalider leur démarche ne nous semble pas impératif. Cependant, examiner comédie par comédie la syntaxe des comédies de Corneille (et, respectivement, de Quinault…) nous semble un travail intéressant en soi, et nous signalons que, si le résultat est qu'une comédie de Corneille peut, en termes de syntaxe, s'éloigner significativement de la moyenne des comédies de Corneille (et, respectivement, de Quinault), ce résultat invalidera le deuxième présupposé des auteurs. Ce serait, pour les auteurs, vu les outils qu'ils ont dû créer pour effectuer leur recherche, une tâche très facile à effectuer..

2. On a vu que l'analyse syntaxique effectuée par les auteurs a extraordinairement échoué à opposer les comédies de Corneille à celles de Quinault. Il pourrait être intéressant de se demander pourquoi. On pourrait ensuite se demander s'il aurait été possible d'adapter l'analyse syntaxique à Corneille et Quinault pour tenter d'obtenir des contrastes syntaxiques plus significatifs entre ces deux auteurs, ou si cela aurait été foncièrement impossible : intéressant champ de travail…