Le logiciel FRANSTAT (en hommage à Charles Muller)

Etienne Brunet[1]

[1] Univ. Nice Sophia Antipolis, CNRS, BCL, UMR 7320, 06300 Nice, France

Abstract

The author develops statistical functions to enable comparative studies of FRANTEXT's documents. Although this platform provides a rare wealth of documentary features, the magnitude of its statistical functions is not always satisfactory. Most of these functions are applied to a word, expression, grammatical structure or thematic environment, that is to say to a particular object that is followed across the texts. There is, however, one function that handles a complete text and produces a dictionary of its vocabulary, in alphabetical order, with associated frequencies. Applying this function concurrently to several texts enables comparisons.
A tool is still required to analyse statistically these frequency dictionaries. FRANstat fulfills this need – it provides the usual range of lexicometry tools: production of a full lexical table, calculation of specificities, intertextual distances, factor and tree based analyses.

Résumé

On se propose de développer les fonctions statistiques de Frantext en permettant l'étude comparée des textes qu'on peut y trouver. Or si Frantext fournit des informations documentaires d'une rare richesse, les fonctions statistiques n'ont pas toujours l'ampleur qu'on souhaiterait. La plupart s'attachent à un mot, à un lemme, à une expression, à une structure grammaticale ou à un environnement thématique, c'est-à-dire à un objet particulier qu'on poursuit à travers les textes. Il en est une cependant qui prend pour objet un texte entier choisi dans le catalogue et en délivre le vocabulaire dans une liste alphabétique, pourvue de fréquences. En appliquant cette fonction parallèlement à plusieurs textes, on a le moyen de les comparer. Encore faut-il disposer d'un outil apte à l'exploitation statistique de ces dictionnaires de fréquence. FRANstat est destiné à cet usage et propose la gamme des outils habituels de lexicométrie : établissement du TLE (Tableau Lexical Entier), calcul des spécificités, distance intertextuelle, analyses factorielles et arborées…

Key words : distance intertextuelle, Frantext, analyses arborées, Charles Muller, lexicométrie, statistique lexciale

Introduction

Il y a près de trente ans Charles Muller (qui vient de s'éteindre à 105 ans) avait proposé au signataire de ces lignes une démarche expérimentale où les méthodes statistiques étaient invitées à reconnaître la plume de trois écrivains à travers leurs textes poétiques, dramatiques et romanesques. Comme ces auteurs avaient été choisis de la même époque et de la même école, la variable temporelle était écartée. Et comme on avait isolé les fréquences hautes, donc des mots grammaticaux, l'incidence du thème était aussi neutralisée. Ne restait que le croisement de deux variables: le genre et la personnalité de l'écrivain. Muller avait choisi tous les paramètres de l'expérience et mon rôle était de la réaliser. Je lâchai la bride à la machine qui ne tarda pas à conclure qu'on avait affaire à trois auteurs, un poète, un dramaturge et un romancier. Le genre l'emportait sur la personnalité propre et les distances intra (entre genres chez le même écrivain) sur les distances inter (entre écrivains). Cette constatation ruinait les espoirs qu'on pouvait attendre des méthodes quantitatives pour l'attribution ou la datation de textes douteux, dès lors que le genre littéraire entrait en ligne de compte.

Mais en trente ans la discipline a évolué, ses méthodes ont gagné en puissance et en précision et le champ des données disponibles s'est élargi à l'infini, jusqu'à atteindre 100 milliards de mots dans *Google Books* pour le seul domaine français. On se propose donc de reprendre et de prolonger l'expérience de Muller avec des outils modernes.

Il importe tout d'abord de ne pas restreindre a priori le choix des mots. Muller en avait retenu 60 (en excluant les pronoms personnels, trop sensibles à la situation du discours). Avec 12 colonnes et 60 lignes, on obtenait un tableau de taille modeste, accessible aux moyens de calcul de l'époque. On considèrera que l'enquête doit s'étendre au *TLE* (ou *Tableau Lexical Entier*), c'est à dire à tous les mots d'un texte sans limite définie pour le nombre de mots (les lignes du tableau).

Pas de limite imposée non plus pour les colonnes, c'est-à-dire le nombre de textes du corpus. On doit pouvoir exercer son choix dans un catalogue très large, afin d'isoler les textes qui appartiennent à un genre particulier, à une époque donnée, à un auteur privilégié ou à quelque autre critère distinctif, l'objectif étant d'éliminer les variables indésirables pour concentrer la lumière sur celle qu'on examine.

De telles conditions n'étaient pas remplies dans deux essais que nous avons tentés dans le passé: le premier (base *AUTEURS*) opposait 75 écrivains entre eux mais la place de chacun était perturbée par le dosage différent des genres dans son oeuvre. Le second (base *CHRONO*), attaché à l'évolution chronologique de la production littéraire, n'était pas à l'abri des effets de mode qui donnent l'avantage tantôt au théâtre, tantôt au roman, tantôt aux essais. Dans l'un et l'autre cas, le genre et la chronologie s'entrecroisaient, en exerçant sur les textes et les auteurs une influence variable et peu analysable. En outre ces deux tentatives s'appliquaient à une partie seulement de *Frantext* (55 millions de mots pour *AUTEURS*, 117 pour *CHRONO*). Il y a lieu d'élargir la gamme à la totalité actuelle de *Frantext*, soit près de 300 millions de mots. Certes d'autres réservoirs linguistiques sont plus larges mais la sélection des textes n'y est pas praticable, seul étant possible le choix des mots, que ce soit dans *Google Book*, *Sketchengine* ou *Wortschatz***.** On a certes proposé avec *THIEF* un outil apte à assurer l'exploitation interactive de *Frantext.* Mais là aussi l'interrogation s'exerçait par un choix de mots dans le dictionnaire. On en obtenait la distribution selon les écrivains, selon les genres, selon les textes ou selon les époques. On pouvait obtenir l'environnement lexical ou grammatical du ou des mots proposés, mais uniquement des mots proposés.

I - L'acquisition des données et création d'une base statistique

Seule une fonction envisageait l'ensemble du vocabulaire du texte sélectionné. On s'est donc employé à généraliser cette dernière fonction en faisant un appel répété à l'option correspondante de *Frantext*. La procédure est simple et rapide et ne nécessite que quelques clics.

1 - Si l'on est habilité à interroger *Frantext* (un abonnement collectif de l'université ou du laboratoire suffit), on doit d'abord préciser le corpus que l'on veut traiter, et qu'on appelle le « corpus de travail ». La figure ci-dessous est familière à tous les utilisateurs de *Frantext.* C'est le passage obligé, avant toute autre démarche. L'option par défaut étend le choix à l'ensemble des textes. Elle ne convient pas ici, puisqu'on veut individualiser les textes ou corpus et les opposer les uns aux autres. Toute une panoplie de critères permet d'exercer son choix, aussi largement ou étroitement qu'on le souhaite.

Figure 1. Interrogation de Frantext. Choix des corpus

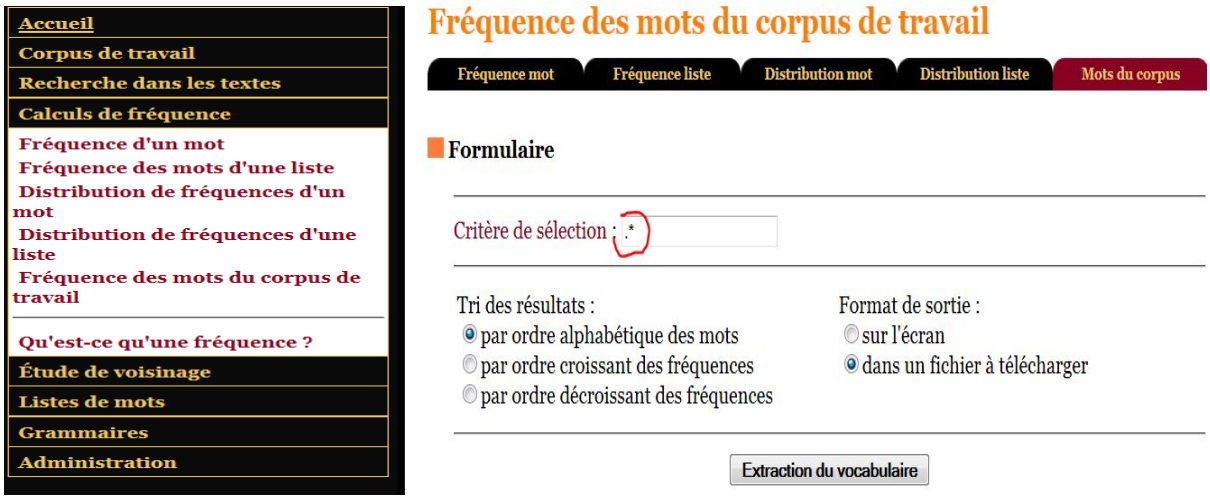
2 - La seconde étape consiste à choisir ses mots. Habituellement la sélection est limitée dans Frantext : elle porte sur un mot ou une liste de mots, ou sur un lemme, une expression , une construction, ou un environnement cooccurrentiel. Ici la demande est exhaustive : on réclame la liste triée de tous les mots du corpus considéré, avec la fréquence de chacun. On choisira donc la rubrique *Fréquence des mots* et la sous rubrique *Fréquence des mots du corpus de travail* (figure 2). Ici la panique guette les utilisateurs peu habitués à la pratique des expressions régulières. Comment exprimer à la machine la requête « tous les mots » qui serait très simple en langue naturelle. La réponse est plus simple encore et se contente de deux symboles **:** **.** * (le point signifiant un caractère quelconque, et l'astérisque une suite de zéro à n caractères quelconques).[1]

[1] La sélection ainsi définie englobe tous les éléments du texte, mots, chiffres et ponctuations. Si l'on souhaite écarter les ponctuations et les chiffres, il faut en dresser la liste (pour l'initiale) précédée du symbole de l'exclusion (^), ce qui peut s'exprimer, par exemple, dans la requête :

[^.,?!;:"%'()-_$£§&0123456789]*

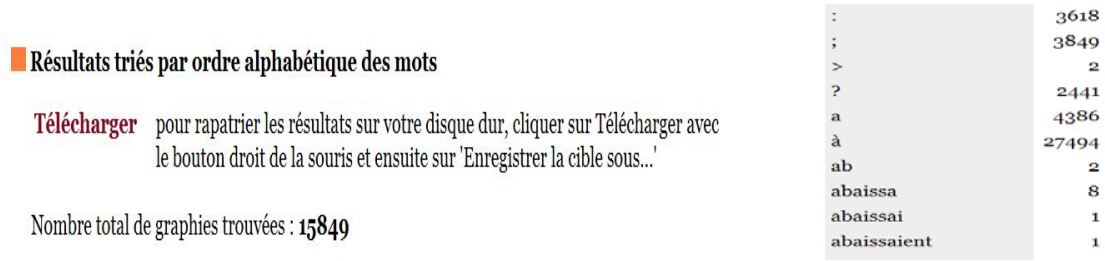
Mais la liste des initiales exclues peut varier d'un texte à l'autre et provoquer des perturbations dans la comparaison ultérieure des résultats. Il vaut mieux s'en tenir à la tradition de *Frantext* qui accepte les chiffres et les ponctuations dans le Tableau Lexical Entier et les incorpore dans le calcul de la taille du corpus – et donc dans les calculs de pondération. Il convient toutefois de prêter attention au traitement particulier des noms propres qui dans la saisie des textes à Nancy ont été pourvus de l'astérisque placé à l'initiale. Fort heureusement le logiciel de *Frantext* les a répartis à leur place dans la liste alphabétique, mais ils n'en continuent pas moins à figurer aussi derrière l'astérisque. Pour éviter de les compter deux fois la requête devrait prendre la forme :

Figure 2. Interrogation de Frantext. Choix des mots



3 - Il ne reste plus qu'à recevoir les résultats et à les enregistrer dans un fichier au format ANSI (figure 3). Le nom du fichier, choisi par défaut (*resultat.txt*), doit être changé et désigner de façon claire le corpus traité. Il servira d'étiquette pour la suite du traitement.

Figure 3. Interrogation de Frantext. Rapatriement des résultats



4 - Quand pour chaque texte, grâce au bouton *Frantext*, on a obtenu son dictionnaire de fréquences (ou TLE), le bouton *Création* assemble les données dans un index général où chaque forme est accompagnée d'autant de sous-fréquences que le corpus compte de textes. Dès lors les fonctions statistiques du logiciel HYPERBASE s'appliquent, qu'il s'agisse de spécificité, de richesse lexicale, d'évolution, ou de distance intertextuelle. Et l'éventail des outils graphiques est disponible: simples histogrammes ou analyses factorielles ou arborées. Voir les fonctions proposées dans le menu de la figure 4.

[^*]*

Mais cette précaution n'est pas nécessaire, le traitement subséquent de *Franstat* assurant automatiquement le nettoyage.

Figure 4. Le menu de Franstat

Si on compare ce menu à celui du logiciel *Hyperbase*, l'offre apparaît cependant beaucoup plus pauvre. Comme le texte est absent, toutes les fonctions documentaires sont inactives. Impossible d'établir une concordance ou de restituer un contexte. Mais certaines fonctions statistiques qui reposent sur la lecture séquentielle du texte deviennent aussi impraticables. On ne peut plus rendre compte des emplois d'un mot à l'échelle du paragraphe, ni explorer les phénomènes de répétition, de cooccurrence, d'attraction thématique ou de dépendance syntaxique. On ne dispose que d'un tableau où chaque mot requiert une ligne pour établir sa répartition dans l'ensemble[2].

Figure 5. Les deux premières lignes du TLE (exemple de Proust)

43485 a , 1 9253 2 916 3 2521 4 1143 5 1288 6 2300 7 1637 8 554 9 628 10 2030 11 1733 12 579 13 2064 14 2589 15 2785 16 772 17 548 18 645 19 805 20 381 21 492 22 3027 23 1115 24 3680

120842 à , 1 20753 2 2696 3 6615 4 2437 5 3640 6 5304 7 5831 8 2161 9 2226 10 7947 11 5166 12 1969 13 7211 14 9379 15 7549 16 2135 17 1338 18 1896 19 3420 20 1860 21 1545 22 7367 23 3356 24 7041

Cela suffit pour dessiner une courbe. Pour obtenir des résultats plus élaborés, on procède à des manipulations sur les lignes et les colonnes. On obtient alors un sous-tableau, offert aux méthodes multidimensionnelles.

II – L'étude d'un genre : la Correspondance dans *Frantext*

Comme l'extraction des données se réduit à des fréquences, obtenues par simple consultation des index et transmises sans la moindre attente, on n'a pas à ménager les ressources du serveur et l'on peut sans scrupule aborder de grands corpus. Dans l'exemple de Proust, l'extraction demande seulement une seconde, et la transmission une seconde seconde. Le fichier obtenu pour toute la *Recherche* dépasse à peine 500 ko, pour 42667 mots différents et 1225316 occurrences, soit 10 fois moins que la taille du texte et 1000 fois moins qu'une concordance complète. Pour une illustration de la méthode, on s'est donc orienté vers un genre largement représenté dans les données de *Frantext* : la correspondance. A priori on

[2] La lecture doit se faire par binômes, dont le premier élément désigne le rang du texte dans la série, c'est-à-dire la colonne du tableau, et le second la sous-fréquence à inscrire dans cette colonne. Ce mode de lecture fait gagner de la place en évitant de dévider un chapelet de zéros quand un mot est peu fréquent et donc absent dans la plupart des textes.

pouvait espérer une certaine solidité dans l'attribution des étiquettes, car des critères formels (un destinataire, un message, une date, une signature) permettent de décider si l'on a affaire ou non au genre épistolaire. En réalité le codage générique établi à Nancy n'a pas été automatique, car il y a des cas où l'appartenance au genre est douteuse ou ambiguë. Que décider quand les lettres n'ont pas de correspondant, quand elles sont fictives, quand elles s'enchaînent dans un exposé pédagogique, un récit de voyage ou une action romanesque ? En beaucoup d'occasions des titres présentés comme des recueils de lettres n'ont pas été retenus dans le genre épistolaire et ont été versés au chapitre des romans[3], des essais ou traités[4], des pamphlets[5], des récits de voyage[6], des mémoires[7] ou même de la poésie[8]. Plus de 40 textes ont été éliminés comme faussaires, soit plus de 2 millions d'occurrences. Il reste alors une masse de 11 millions à franchir le sas sans encombre.

1 - Sans trop s'interroger sur le bien-fondé des étiquettes maintenues sous la rubrique « Correspondance », on a soumis cette masse au traitement statistique de *Franstat.* Un calcul de distance intertextuelle montre, comme on pouvait s'y attendre, que le genre épistolaire évolue avec le temps. Les écrivains s'orientent de la droite vers la gauche selon un croissant qui suit grossièrement l'ordre chronologique. À chacun des 44 textes ou sous-ensembles une nuance est attachée entre le bleu et le rouge, le rouge désignant le début de la chronologie et le bleu la fin. L'œil reconnaît sans peine un mouvement progressif qui passe de l'un à l'autre à travers le camaïeu des couleurs violacées. Mais il y a des échappées erratiques comme celle de Du Camp ou de Gobineau, en haut à droite. Et l'analyse souffre de graves déséquilibres : certains auteurs sont peu représentés et n'ont que 10000 occurrences, comme d'Alembert, Bernanos ou Charles-Joseph de Ligne. D'autres au contraire, comme Sand et Flaubert, en ont cent fois plus, et accaparent le tiers de l'ensemble, au point de rendre nécessaire une partition en décennies de leur correspondance. Enfin le doute naît sur la nature de l'évolution : s'agit-il d'une mutation du genre, de changements dans le contenu ou du renouvellement de la langue ?

Or l'examen des spécificités montre que l'originalité et l'éloignement des premiers textes tiennent principalement aux variations de l'orthographe. Les mots les plus spécifiques appartiennent au magasin des graphies anciennes, comme le montre la série des têtes de

[3] Ainsi en est-il des *Lettres portugaises* de Guilleragues, des *Lettres persanes*, de Montesquieu, des *Lettres angloises…* de l'abbé Prévost, des *Lettres de mon moulin* de A. Daudet et de bien d'autres lettres romanesques publiées par Crébillon, Graffigny, Riccoboni, Dorat, Léonard, Charrière, Latouche. Ce genre littéraire illustré par le *Werther* de Goethe a eu beaucoup de succès en France avec les *Liaisons Dangereuses*, la *Religieuse*, la *Nouvelle Héloïse*, *Mademoiselle de Maupin*, etc…

[4] Les *Lettres philosophiques* de Voltaire sont classées parmi les essais, ainsi que les *Lettres juives* du marquis d'Argens, les *Lettres de Dupuis et Cotonet* de Musset et quelques signatures de Bougainvilliers, Aubert de la Chesnaye des Bois, Palissot de Montenoy, Roland de Charbonnières, Horace Monod. En réalité une mode a été lancée qui consiste à publier une proposition ou un réquisitoire sous la forme d'une *lettre publique* dont le public prend connaissance en même temps que le destinataire, l'éditeur se substituant à la poste.

[5] Derrière l'exemple illustre des *Provinciales* de Pascal, on relève, un peu abusivement, les *Lettres écrites de la montagne* de Rousseau.

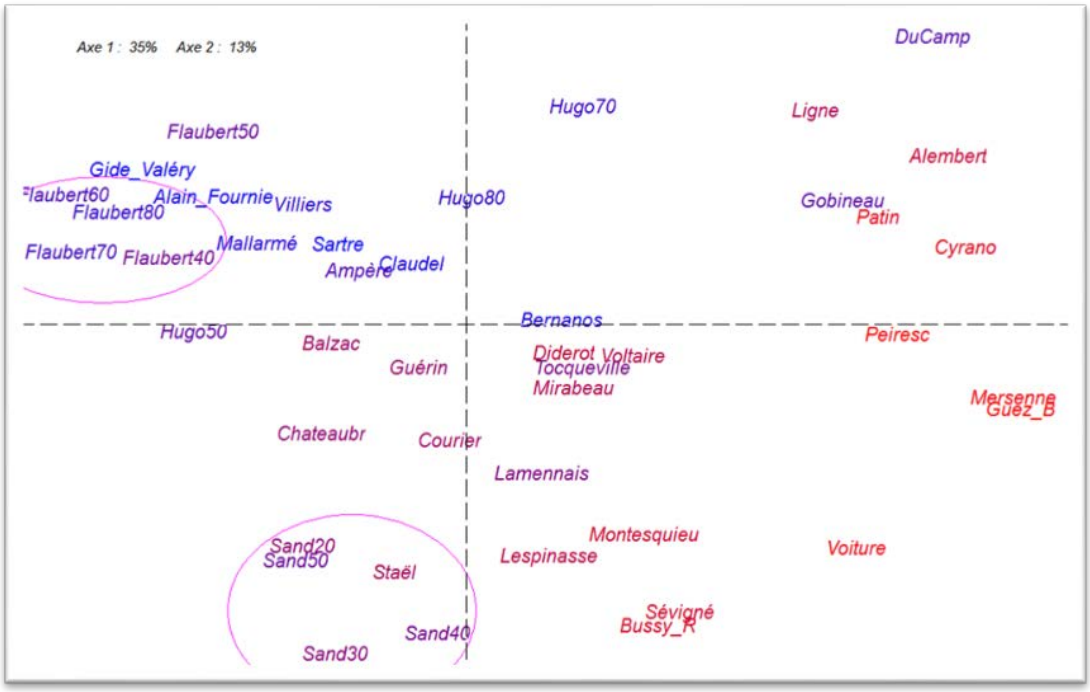
[6] *Voyage littéraire de la Grèce* de P.A. Guys. On aurait dû ajouter *Le Rhin* de Victor Hugo.

[7] *Lettres d'amour en héritage*, de Lydia Flem.

[8] *Lettres d'hivernage* de L. Senghor.

liste *: ny* chez Guez de Balzac, *luy* chez Peiresc, *vostre* (Voiture), *icy* (Patin), *estes* (Cyrano), *sçai* (Bussy-Rabutin). La stabilité de l'écriture n'est acquise qu'à partir de Madame de Sévigné. On a donc renoncé à poursuivre plus avant un corpus que rendaient fragile l'instabilité de l'orthographe et de trop grands écarts de taille.

Figure 6. La distance intertextuelle. Analyse factorielle (corpus brut de Frantext)

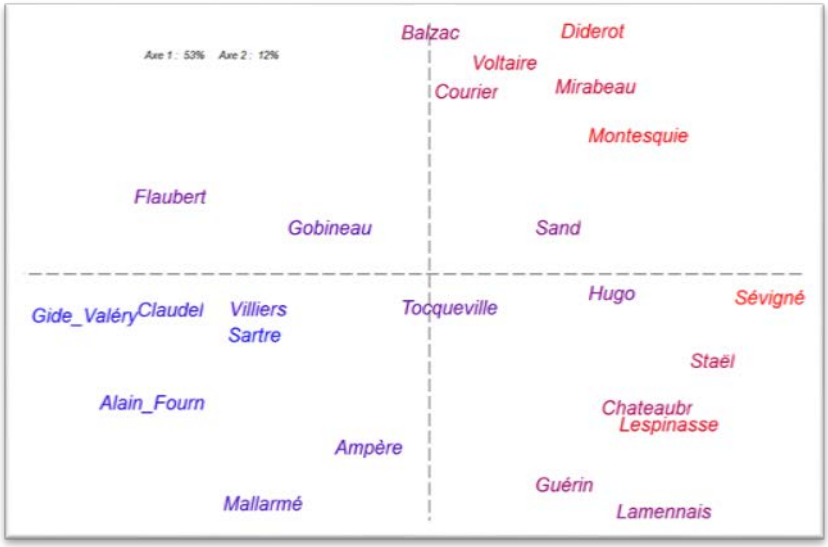


2 – Dans une deuxième approche, le corpus épistolaire élimine les sept premiers textes et commence avec Madame de Sévigné dont l'édition retenue est exempte de graphies anciennes. Mais certaines survivent encore dans les éditions de Diderot, de Mlle de Lespinasse et de Mirabeau (imparfaits en *oit,* substantifs en *ans, ens, ems* au lieu de *ants, ents, emps*). En outre un examen plus poussé du texte de Du Camp montre qu'il s'agit d'un récit de voyage en terre nordique, analogue à celui que le même auteur a publié plus au sud sous le titre *Le Nil, Égypte et Nubie.* On l'a donc exclu pour la même raison que le *Rhin* de Hugo. Exclus également trois corpus trop peu fournis : ceux de d'Alembert, de Ligne et de Bernanos. Pour la raison inverse on n'a conservé qu'une fraction des corpus envahissants de Sand, de Hugo et de Flaubert : celle qui concerne la décennie 1850. Restent en piste 24 sous-ensembles, sur les 44 proposés au départ.

La figure 7 comme la précédente est une analyse dite « de correspondance », ce qui n'a rien à voir avec le genre épistolaire. Elle rend compte pareillement de la distance intertextuelle, c'est-à-dire des liens étroits ou distendus que les textes ont entre eux quand on considère leur vocabulaire. Le contraste entre les auteurs en rouge à droite et ceux qui virent au bleu à gauche s'explique pareillement par la chronologie. L'épuration des données n'a guère changé le résultat, non plus que le recours à un calcul différent. Car ici nous retrouvons la méthode de Muller[9], appliquée aux basses fréquences.

[9] Cette méthode, fondée sur la loi binomiale, est proposée par Muller dès 1968 dans son *Initiation à la statistique linguistique*, p. 210. Plutôt qu'une distance, elle établit une « connexion » ou proximité lexicale. Privé de

Figure 7. Analyse « de correspondance » de la Correspondance. La connexion de Muller appliquée aux basses fréquences



Une fois évacuée la variation de l'orthographe, les mouvements de l'histoire imprègnent la correspondance comme une éponge, au point que les historiens s'appuient volontiers sur les témoignages épistolaires, souvent plus sûrs que les documents officiels. Certains genres comme la poésie peuvent échapper aux variations de l'actualité, parce que le stock lexical auquel ils s'alimentent varie peu : le même « soleil » éternel éclaire les poèmes du monde entier, de l'antiquité à nos jours. La correspondance au contraire est sensible aux événements particuliers, aux lieux, aux personnes, aux modes, aux questions - et donc aux mots - que l'actualité met à l'ordre du jour. Les lettres sont toujours, à des degrés divers, des éphémérides. Écrites au jour le jour, dans le désordre, et envoyées aux quatre points cardinaux, elles subissent un classement, lors de la publication, qui est le plus souvent d'ordre chronologique. Or cette trace de l'actualité s'observe en priorité dans les mots de basse fréquence[10], isolés dans le graphique 7. Ce sont d'abord des noms propres, lieux familiers ou personnes proches, parmi lesquelles l'allocutaire lui-même et le destinataire. Ces noms circulent peu d'un corpus à l'autre et on les retrouve concentrés dans la zone hautement significative des spécificités de chaque correspondance. Ainsi la liste caractéristique de Staël commence par *Staël, Narbonne, Coppet, Zurich, Lausanne* et il faut attendre à la treizième place le premier nom commun (*père*). En dehors des correspondances croisées où les

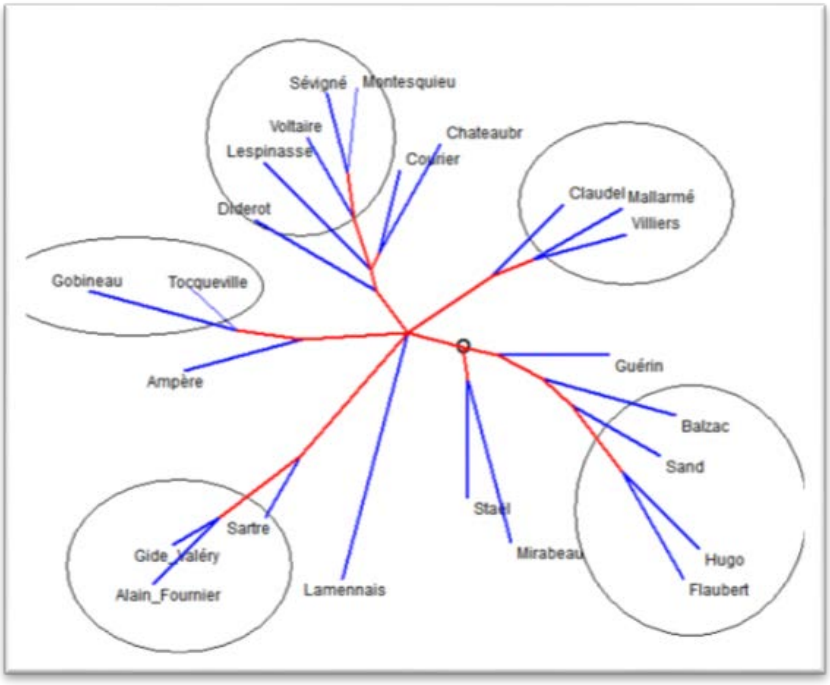
moyens de calcul, Muller n'a pu à l'époque en fournir une application à grande échelle - ce que nous avons fait à plusieurs reprises, à propos de Hugo ou du théâtre classique. Voir E. Brunet, « Une mesure de la distance intertextuelle : la connexion lexicale », in *Revue, Informatique et Statistique dans les sciences humaines*, n° 1 à 4, C.I.P.L., Liège, et E. Brunet « Muller le lexicomaître », in *Mélanges offerts à Charles Muller pour son centième anniversaire*, Conseil International de la Langue française, Paris, 2009, p.99-119.

[10] Les basses fréquences vont ici de 1 à 50. Les hapax sont pris en compte dans le calcul de Muller, comme dans celui de Jaccard, alors qu'ils ne le sont pas dans l'algorithme de Labbé.

interlocuteurs s'adressent l'un à l'autre[11] (c'est le cas de Gobineau et Tocqueville), chaque sous-ensemble est un univers fermé qui ne communique guère avec les autres sous-ensembles, dont il est séparé par l'espace, le temps ou les circonstances. L'analyse de ces individualités juxtaposées ne propose guère de regroupement exploitable. Un seul facteur emporte l'ensemble : la dérive du temps.

Mais le même algorithme proposé par Muller (figure 8) conduit à des interprétations moins triviales quand il s'applique aux hautes fréquences On y voit se rapprocher Gobineau et Tocqueville non pas parce qu'ils appartiennent à la même époque mais parce qu'ils partagent les mêmes questions sinon les mêmes réponses. On voit inversement s'éloigner des gens qui sont du même temps mais dont les discussions divergent, plus poétiques chez Claudel et Mallarmé, plus largement littéraires ou morales chez Gide ou Alain-Fournier. Quant à Flaubert il rejoint son camp naturel, près de Sand, Hugo et Balzac. Si les écrivains antérieurs à la Révolution restent groupés, ceux qui ont vécu 1789 et l'Empire, se dispersent, Mme de Staël et Mirabeau d'un côté, Chateaubriand et Courier de l'autre.

Figure 8. Analyse arborée. La « connexion » de Muller établie sur les hautes fréquences



Reste une difficulté pour approfondir l'analyse : chaque recueil n'a pas la cohérence qui serait nécessaire pour établir un profil. Suivant les destinataires des lettres d'amour s'y mêlent à des questions de travail, de voisinage, de carrière, de santé, à des considérations esthétiques ou morales, à des réflexions générales comme à des expériences vécues. Dans la plupart des cas la lettre n'a pas d'intention littéraire et ne répond pas à une poussée créatrice. Et la dispersion qu'on observe au niveau du recueil peut aussi se produire dans la même lettre où le ton peut varier comme les sujets abordés.

[11] La composition des recueils est variable sur ce point, et peut regrouper plusieurs paquets de lettres échangées d'un correspondant à l'autre. C'est le cas de la correspondance d'Alain-Fournier et Rivière, de celle de Gide et Valéry, de Claudel et Gide, et des deux Ampère, père et fils

Le genre épistolaire apparaît ainsi mou, invertébré, prêt à toute les contorsions. Mais quand on le compare aux autres genres à partir du panorama de *Frantext* il montre des caractéristiques précises et constantes dont certaines se rapprochent de la langue orale et dont rend compte la liste des spécificités reproduite partiellement dans la figure 9: domination de la première personne (*je, j', me, moi, m', ma, m*on) et, à un moindre degré, de la seconde (*vous, vos, te, t'*), installation dans le présent (*ai, a, est, suis, avez, crois, dites, aujourd'hui, présent*), le passé immédiat (*hier*) ou le futur proche (*sera, serai, aurai, écrirai, demain, prochain*), relâchement du style avec l'abondance des présentatifs neutres (*c', ce, cela, voilà, voici*), des parenthèses et des points de suspension, propension à l'hyperbole (*très, beaucoup, partout, toujours, jamais, tout, bien, fort*), propos complaisants où tout est *beau, bon* et *admirabl*e et où la sincérité peut souffrir des convenances de l'urbanité. Comme à l'oral les verbes abondent, surtout les auxiliaires et assimilés (*être, avoir, falloir, pouvoir, vouloir, aller*) et s'accommodent volontiers de la négation (*ne, n', pas, point, rien, guère*), et de la conjonction *que.* On n'insistera pas sur la thématique épistolaire qu'imposent les conventions du genre et particulièrement les clausules de politesse qui s'accumulent au début et à la fin du message.

Figure 9. Les spécificités de la correspondance

324.95	818177	109913	vous
264.24	1252134	133526	je
245.62	334	1028	nohant
217.63	2715	2697	grignan
193.05	248484	34949	ai
176.01	290	690	castor
168.28	39103	9580	lettre
165.40	110162	18247	votre
160.58	454107	48645	j'
142.36	1540412	122170	que
124.30	120934	16197	suis
119.56	106589	14486	m
118.13	5832	2342	embrasse
117.80	44116	7928	paris
116.99	17980	4492	chère
116.97	14690	3976	adieu
114.60	15209	3994	écrire
110.70	28162	5687	cher
109.75	477329	42600	me
108.78	4736	1938	écris
107.89	736	698	coulanges
107.83	309477	30055	mon
107.78	986675	76878	est
105.32	319425	30484	m'
103.08	504761	43485	a
100.33	927	737	villiers
97.90	91853	11591	(
97.83	93294	11715	)
97.12	318200	29351	bien
96.82	20917	4270	lettres
95.91	111411	13164	cela
95.13	218211	21674	ma
94.73	1019	735	claudel
94.59	228778	22421	moi
91.47	12431	2960	..
89.91	37206	5893	mme
89.34	32299	5341	crois
87.41	16605	3423	écrit
87.20	43921	6462	bonne
86.89	298766	26671	c'
85.94	997	664	montesquieu
83.54	4615	1524	envoie
80.85	1955	909	amitiés
77.88	1525	766	écrivez
76.88	57290	7206	vos
76.73	63506	7749	avez
76.07	9722	2208	santé
75.10	9928	2215	voudrais
74.84	11771	2457	reçu
74.54	32772	4790	ami
74.40	7083	1788	envoyer
73.80	868891	61965	ne
73.75	83431	9257	très
73.58	1177	631	écrirai
73.46	17052	3071	hier
73.19	10023	2185	prie
72.09	2309	903	compliments

Mais on a peu d'appui pour dépasser ces observations un peu superficielles et presque triviales[12]. Par manque de structures claires et d'informations métalinguistiques, le corpus épistolaire se dérobe à l'analyse et nous n'irons pas plus loin dans l'exploitation de ce genre littéraire[13].

[12] Les masses globulaires qu'on peut estraire d'internet et des réseaux sociaux sont souvent moins structurées encore que notre corpus. Mais, les mots étant accessibles dans leur environnement, la statistique peut s'appuyer sur leurs liaisons et dégager des lignes de force. Dans un corpus réduit comme ici à un dictionnaire de fréquences, les mots n'ont plus de rapport les uns avec les autres et les méthodes cooccurrentielles sont interdites.

[13] Qu'on ne s'indigne pas de cette dérobade. Quand un corpus est mal fagoté et une recherche mal engagée, il est vain d'aller plus avant. La valeur d'une méthode ne tient pas à la constance de ses réussites mais au signal qu'elle donne quand on va à l'échec.

III. Un exemple plus probant : les comédies en vers de Corneille et Molière

Nous retrouvons ici Muller non plus seulement comme promoteur d'une méthode mais aussi comme fournisseur des données. Il s'agit de son cher Corneille. Le texte en fut saisi à Besançon dans les années 50, sur cartes perforées, et après avoir servi de matière aux deux thèses de Muller, il fut transféré au Trésor de la Langue Française avant de servir de nouveau à la polémique suscitée par Dominique Labbé autour de Corneille et Molière. Muller n'a pas jugé utile d'entrer dans cette querelle car dès 1967 la conclusion de son étude contenait un avertissement prémonitoire : « Notre étude ne promettait ni révélations ni solutions inédites. Ne serait-elle pas de nature, plutôt, à mettre en garde ceux qui, en l'absence de renseignements historiques, attendent de la statistique lexicale des certitudes en matière de datation et d'attribution ? »[14]

Nous avons dit notre sentiment sur cette affaire[15] en regrettant qu'une interprétation abusive et péremptoire ait perverti les observations. Quand les calculs montrent la proximité de deux textes cela n'implique pas nécessairement qu'ils relèvent de la même plume. D'autres facteurs peuvent jouer, dont le plus influent est le genre. En mêlant prose et vers, comédies et tragédies, Corneille et Molière, certains rapprochements sont prévisibles, s'ils concernent des textes soumis aux mêmes contraintes, par exemple la double exigence de la versification et de l'intention comique. Quand s'exercent de telles forces, la proximité atteint ou dépasse le seuil arbitraire établi par Labbé, et cela est observé pour d'autres écrivains de la même époque, comme Quinault ou Regnard[16].

Si l'on veut comparer deux écrivains et éventuellement prouver qu'il s'agit d'un seul, il faut s'attacher à isoler un seul facteur en neutralisant les autres et soumettre les textes aux mêmes conditions. Isolons donc les comédies en vers, qui sont soumises aux mêmes contraintes de genre, et voyons si celles de Molière et de Corneille se confondent. La réponse, évidente, est dans la figure 10. Qu'elle soit fondée sur la fréquence des mots (méthodes Labbé et Muller) ou sur la présence-absence (méthodes Jaccard et Evrard), l'analyse montre une séparation radicale des textes de l'un et l'autre écrivains[17]. Pour faire bonne mesure, ajoutons qu'en suivant une autre méthode, celle que prône André Salem pour le Tableau Lexical Entier, l'analyse factorielle aboutit à une décantation tout aussi claire. Si donc Corneille a écrit les pièces de Molière, il a bien caché son jeu, en se départissant de sa manière propre, pour imiter celle de Molière !

---

[14] Ch. Muller, *Étude de Statistique lexicale. Le vocabulaire de Pierre Corneille*, Paris, Klincksieck, p. 273.

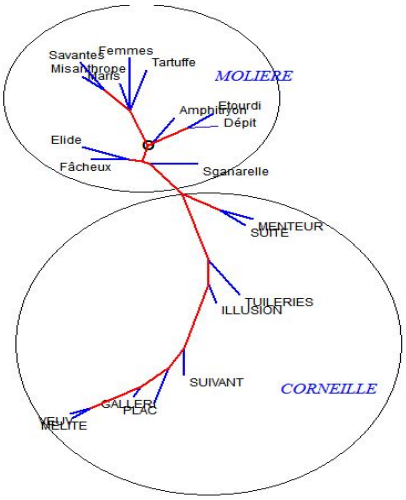
[15] Brunet, « Où l'on mesure la distance entre les distances », in revue *Texto*, mars 2004, http://www.revue-texto.net/Inedits/Brunet/Brunet_Distance.html

[16] Voir l'étude comparative de Ch. Bernet, « La distance intertextuelle et le théâtre du Grand Siècle », in *Mélanges offerts à Charles Muller*, CILF, Paris, 2009. « Il ressort de ces observations que la proximité lexicale, toute relative, entre Corneille et Molière, n'a rien d'exceptionnel. Appliqué aux textes dramatiques examinés ici, le barème proposé par Dominique Labbé conduirait à des spéculations infondées et à des hypothèses invraisemblables. » (p. 90).

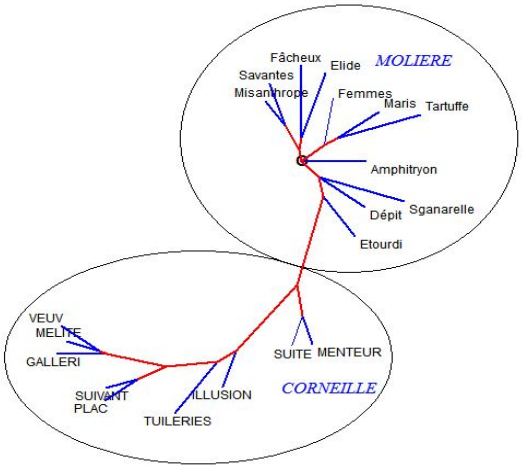
[17] Une analyse arborée, en tous points semblable, figure dans l'article de Ch. Bernet, p.97. Établie sur un nombre moindre de textes et sur un objet plus précis (les mots à la rime), elle montre pareillement la différence entre les deux écrivains.

Figure 10. La distance intertextuelle des comédies en vers. Quatre méthodes convergentes

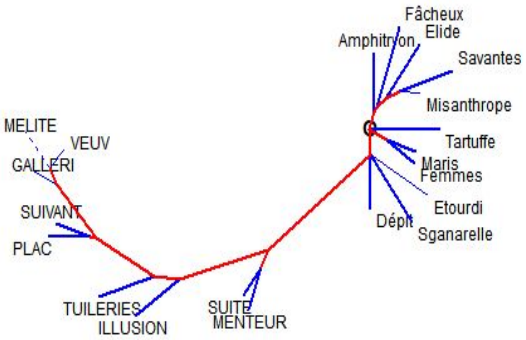
Méthode Labbé



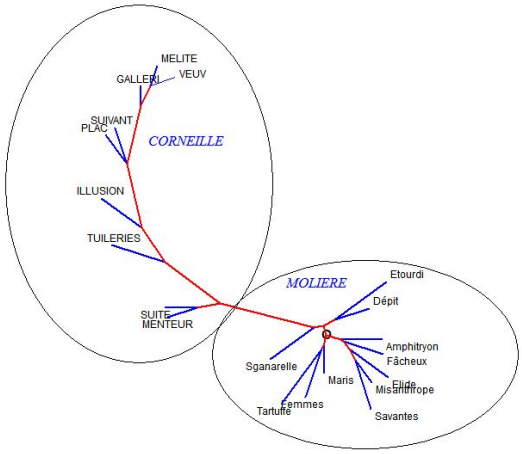
Méthode Muller



Méthode Jaccard



Méthode Evrard



Conclusion

On pourrait produire d'autres exemples de résultats obtenus avec *FRANstat*, en variant les corpus et les questions et en vantant la facilité et la puissance du traitement. Mais il convient de souligner aussi les limites de la méthode. Comme les données traitées ne portent que sur les fréquences, le retour au texte est impossible et donc tout contrôle effectif. La sécurité réside tout entière dans l'exactitude froide des calculs. Quand survient un doute, un soupçon, ou une idée d'interprétation, le texte se dérobe. En outre on n'atteint que les graphies: dans l'état actuel de *Frantext,* les lemmes échappent au recensement statistique. Enfin rien ne

permet de saisir la place des mots dans la chaîne discursive. Les éléments traités sont des pièces détachées qui interdisent toute relation de voisinage et tout calcul de cooccurrences. Pour aller plus loin, il faudrait obtenir de *Frantext* la transmission du texte. Mais le copyright et la réticence des éditeurs s'y opposent[18].

On en revient donc à la situation où se trouvait Muller quand il étudiait Corneille, il y a cinquante ans, s'escrimant avec un tableau de milliers de lignes et de 32 colonnes. Il lui avait fallu des années pour le constituer et plusieurs saisons pour l'exploiter. Aujourd'hui, avec *Frantext* et *Franstat* et une machine ordinaire, quelques minutes suffiraient. A condition d'y ajouter l'expérience et la finesse du "lexicomaître"[19].

Mais ici, parmi les remue-l'air,
On ne voit plus paraître
Le maître
Muller[20].

.

[18] *Frantex*t n'a pas vocation à la diffusion des textes, même libres de droits. Cette fonction a été sous-traitée au site *CNRTL* où l'on peut télécharger les textes dans le format *TEI.* Malheureusement le catalogue est encore trop restreint. Par exemple il ne propose qu'une vingtaine de textes dans le genre « Correspondance ». On en obtient nettement plus dans *Wikisource*.

[19] Me permettra-t-on, en guise de post scriptum, quelques mots encore à la mémoire de Muller. J'ai l'âge où il est prudent de se taire et cette communication (la dixième) est définitivement la dernière que j'aurai faite aux *JADT* (et « Tous comptes faits » le dernier ouvrage avant que m'atteigne l'âge du retrait et du silence). J'ai souvenance d'une situation semblable où se trouvait Muller quand il avait cet âge. Il venait de publier, en 1993, ce qu'il considérait comme son meilleur ouvrage, *Débats et Bilan,* et qu'il qualifiait de « DéBile » en abrégé. Comme je venais d'en faire un compte-rendu très louangeur, il tint à me raconter l'histoire de Gil Blas, secrétaire et confident de l'archevêque de Grenade, tombé en disgrâce pour avoir osé critiquer la dernière homélie de l'archevêque vieillissant. Après avoir reçu de Muller des leçons de calcul et des leçons de style, j'ai apprécié comme il convient cette leçon de vie et je n'ai plus osé après cela faire le recensement de ses autres ouvrages. Il est vrai que ses productions ultérieures portaient sur l'orthographe, où ma compétence est limitée. Il y a pourtant dans l'un de ses derniers livres, *Monsieur Duquesne et l'orthographe*, un personnage qui s'adonne gentiment à la statistique et dont le nom m'a intrigué : il s'appelle Charles-Etienne Bruller. Et Muller de commenter malicieusement ce choix : «Quand à mon double prénom, on m'a dit qu'il est dû à l'amicale rivalité de deux parrains possibles ; ils étaient amis, et du reste tous deux linguistes. » (p. 160). Je ne puis cacher un peu d'émotion devant ce témoignage, même s'il apparaît sous forme de plaisanterie. Muller aimait plaisanter et raconter des histoires, comme celle-ci que je tiens de lui et qui met en cause des mathématiciens réunis dans un colloque analogue aux *JADT.* On apprend que l'un d'entre eux vient d'être père. On lui demande « une fille ou un garçon ? » et le mathématicien de répondre « oui ».

[20] Ce quatrain est extrait de la correspondance que Muller et moi avons entretenue de longues années et qui avait souvent recours à la versification, de part et d'autre. Ce court billet a été envoyé de Saint Jacques de Compostelle, où se tenait le 19ème Congrès de Linguistique Romane. Une bonne trentaine de linguistes éminents y avaient apposé leur signature.