La textométrie en question

Bénédicte Pincemin

In Véronique Magri (dir.),
« Linguistique et traitements quantitatifs »
Le français moderne*, 2020, n°1, p. 26-43.*

[Version auteur acceptée]

La textométrie en question

Bénédicte PINCEMIN

1. Introduction

Face aux approches quantitatives de la langue et des textes, les avis sont partagés. Certains y voient une opportunité pour se doter d'outils nouveaux, au service de leur problématique de recherche. D'autres identifient rapidement un certain nombre de limites évidentes par rapport à la finesse de leur objet d'étude. Toutefois – et c'est là le propos de notre article-, pour peu qu'on y regarde de plus près, ces évidentes limites semblent quelquefois davantage relever du quiproquo. Dans ce numéro s'adressant à une large communauté de linguistes et dédié aux traitements quantitatifs, il semblait donc opportun de proposer un point actualisé sur cette question, plus spécifiquement centré sur la textométrie, rappelant de précédents argumentaires de référence en la matière (Guiraud 1960, Brunet 2016, Mayaffre 2007, 2010) et ajoutant quelques considérations originales. Il s'agit bien de tenter de répondre aux objections de fond sur les principes mêmes de la méthode, sans s'attarder sur les mésusages dus à la méconnaissance de la méthode, tel le réquisitoire critique entrepris par Carbou (2017a, 2017b). Si donc nous parvenons à dissiper l'essentiel des malentendus, il n'y aura plus de mauvaises raisons de ne pas faire de textométrie, laissant place aux échanges les plus intéressants et les plus constructifs sur les voies, quantitatives ou non, de la recherche linguistique.

Notre propos sera centré sur la textométrie, en tant que méthodologie riche et cohérente pour l'analyse de corpus textuels en sciences humaines et sociales. L'objectif de la méthode est de fournir des outils pour une lecture renouvelée et une exploration systématique d'un corpus d'étude. En linguistique, l'observation outillée peut porter tout particulièrement sur des questions contrastives (diachronie, diatopie, diastratie, etc.). Appelée d'abord *lexicométrie* (dans les années 80), s'agissant de *mesures* basées sur des décomptes de mots, la discipline a vu son nom évoluer en *textométrie* ou *logométrie* au début des années 2000, de façon à mieux rendre compte de sa capacité à considérer non seulement le lexique (le décompte et la répartition des mots au sein des textes d'un corpus) mais aussi d'autres paliers de description linguistique et textuelle (la morphosyntaxe, les structures textuelles, etc.). L'approche se caractérise principalement par l'articulation de traitements quantitatifs, simples décomptes de fréquences ou modèles statistiques adaptés, et de traitements qualitatifs, avec différents outils de navigation pour la lecture méthodique des contextes d'emploi, dite « retour au texte ». Les fonctionnalités centrales sont le calcul de concordances (pour une consultation systématique des contextes d'un motif linguistique et la mise en évidence de régularités de constructions), le calcul de spécificités (pour évaluer les variations quantitatives d'emploi entre les différentes sous-parties d'un corpus), et l'analyse factorielle des correspondances (pour obtenir une visualisation globale et synthétique des principales dimensions de contraste qui structurent le corpus) (Lebart & Salem 1994, Lebart *et al.* 2019). Le caractère endogène (l'essentiel de l'information utilisée est construite à partir du corpus lui-même), contextuel (pour la caractérisation des mots comme des textes), et contrastif de la méthode explique ses affinités fortes avec la sémantique différentielle et interprétative de Rastier (Pincemin 2012a).

2. De quelques quiproquos

2.1. L'automatisation : Pouvez-vous faire tourner le logiciel sur mon corpus ? Je regarderai les résultats pour voir ce que cela apporte.

Dans un certain imaginaire, le fait de recourir à un logiciel automatise l'analyse, et remplace le travail de recherche. Le chercheur déposerait ses données dans l'appareil, puis appuierait sur quelques boutons, et récupérerait le résultat. La qualité des résultats dépendrait du degré de perfectionnement du logiciel. Il s'agirait alors essentiellement d'une procédure technique, relevant de l'ingénierie et non de la recherche, avec quelques transformations de fichiers, de la programmation informatique, la mise en œuvre d'algorithmes pour certains issus des mathématiques. Un tel point de vue est déploré par de nombreux collègues (Jenny 1997, Chateauraynaud 2003, Demazière et al. 2006, Guerreau 2012).

On peut comprendre qu'un tel modèle soit visé par certaines sociétés de *text mining* pour leurs prestations d'intelligence économique, l'objectif étant de transformer des flux d'informations en actions. Ce n'est évidemment ni la réalité et l'intérêt du travail de recherche en sciences humaines (où il s'agit de comprendre le texte, et non de l'exploiter), ni la démarche de la textométrie. Ce que prend en charge la partie logicielle de la méthode textométrique, c'est la mémorisation complète et détaillée des textes, et l'effectuation rapide et efficace des calculs : bref, précisément la partie machinale de l'analyse, le rôle de l'exécutant. Il reste au chercheur toute la part interprétative, tous les choix d'analyse, le contrôle des opérations et l'intelligence des observations. À commencer par les données, qui ne sont justement pas données : quels textes faut-il considérer ? L'ensemble des textes sera déterminant pour les possibilités d'observation et de découvertes, puisqu'ils construisent la référence par rapport à laquelle sont mis en évidence les contrastes. Quelle version, quelle édition, quelle inscription numérique de ces textes ? Il n'y a pas de représentation pure, brute, neutre ; et donc la préparation et mise en forme des données ne se réduit pas à des opérations techniques de formatage, mais engage surtout une réflexion qui relève de la philologie numérique. Puis, quelles relations des textes entre eux, quelles contextualisations ? Là encore, les informations que choisit d'apporter le chercheur, la structure qu'il modèle pour son corpus, ouvriront certains possibles pour l'analyse (Dalud-Vincent 2011, Cointet & Parisie 2018).

Puis l'analyse elle-même suppose le choix de points d'entrée, et la détermination des types de vue et de calculs permettant de progresser. Tout calcul, aussi performant et perfectionné soit-il, ne fait pas forcément sens. Ainsi, on peut étudier les reprises de l'*Encyclopédie* de Diderot et d'Alembert à des dictionnaires de l'époque (le *Trévoux*, le *Moréri*) : mais il ne fait pas sens d'utiliser un logiciel de repérage de plagiat, de mesurer l'ampleur quantitative des emprunts, et d'en conclure à la malhonnêteté des auteurs de l'*Encyclopédie*. En effet, il faut prendre en compte les usages de l'époque, pour lesquels il était complètement assumé que ces compilations savantes exploitent les ouvrages de référence existants. Et donc la question pertinente n'est pas celle de l'existence et de la quantité des reprises, mais celle des différences -réécritures, innovations, choix- opérées par les auteurs (Leca-Tsiomis 2014), qui pourrait être assistée par d'autres types d'interrogations et de calculs d'ailleurs (recherches de motifs complexes, alignement, traits caractéristiques des ajouts et des retraits, etc.).

Quant aux sorties générées par les calculs, il s'agit de résultats de calculs, mais pas encore de résultats scientifiques ni de réponses à la problématique de recherche. Ainsi le volume des sorties n'est pas un indicateur de l'avancement et de la productivité de l'analyse.

> [Les] résultats statistiques sont plus souvent ignorés que condamnés. *Graecum est non legitur*, disent les esprits littéraires devant une formule mathématique, même enfantine. À plus forte raison renâclent-ils devant ces listes dont on ne voit jamais en même temps la tête et la queue, devant ces tableaux de nombres accueillants comme des buissons d'ajoncs et ces monceaux d'index et de concordances qu'on voit à l'abandon sur les quais de la recherche. On doit avouer que ces filets pleins de « résultats » n'ont souvent rien d'engageant. Et il arrive que leur auteur ne soit guère engagé, amassant des matériaux qu'il livre à l'état brut ou à peine transformés. Il est vrai que dans le passé l'effort était fort long pour obtenir un résultat et ce qui eût dû être l'étape initiale devenait le terme de l'épreuve. Mais les résultats qui ne sont pas interprétés par celui qui les a obtenus, les listes qui ne sont pas analysées, les concordances qui ne sont pas exploitées, tout cela appartient-il à la statistique linguistique ? N'est-ce pas plutôt seulement une prestation de service documentaire ? (Brunet 2016 : 375)

L'ampleur des choix à effectuer, la variété des paramètres à régler (Brunet 2011, 279-280, 2016, 374-375) (y compris le choix du logiciel lui-même, qui peut donner une implémentation originale de la lemmatisation – cf. méthode Reinert, dans Alceste et Iramuteq- ou d'une mesure statistique – ex. le calcul des spécificités dans Hyperbase, « réajusté » à l'échelle de l'écart-réduit pour des raisons d'ergonomie) peut inquiéter le chercheur, quant au caractère relatif et artéfactuel des résultats. À cela plusieurs éléments de réponse. Tout d'abord, la nécessité, effectivement, de bien connaître les différentes options pour faire un choix éclairé et ajuster en conséquence sa compréhension des sorties produites. Il s'agit bien entendu d'éviter un parcours d'analyse par trop empirique, testant différentes configuration de traitement et s'arrêtant lorsque la sortie « plaît » : la démarche scientifique ne prend de fait sa valeur qu'en donnant sens au lien entre données et résultats. Ensuite, un second élément de réponse tient à la mobilisation des outils de contrôle de stabilité et de validation apportés par les procédures statistiques elles-mêmes : ainsi, l'indice de

spécificité intègre le fait que la déviation est confirmée par un grand nombre de cas ou non ; ou encore, les aides à l'interprétation de l'analyse factorielle des correspondances renseignent sur la déformation créée par la projection plane ; et bien entendu, les ellipses de confiance sont un outil justement dédié à l'évaluation de la stabilité des observations (Lebart 2004). Enfin, troisième élément de réponse à cette inquiétude sur la sensibilité des analyses, la confirmation par l'expérience que les structures globales mises au jour par les calculs s'avèrent remarquablement stables (Brunet 2007).

2.2. Le traitement à grande échelle : La textométrie, c'est pour les gros corpus : pour les analyser plus vite et en économiser la lecture.

La textométrie propose la lecture renouvelée d'un corpus, en outillant l'observation systématique des mots (ou d'autres unités linguistiques : traits, motifs). À petite échelle (pour l'étude d'un sonnet par exemple), l'approche n'est pas impossible, mais son apport sera faible sinon nul, pour compléter la lecture experte d'un chercheur qui aura pu mémoriser complètement le texte. C'est lorsque le corpus embrasse un volume de textes tel que la mémorisation humaine devient plus synthétique qu'analytique, que la complémentarité devient intéressante.

> à l'échelle d'un texte […] la conscience humaine est mieux à son affaire et […] peu de secours sont à attendre de la machine, sinon la confirmation, paradoxalement réjouissante et adjuvante, de sa myopie face aux effets de sens les plus évidents. Il en va autrement lorsque de grands espaces s'ouvrent à la soif dévorante de la machine. […] Là où le lecteur parcourt en marchant l'espace littéraire, dans la succession changeante et l'effacement progressif des paysages, l'ordinateur saisit d'un coup le même espace, comme on lit une carte géographique ou stratégique, tous les points étant à plat, offerts à l'œil en même temps. En réalité les textes, si écrasés qu'ils soient par la perspective plongeante d'un observateur posté sur Sirius, acquièrent une lisibilité qu'ils n'ont pas pour l'explorateur engagé dans le maquis de la lecture. Au ras du sol, en enjambant les ruisseaux, on peut difficilement délimiter la ligne de partage des eaux. Mais d'en haut le paysage littéraire se découvre avec l'orientation des chaînes, les pentes, les ruptures et tous les mouvements de terrain produits par l'histoire. (Brunet 2016 : 371)

Toutefois, le principe même de la méthode consiste à connaître son corpus pour pouvoir interpréter les résultats obtenus, puisque le corpus est l'univers de référence que l'on se donne pour effectuer et qualifier les observations. Si la composition ou les frontières du corpus sont mal définies, les décomptes effectués ne peuvent pas être interprétés, car ils sont relatifs, mais on ignore par rapport à quoi : ce genre de difficulté est illustré par l'initiative « culturonomique » (et clairement pas textométrique) de l'application N-gram de Google (Chateauraynaud & Debaz 2012).

Pour cette même raison, le corpus ne pourra donc pas être trop gros, au sens où le chercheur n'arrive plus à se faire une idée suffisamment claire de son contenu d'ensemble (Geffroy & Lafon 1982). En textométrie, il importe d'avoir une certaine familiarité avec « son » corpus : l'avoir construit, saisi, lu, parcouru, fréquenté, sinon l'aimer (Rastier 2011 : 34). Ainsi, il apparaît clairement que la textométrie se démarque d'un des courants de la linguistique de corpus qui fait du volume des données un facteur direct de qualité, selon la maxime *Big is beautiful*, « Gros, c'est beau ».

Le corpus sera donc lu à toutes les étapes de l'analyse : lors de sa préparation ; pendant l'analyse, pour suggérer de nouvelles pistes d'investigations comme pour vérifier l'interprétation à donner à tel fait saillant ; et sans doute après le traitement textométrique, car la lecture peut être reprise et renouvelée par les nouveaux éclairages apportés par l'exploration systématique et non linéaire. L'approche textométrique n'est donc pas une opération de découverte d'un gisement inconnu d'informations, c'est un outil de re-lecture, de re-découverte, de données que l'on veut approfondir ; le corpus n'est pas une *ressource* que l'on exploite, mais une *source*, qui reste au cœur de l'analyse et qui fait référence pour toutes les interprétations (Valette 2016).

> [La logométrie] est une lecture révolutionnaire mais non destructrice qui cherche à adjoindre à la lecture *naturelle, linéaire, qualitative, traditionnelle* du texte [...], une lecture *hypertextuelle, quantitative, tabulaire, réticulaire* que seul autorise le numérique […]. (Mayaffre 2010 : 23)

> La lecture informatique ne vaut pas plus que la lecture humaine -elle vaut même plutôt moins- : c'est le renforcement d'une lecture par l'autre qui est productif. (Mayaffre 2007)

Par ailleurs, au plan des traitements, il est vrai que la statistique fournit un instrument d'observation privilégié pour les hautes fréquences : en effet, ces hautes fréquences pèsent de façon plus influente dans les calculs, et les fluctuations quantitatives d'emploi sont aussi davantage mises en valeur, car le grand nombre d'attestations renforce le jugement d'écart. Mais la méthodologie textométrique donne aussi quelques points d'accès aux singularités et aux phénomènes plus rares qui peuvent intéresser particulièrement le chercheur, avec les fonctionnalités qualitatives (consultation de contextes ciblés) mais aussi quantitatives (repérage d'exceptions et d'écarts dans les vues globales, signalement d'absences remarquables, cf. le concept de nullax au paragraphe 5).

2.3. Les mots : La textométrie travaille sur les formes graphiques : linguistiquement, il est évident que ce ne sont pas les bonnes unités d'analyse.

Si la textométrie se base sur les mots tels que figurant à la « surface » du texte, elle devrait être évidemment affectée par tous les phénomènes lexicaux bien connus des langues, de décalages entre le signifiant et le signifié : homonymie, polysémie, synonymie, implicite, etc. (Carbou 2017a).

Les unités sur lesquelles travaille traditionnellement le linguiste pour ses analyses sont construites et elles intègrent les connaissances sur la langue, pour leur délimitation, leur structuration, leur contenu sémantique, leur domaine de variation, leurs conditions contextuelles. En revanche, les unités sur lesquelles travaille traditionnellement la textométrie sont des fragments du corpus initial, des traces ou indices tirés de la matérialité graphique du texte. Elles n'ont pas une valeur directement descriptive comme les unités construites par le linguiste, mais elles peuvent se manifester dans des configurations auxquelles le chercheur peut donner un sens. Autrement dit, la textométrie introduit un niveau intermédiaire de représentation et de travail, à travers lequel elle assume un jeu, un écart contrôlé, vis-à-vis de notre perception du fonctionnement complexe de la langue. Elle reconnaît se baser sur une « indexation minimale » qui n'est qu'une approximation, qu'un point d'appui, vers la mise en évidence de structures plus complexes (Geffroy et al. 1974, Lafon & Salem 1983). Dans le même esprit, dans ses analyses thématiques assistées par ordinateur, Rastier (2001) distingue ainsi les *cooccurrents* (pointés par le calcul) des *corrélats* (à valeur sémantique, précisés et qualifiés par le chercheur).

C'est ainsi que l'on peut comprendre et expliquer le désarroi de certains linguistes devant les unités sur lesquelles une approche textométrique peut baser ses analyses. Par exemple, étudiant les débats de l'entre-deux-tours des élections présidentielles françaises, Kerbrat-Orecchioni (2017) dénonce dans les travaux d'obédience quantitative le « problème de l'identification et du regroupement des marqueurs : les listes qui nous sont proposées d'items censés véhiculer une même valeur laissent parfois perplexe ». Par exemple, « cette affirmation de Dupuy & Marchand (2011 : 144) où l'on voit figurer parmi les « marqueurs de désaccord » les items *oui* et *accord*, et c'est sur une telle base qu'est censée être démontrée l'idée que l'expression du désaccord ne ferait que croître d'un débat à l'autre : "[...] les duellistes marquent plus facilement que par le passé leur désaccord avec l'adversaire, comme le montre l'emploi croissant des formes *non*, *faux*, *mais*, *oui*, *accord* (figure 9)" ». Pour le linguiste, il est évident que les valeurs sémantiques et argumentatives de *oui* et *accord* s'opposent antonymiquement au désaccord et ne devraient pas figurer dans la même classe, et qu'il faut articuler finement la description en distinguant ces items. Mais pour le textomètre, il s'agit d'un ensemble de « traces » qui, dans le corpus étudié, vont fonctionner ensemble, en système, et correspondent à des passages qu'il interprète comme marquant globalement le désaccord. Les auteurs de l'analyse expliquent d'ailleurs ce lexique mêlé par une forme d'argumentation composée, en deux temps : « N. Sarkozy a largement recours aux formules de négation (*non*, *mais*) et au désaccord non explicite. Cette stratégie consiste dans un premier temps à affirmer son accord avec les propos de l'adversaire puis, dans un second temps, à nuancer cet accord (*oui mais*, *d'accord mais*, *vous avez raison mais*) » (Dupuy & Marchand 2011 : 144). Dans le même passage critique, à propos d'une autre liste d'indices, « (Madame, Royal, elle, est-ce que vous dites) », Kerbrat-Orecchioni note : « Dans le texte d'origine, une virgule sépare les mots *Madame* et *Royal*, mais il ne peut s'agir que d'une erreur typographique étant donné que si *Madame* peut fonctionner seul (comme terme d'adresse), il n'en est pas de même pour *Royal*. » Ainsi, l'unité linguistique qui pourrait faire marqueur, c'est « Madame Royal », mais les fragments de texte sur lesquels s'appuie l'analyse textométrique ce sont bien d'abord « Madame » et « Royal », au-delà desquels l'étude des contextes et l'interprétation peuvent effectivement établir un lien, mais dans un second temps.

Toujours dans le même propos de (Kerbrat-Orecchioni 2017), on relève une troisième fois encore cet écart entre les catégories linguistiques construites et les structures induites du corpus sur lesquelles le textomètre prend appui : « Mayaffre (2012a : 51) [...] voit apparaître au tournant des années 1980 [...] le passage d'un "discours nominal" à un "discours verbal", tous les pronoms étant curieusement comptabilisés avec les verbes et adverbes… » Ce propos de Mayaffre s'inscrit de fait dans un contexte où tant un calcul de spécificités qu'une analyse factorielle font apparaître très nettement une opposition d'usage entre verbe / pronom / adverbe d'une part, et nom / adjectif / déterminant, d'autre part, de telle sorte que selon les périodes on observe systématiquement le sur-emploi d'un des trios et le sous-emploi de l'autre. Ces groupements, étonnants pour le linguiste qui n'y reconnaît pas la relation paradigmatique du nom et du pronom, font en fait partie des observables récurrents de l'analyse textométrique (Brunet 2016 : 147-148, et Mayaffre 2012a le signale aussi à la page précédente, en note 12), il ne s'agit pas d'un commentaire un peu rapide mais bien d'un fait de corpus, sans prétendre livrer en l'état une connaissance en langue.

Le caractère intermédiaire et fruste des unités de travail du textomètre n'est pas qu'un pis-aller technique face au volume des données. À l'origine du parti pris pour les formes graphiques (assumées pour n'être qu'une approximation des « mots » du linguiste), pour la « surface » des textes, il y a la méfiance des biais et réductions induits par la projection de catégories *a priori*, de connaissances externes au corpus, qui viendraient appauvrir et perturber les observations. L'ouvrage de référence de la discipline (Lebart & Salem 1994) présente précisément la textométrie comme une solution à l'analyse des réponses aux questions ouvertes dans les questionnaires, dans la diversité de leurs formulations, dans le but d'éviter un post-codage qui standardise les réponses et neutralise des différences qui s'avèrent en fait significatives. Par exemple, le « manque d'argent » évoqué par les jeunes sans diplôme n'est sans doute pas équivalent aux « raisons financières » mentionnées par les jeunes diplômés, alors qu'un post-codage les aurait probablement ramenés à la même catégorie (Lebart & Salem 1994 : 188). De même, les pionniers de l'analyse textométrique qui militaient pour ne faire aucune lemmatisation[1] préalablement aux calculs, avaient observé que l'opposition singulier / pluriel pouvait être corrélée à des usages différents et marquer dans leurs corpus politiques une opposition abstrait / concret (travail *vs* travaux, société *vs* sociétés) (Geffroy et al. 1974).

Les progrès de l'édition numérique (avec la mise au point de standards de codage structurés comme XML) ont assoupli cette question du choix de la représentation initiale des unités du corpus, en permettant que plusieurs représentations coexistent et soient mobilisables en fonction des besoins. Ainsi, la méthode textométrique n'impose pas un type d'unités (« il faut lemmatiser », ou « on ne considère que les formes graphiques »), mais elle est vigilante sur deux points : la régularité de la représentation (cf. la norme lexicologique (Muller 1977)), et le risque de gommer des variations significatives au sein du corpus en projetant des interprétations externes, indépendantes du corpus, et surtout *a priori* (informations issues de dictionnaires, d'ontologies, ou de connaissances expertes du domaine).

Ainsi, si l'annotation des corpus peut être articulée avec l'approche textométrique (Rizkallah 2013), c'est typiquement dans un second temps de l'analyse, après avoir vérifié que les catégories introduites ne viennent pas masquer les différences qu'il s'agit justement de découvrir. C'est ainsi que le sociologue Chateauraynaud, d'abord en désaccord avec l'approche très formelle des textes qui consiste à s'appuyer sur leur matérialité graphique plutôt que sur un premier enrichissement apporté par la lecture et les connaissances du chercheur, comprend finalement que cela peut être vu comme deux moments successifs de l'analyse, fondamentalement différents mais complémentaires :

> Lebart et Salem [1994] se donnent pour but de "retarder le plus possible le saut interprétatif", l'idée étant de produire des tableaux donnant une vue objective du corpus. [...] [Ces travaux] portent à mettre entre parenthèses tous les "préjugés" et autres "prénotions", refusant de faire confiance *a priori* aux "intuitions" ou aux "interprétations" du chercheur. […] Faire table rase de cette richesse sociologique et se priver des modes d'accès privilégiés qu'elle procure aux modes sociaux et à leurs transformations, pour faire "régresser" l'analyse vers de pures associations verbales, [...] c'est demander un sacrifice trop grand au chercheur : celui de son "intelligence du social". […] Pour conserver quelques points de convergence, on dira que les résultats

[1] La lemmatisation consiste à représenter chaque mot par l'entrée du dictionnaire correspondante : l'infinitif pour les verbes, le masculin singulier pour les adjectifs qualificatifs, etc., de sorte de reconnaître comme un même mot différentes variantes flexionnelles.

> produits par la statistique textuelle constituent de bons points de départ, les associations de mots et leur distribution fournissant les premiers repères du travail interprétatif grâce auxquels on explore, graduellement, les propriétés marquantes d'un dossier. La démarche introduite par Prospero n'est donc pas directement concurrente de la statistique textuelle. Elle opère sur un autre plan, à un autre niveau. (Chateauraynaud, 2003 : 61-63)

Pour le linguiste, la question pourrait être celle d'affiner, au fur et à mesure de l'étude et de l'observation du corpus, les unités d'analyse, de sorte à passer d'un découpage systématique et rudimentaire à des unités linguistiquement plus précises. Il ne faut pas se cacher que, même assistée par des outils facilitant l'annotation par lot de cas analogues (par exemple l'annotation via la concordance proposée par Le Trameur ou TXM), la tâche est énorme et délicate. Car la définition des bonnes unités n'a aucune évidence. En pratique, ce genre d'initiative doit généralement assumer son caractère partiel (on affine le codage sur les occurrences concernant le phénomène étudié) et relatif (on escompte une amélioration en sachant qu'on n'atteint pas l'idéal).

Par ailleurs, une autre source d'informations importante qui intervient dans les traitements textométriques est celle des métadonnées, autrement dit toutes les informations de contexte textuel et intertextuel qui servent tant à situer les observations (en concordance par exemple) qu'à structurer le corpus (construction de contrastes : entre locuteurs, entre périodes, entre genres textuels, etc.). Ce serait donc une erreur de croire que les « formes graphiques » de la textométrie sont des mots totalement dénués d'ancrage textuel et situationnel, au contraire, ces informations participent à l'élaboration des contextes qui sont un élément essentiel et moteur de la méthode, puisque celle-ci procède par l'observation systématique des contextes et par leur mise en contraste.

Enfin, les développements textométriques les plus récents, avec la capacité à travailler sur des corpus structurés, enrichit encore la représentation de la textualité en gardant l'accès au document source (fac-simile, enregistrement avant transcription) et en permettant de détailler des composants intratextuels : phrases, vers, etc. mais aussi distinction enquêteur / enquêté par exemple.

2.4. Le quantitatif : Le fonctionnement de la langue ou des textes n'a rien à voir avec des dénombrements, une approche quantitative est inappropriée.

L'approche quantitative ne s'impose certes pas, en tout cas elle n'est clairement pas exclusive, la textométrie fait toute sa place aussi à une approche qualitative, complémentaire (certains corpus, non conçus en termes de représentativité d'un tout, peuvent d'ailleurs ne se prêter qu'à des explorations qualitatives, cf. Lejeune & Bénel 2012).

Cependant, l'observation montre que la langue présente de très nombreuses propriétés quantitatives (Guiraud 1960) ; et beaucoup de jugements ont à voir avec des aspects quantitatifs, même s'ils ne manient pas de chiffres précis, si bien que « le nombre fait sens » (Mayaffre 2007) :

> On ne compte […] que ce qui est quantifiable : les prix et les produits, les carottes et les avions. On ne compte pas les idées… Pourtant la démocratie s'exerce en comptant les votes, les opinions, les hommes. Et le jugement littéraire lui-même n'est pas tout à fait dépourvu de compteurs inconscients : beaucoup des jugements qu'on croit qualitatifs sont inspirés par une statistique implicite qui n'avoue pas son nom et autorise l'emploi des mots *typique*, *spécifique*, *caractéristique*, si fréquents sous la plume de la critique lorsqu'elle analyse un auteur, un genre ou une époque. Le mot *fréquent* lui-même relève de cette approche, comme aussi *rare*, *original*, *banal*, *courant*, *cliché*, *surprise*, *rupture*. Les littéraires parlent d'horizon d'attente quand les statisticiens parlent d'espérance mathématique. L'espérance des uns et l'attente des autres, ce n'est qu'une prévision fondée sur les observations répétées que la conscience enregistre. (Brunet 2011 : 312-313)

Regarder le texte au prisme des fréquences de ses unités lexicales (ou de ses « formes graphiques »), c'est faire l'hypothèse d'une certaine identité du mot au fil de ses occurrences. Là encore, le linguiste peut, à bon droit, émettre de fortes réserves : la modélisation semble complètement éluder les variations d'énonciateur, et les univers d'assomption différents qui diffractent le sens en plans multiples. Le textomètre ne peut pas non plus ignorer cette diversité profonde de la langue, qui fait que chaque occurrence est unique. De fait, le principe même des traitements textométriques, c'est la caractérisation par les contextes, or ceux-ci se manifestent bien par leur diversité. Ainsi, la textométrie donne également des outils pour observer

les variations ou évolutions d'emploi d'une même forme graphique : par exemple, en calculant ses cooccurrents dans différentes sous-parties du corpus ; ou en observant son appartenance à différentes classes thématiques construites par la méthode Reinert.

L'objection est recensée par Pierre Guiraud (1960), qui explique ainsi comment peut se résoudre le paradoxe :

> *3ème objection* les mathématiques ne peuvent opérer que sur des choses ou des notions rigoureusement identiques : or le langage est une collection d'accidents originaux et essentiellement différents les uns des autres.
>
> On peut ajouter 1 pomme + 1 pomme, mais non 1 pomme + 1 cerise. Ce qui est vrai et faux dans les deux cas ; car, d'une part, il n'y a pas deux pommes absolument semblables ; d'autre part, 1 pomme + 1 cerise = 2 fruits. […] On n'opère jamais sur cette pomme-ci mais sur quelqu'un de ses attributs : sa qualité de fruit, sa forme, sa couleur, son poids etc…. […] On peut refuser de compter combien de fois Baudelaire a employé le mot 'ombre' et à plus forte raison d'ajouter ces 'ombres' baudelairiennes à celles de Hugo ou de Racine ; car à supposer qu'on ait parfaitement distingué entre les diverses significations de ce mot (ombre des arbres, ombres de la nuit, ombres des enfers) il n'y a pas deux exemples de cette collection qui, replacés dans leur contexte, aient une valeur identique. Mais précisément on n'ajoute que ce que ces mots ont en commun : une même figure phonique, une étymologie et un sens de base voisins ; et dans ces limites l'opération est parfaitement légitime ; mais on ne doit pas perdre de vue qu'on n'a pas additionné des signes dans la totalité de leur valeur, mais certains caractères qui leur sont communs. (Guiraud 1960 : 21-22)

Peut-être plus précisément encore, il nous semble que l'idée-clé, c'est que le textomètre sait que, s'il a laissé le logiciel indexer sommairement son texte, ses décomptes ne portent pas sur des mots –et encore bien moins sur des sens ou sur des concepts-, mais sur des « formes graphiques ». Ainsi, l'interprétation ne s'arrête pas à l'entité dénombrée, mais elle en pressent la plus ou moins grande « épaisseur » sémantique, voire l'hétérogénéité ; la consultation des contextes est de règle, pour éviter les évidences intuitives et simplifiantes, et détailler et nuancer l'interprétation, voire la suspendre (une stratégie d'analyse peut identifier, évaluer à part, puis écarter de l'analyse, certaines unités trop ambiguës ou incertaines, pour poursuivre l'analyse sur un terrain clarifié). Complémentairement, l'expérience textométrique montre que les structures globales mises en évidence par les calculs sont remarquablement stables par delà différents choix de représentation des textes (graphies, lemmes, catégories grammaticales, voire réduction à des séquences de consonnes et voyelles, Brunet 2007, 2016 : 149-150). Ces choix de représentation sont comme différents filets jetés sur les textes, filets aux mailles plus ou moins fines, plus ou moins ajustés aux « vraies » unités linguistiques, mais sans doute jamais complètement décalés non plus ; l'information linguistique apparaît riche, redondante, si bien que ce que captent partiellement et imparfaitement les mailles redonne cependant globalement une bonne image d'ensemble.

2.5. La complexité : Pourquoi ces modèles statistiques, et d'ailleurs lequel faudrait-il adopter ? Un simple calcul de pourcentages répond aux questions d'irrégularité d'emploi de façon claire et directe.

La simple observation des fréquences peut donner une idée des irrégularités quantitatives d'emploi entre différents sous-ensembles de textes, lorsque ceux-ci sont de tailles analogues. Lorsque les tailles sont inégales, une règle de trois (rapportant la fréquence à la taille de la partie, pour en faire une fréquence relative, normalisée) semble une réponse simple et adaptée. Pourtant, à y regarder de plus près, cette opération mathématique suppose une sorte de fonctionnement homothétique de la langue : dans un texte équivalent mais deux fois plus long, les mots devraient tous avoir un usage doublé. Au vu des corpus, il est bien évident que cela ne correspond pas au fonctionnement linguistique, chaque texte comprenant notamment une très large part de *hapax*, mots à usage unique dans le texte. C'est en reformulant mathématiquement précisément le questionnement linguistique (« je cherche les mots qui semblent marquer une attirance –ou une répulsion- pour une partie du corpus ») que le calcul des spécificités a été proposé. Mesurant des probabilités, il peut non seulement mettre en évidence des variations d'usage, mais il peut également apprécier l'importance de celles-ci, leur caractère plus ou moins notable (est-ce que cela fait partie des fluctuations courantes ou est-ce que cela marque véritablement un écart). Par exemple, une fréquence nulle a toujours le même pourcentage (0 %), quelle que soit la fréquence par ailleurs du mot en corpus et quelle que soit la taille de la partie ;

alors qu'une évaluation de cette même fréquence nulle en spécificité fournit un indice qui permet de faire la part entre l'absence usuelle et l'absence marquée (nullax).

Le rapport au modèle statistique n'est pas un rapport confirmatoire (où il s'agirait de modéliser la langue pour prédire ses réalisations textuelles), mais un usage exploratoire. Par exemple, pour le modèle des spécificités utilisé pour mesurer les sur- ou sous-emplois, la modélisation représente une abstraction que l'on se donne comme repère (une répartition aléatoire des mots parmi les textes du corpus), et par rapport à laquelle on peut évaluer les écarts des répartitions observées. Le modèle ne prétend à aucune réalité linguistique, mais n'en est pas moins utile pour caractériser des usages.

> [C'est] l'application du schéma d'urne [que Bratley] conteste radicalement, comme tout à fait inadéquat au domaine des mots. Le schéma d'urne suppose des tirages indépendants. Or, les mots dans la chaîne du discours sont interdépendants. Ainsi l'article appelle un substantif subséquent et le mot *chat* une fois tiré exclut un second tirage immédiat du même mot. Le modèle est donc faux dans son principe. […] [En revanche] le schéma d'urne et le modèle probabiliste [peuvent fournir] la référence d'où procède la mesure. Si je veux vérifier qu'une ligne est droite ou non, qu'une surface est plane ou non, je me sers d'une règle. Si la surface a des creux et des bosses, ou si la ligne a des sinuosités, je ne vais pas casser la règle, sous le prétexte qu'elle ne convient pas aux données, que la nature est rebelle aux figures idéales et que la « prévisibilité » de la règle est toujours démentie par des faits. Il y a beaucoup à dire sur la notion de prévisibilité : la règle dont je me sers ne permet pas de prévoir si la ligne que je suis va tourner à droite ou à gauche, pas plus que le thermomètre ne me permet de savoir quelle température il fera demain. En matière lexicale la règle statistique ne permet, elle aussi, que la mesure. II ne s'agit que de décrire, nullement d'expliquer, moins encore de prévoir. (Brunet 2016 : 368-369)

Ainsi, les modélisations mathématiques adoptées par la textométrie (le calcul des spécificités, l'analyse factorielle des correspondances) correspondent à des choix de modélisation précis, explicitables, adaptés aux données textuelles. Il n'y a pas une complexité arbitraire, à admettre telle quelle (argument d'autorité), ou à tester (validation empirique, « telle formule marche mieux / moins bien »). L'enjeu pour l'utilisateur c'est de comprendre les principes sous-jacents (que vise la modélisation, que prend-elle en compte, de quelle façon) et de comprendre en conséquence comment lire les résultats (Guerreau 2012, Mairey 2017, Demazière et al. 2006, Marchand & Ratinaud 2016).

Dans cet esprit, nous pouvons apporter des éléments de clarification par rapport aux critiques de Carbou (2017a) vis-à-vis de la classification Reinert. Est-elle thématique ? Le choix de construire des unités de contexte de l'ordre d'une à quelques phrases et de se focaliser sur les lexèmes est efficace pour capter particulièrement des relations thématiques, donc le résultat aura une dominante thématique, mais il resterait abusif d'en conclure qu'on produit exactement des thèmes (voire les thèmes). Certaines classes peuvent être hétérogènes ? Oui, cela dépend de l'algorithme, on peut anticiper cela. Si l'on veut réduire l'hétérogénéité des classes il faut adapter l'algorithme, ou peut-être simplement certains seuils. Une interprétation basée sur les éléments dominants des classes serait biaisée ? Mais souvent ce ne sont pas les points dominants que l'on retient comme caractéristiques, mais les éléments les plus discriminants. En tout cas il faut comprendre le mécanisme de sélection des représentants pour comprendre les critères en jeu (plutôt internes à la classe ou plutôt externes). La classification dépend du corpus global ? Mais oui, pourquoi pas, c'est prendre acte du fait que le corpus joue le rôle de référence ; l'impact de la composition du corpus dépendra aussi de l'algorithme (ce sera plus sensible en classification descendante qu'en classification ascendante, où les regroupements sont d'abord locaux) et du nombre de classes demandé.

2.6. L'objectivité : Grâce aux mathématiques, au traitement formel et informatisé, on a une analyse neutre, objective.

La valeur de l'analyse, son caractère scientifique, rigoureux, éclairé, ne reçoivent aucune garantie du simple fait d'avoir mobilisé un appareillage mathématique ou informatique. Le recours à la modélisation formelle et à l'expérimentation, qui obligent à fixer le corpus, les unités d'analyse, les traitements précis effectués, peuvent bien sûr contribuer directement à la mise en œuvre d'une analyse méthodique et systématique. Mais encore faut-il conduire l'analyse.

Cette fascination pour le caractère ésotérique et savant des formules statistiques a été clairement dénoncée, voire pour l'analyse quantitative comme traitement à la mode (Lemercier & Zalc 2008).

> Parmi les linguistes et surtout les « littéraires », c'est une minorité qui pratique ces méthodes [quantitatives] […]. Chez certains, c'est une abstention polie, teintée d'ironie, et qui se retranche derrière l'allergie bien connue, de tout vrai littéraire pour toute mathématique […]. Chez d'autres, c'est une franche hostilité qui les pousse à condamner ces techniques grossières et réductrices […]. Il y a peut-être pire […] : à l'inverse de la franche hostilité, j'ai dû constater l'ignorance béate et confiante, pour qui tout ce qui est calcul ou formule est vérité d'évangile. (Muller 1986 : 10-11)

> [Il] y a beaucoup à faire encore, pour conjurer les deux dangers qui guettent les humanités : l'exclusion stupide de l'ordinateur ou une sotte allégeance à son endroit. […] [L'inquiétude des premiers] est rassurante : pour convaincre les sceptiques il suffit de montrer que l'homme anime la machine et que la main tient l'outil. On peut craindre davantage le danger opposé, qui donne tout pouvoir à la machine […]. (Brunet 2016 : 364)

> La méthode [logométrique] souffre surtout, dans l'usage vulgaire, d'une réduction voire d'une perversion de ses objectifs. Assignée seulement à administrer la preuve chiffrée grâce au décompte des mots, elle manque l'essentiel, se met sous une saine critique de scientisme et se condamne à n'être qu'un gadget convoqué par les littéraires, les historiens ou les politologues en mal de critères objectifs (i.e. quantifiés) pour sceller leur démonstration. (Mayaffre 2010 : 22-23)

François Rastier (1991, 2011) a décrit les calculs et autres traitements automatiques formalisés comme un moment de suspens de l'interprétation, entre deux phases où l'interprétation s'impose, à la fois scientifiquement mais aussi culturellement et sémantiquement (l'homme ne peut s'empêcher de donner du sens). Si bien que pour le chercheur analysant ses données, il s'agit de contrôler ces différents moments en préparant les données pour les calculs, en comprenant les principes des manipulations opérées, en connaissant les règles de lecture des représentations produites (herméneutique des sorties logicielles, Rastier 2011 : 44), et en prenant méthodiquement appui sur les observations et la consultation systématique du corpus pour établir un faisceau d'indices convergents à l'appui de la lecture proposée des résultats.

> Le suspens de l'interprétation est impossible dans les sciences de la culture (j'y inclus les sciences historiques, humaines ou sociales), dont les objets sont des œuvres humaines. Nous sommes condamnés au sens, soit : il ne s'agit cependant pas pour les sciences de la culture de renchérir sur le mode compulsif de l'interprétation, mais bien plutôt d'établir une distance critique à l'égard des interprétations, d'en spécifier les conditions et d'en interroger la légitimité.

> La formalisation, certes, suspend l'interprétation, mais pas le temps du calcul, dont les résultats devront être interprétés. Il ne s'agit donc pas d'éliminer l'interprétation, mais bien de la problématiser. Le détour du calcul ne supprime pas la dimension critique, mais engage au contraire à la redoubler: qui a manié des calculs statistiques peut aisément s'en convaincre. (Rastier, *in* Lacour 2005)

De fait, l'une des faiblesses possibles de certaines analyses textométriques ou quantitatives, c'est bien l'interprétation subjective et trop libre des résultats : en s'en tenant à une lecture intuitive et directe des sorties graphiques ou chiffrées pour elles-mêmes (sans les indicateurs complémentaires, les fréquences absolues, le nombre de textes concernés, les tailles des parties, les seuils, les rangs, les ordres de grandeur, etc.), et en « comprenant » spontanément les fragments lexicaux mis en évidence et se recontextualisant les uns les autres sans vérifier dans le corpus les relations attestées entre ces mots (par exemple, s'agit-il de syntagmes avérés ou de moirages paradigmatiques (Tournier 1986)). Le résultat textométrique, ce n'est pas que tel mot soit apparu avec telle fréquence exacte, ou atteigne tel score dans telle partie du corpus : c'est plutôt de mettre en évidence un faisceau d'indices concrets, convergents, significatifs, en lien avec une interprétation. Et comme le note Condamines[2], l'expert du domaine représenté par le corpus n'est pas moins exposé que lecteur naïf, car quelques mots suffisent à stimuler ses connaissances et à ébaucher un discours faisant sens, mais donc pas forcément congruent avec les structures effectivement mises au jour par les traitements.

[2] Communication personnelle, 16 novembre 2018.

2.7. Les résultats : Cela fait de beaux graphiques pour la communication (en revanche, scientifiquement, cela ne va pas bien loin)

Les outils sont par nature productifs. Il faudrait alors privilégier l'usage de ces méthodes pour ce qu'elles peuvent apporter de nouveau, de renouvelant. Sinon, le risque serait de conclure que c'est une montagne qui accouche d'une souris : on mobilise un lourd appareillage mathématique, et finalement on trouve des évidences, on enfonce des portes ouvertes (Carbou 2017b).

Effectivement, tout traitement textométrique ne génère pas forcément une révélation, une découverte. Dans beaucoup de cas, l'interrogation confirme ce que des travaux antérieurs avaient déjà perçu –ce qui en soi est rassurant sur les capacités de la méthode. Nonobstant, l'exploration systématique d'un corpus souvent plus étendu permise par l'outil textométrique permet de repérer des éléments factuels pour confirmer une intuition ou de premières observations, et d'affiner la description (Prévost 2005). Et dans certains cas (pour les analyses factorielles typiquement), ce qui est reconnu comme évident ne l'est peut-être pas tant que cela : aurait-on vraiment cartographié le corpus de la même façon, ou repéré et hiérarchisé les mots caractéristiques sans inversion ni oubli ? N'y a-t-il pas dans le résultat du calcul des « détails » qui nous auraient échappé ?

Il reste aussi plus clairement des cas de découverte, ou du moins d'attention attirée sur certains aspects, avec la mise en évidence de points singuliers au milieu des régularités confirmées. Par ailleurs, les unités plus formelles que linguistiques sur lesquelles se base l'analyse, et leurs reconfigurations par les calculs, génèrent de nouveaux observables (Rastier 2011).

La démarche textométrique répond au désir de lire autrement (Lejeune 2010) : d'étudier le texte sous un autre angle, de prendre ses distances par rapport au parcours ménagé par le texte, aux mots qui se fondent dans l'enchaînement du texte, aux structures marquées. La délinéarisation du texte et la restructuration du contenu désamorce certains automatismes et fait quitter un terrain familier. La formalisation conduit à expliciter des présupposés, à décentrer le lecteur pour se recentrer sur le texte, à dépasser la lecture empathique des textes (Mayaffre 2007).

Il faut aussi accepter les limites inhérentes au travail sur corpus numérique : comme celui-ci oblige à préciser concrètement ses données, on doit généralement constater dans son corpus des lacunes, des hétérogénéités, des incertitudes, des biais. Mais l'imperfection du corpus ne compromet pas nécessairement l'intérêt des observations. Ce qui importe, c'est la bonne connaissance du contenu du corpus et de ses limites, son *interprétabilité*, de sorte à pouvoir ajuster en conséquence la lecture des résultats des traitements (Pincemin 2012b).

Par comparaison avec le domaine du traitement automatique des langues (TAL), Valette (2016) s'inquiète de l'absence de procédure claire d'évaluation en textométrie. Certains travaux de textométrie donnent l'impression d'une analyse très personnelle, peu contrôlable ; alors qu'en TAL, les propositions peuvent être situées par des indicateurs chiffrés précis de performance par rapport à des bancs d'essais et des résultats de référence externes. Il faut cependant d'abord souligner que les objectifs du TAL et ceux de la textométrie ne sont pas analogues : le TAL consiste précisément à automatiser une certaine tâche d'analyse linguistique (par exemple l'analyse morphologique, l'analyse syntaxique, la reconnaissance d'entités nommées, l'extraction de certaines informations) pour laquelle on considérera que le résultat est unique et défini. Le principe de l'évaluation peut donc consister à mesurer l'écart entre le résultat d'un traitement et la bonne réponse attendue, pour un certain nombre de cas préparés dans une batterie de tests. En revanche, la textométrie s'intéresse à ouvrir de nouveaux parcours de lecture dans un corpus d'étude : or cette lecture est une activité fondamentalement sémantique et interprétative, qui intègre par nature la pluralité des résultats. Pour un corpus donné, il n'y a pas « la » bonne analyse textométrique qui pourrait servir d'étalon de contrôle, mais une multiplicité de lectures et de points de vue sur les textes. Cela étant, l'exigence de rigueur scientifique et d'évaluabilité reste entière : elle ne réside pas dans le résultat produit, mais dans la cohérence et la pertinence du parcours d'analyse, et dans les moyens donnés pour sa reproductibilité (précision de l'édition, explicitation des conventions de transcription, choix du logiciel, détail des paramètres des traitements, accès aux contextes d'emploi pour vérification, etc.).

3. Bilan et perspectives

Au fil de notre parcours, les doutes et craintes du linguiste vis-à-vis des approches quantitatives qui ont pu être examinés sont : l'automatisation de l'analyse et l'abandon du travail de recherche, le primat qui pourrait être donné au volume, l'indigence linguistique des unités de base de l'analyse, la nature fondamentalement non quantitative de la langue, la complexité ou l'inadéquation des modèles mathématiques, la neutralité des traitements formels et la place de l'interprétation, la qualité scientifique des résultats et les apports originaux de l'approche textométrique. Par rapport à ces questions de fond, nous espérons avoir apporté une meilleure compréhension des principes de la textométrie, et montré que celle-ci n'élude pas ces points critiques mais les intègre de façon contrôlée.

En pratique, le chercheur qui expérimente l'approche textométrique pourra dissiper de lui-même assez vite certaines illusions : clairement, l'analyse ne se fait pas toute seule (il n'y a pas automatisation de l'analyse), et le calcul n'économise pas la lecture des textes (il faut bien connaître son corpus et souvent reconsulter les contextes d'emploi pour comprendre les résultats des calculs et éviter les erreurs d'interprétation). Il s'agit en quelque sorte ici de quiproquos importants mais relativement simples, principalement issus d'*a priori* et de méconnaissances. À l'opposé, d'autres interrogations accompagneront le textomètre tout au long de sa pratique, et continueront à faire l'objet des débats entre pairs et d'un approfondissement scientifique : comment penser les correspondances (indirectes mais effectives) entre forme et contenu, surface et profondeur, localité et globalité, lexique/morphosyntaxe et texte/intertexte ? Jusqu'où aller dans l'édition des textes et l'annotation des corpus, quel équilibre et quelle dynamique trouver entre l'observation de données « brutes » et leur mise en forme savante, experte ? Et du côté des résultats, ce sont aussi les avancées et les échanges de toute la communauté scientifique qui pourront alimenter la mise en évidence des apports originaux de l'approche : en la matière, un exemple isolé n'est jamais suffisant, puisque tout est affaire d'appréciation progressive, et non de validation binaire et univoque (il y a apport ou non).

L'enjeu est bien de clarifier les apports de la textométrie pour l'étude scientifique de la langue et des textes, pour explorer et discuter les perspectives ouvertes par ses orientations les plus récentes. Avec l'effervescence des Humanités numériques et les nouvelles éditions de corpus structurés et annotés, la complémentarité maîtrisée et l'articulation fine des approches quantitatives et qualitatives est plus que jamais d'actualité. La formation des (jeunes) chercheurs et les échanges scientifiques ont un rôle décisif.

Univ. Lyon, CNRS, UMR 5317 IHRIM.

Remerciements : Je tiens à remercier Sophie Audidière, Max Beligné, Isabelle Chesneau, Anne Condamines, Alain Guerreau, Hélène Ledouble, Aude Mairey, Michèle Monte, Francesca Rebasti, Mathieu Valette, pour les éléments de bibliographie, d'expérience et de réflexion dont ils m'ont fait part et qui ont nourri la rédaction de cet article.

Bibliographie

BRUNET, Étienne (2007), « Le corpus conçu comme une boule », *Corpus en Lettres et Sciences sociales. Des documents numériques à l'interprétation. Actes du XXVIIe Colloque d'Albi Langages et Signification*, François Rastier et Michel Ballabriga (dir.), Carine Duteil-Mougel et Baptiste Foulquié (éds), Presses universitaires de Toulouse. Actes également publiés en ligne sur le site Texto! Textes & Cultures, http://www.revue-texto.net/1996-2007/Parutions/Livres-E/Albi-2006/Sommaire.html. Texte réédité dans Brunet (2011), chapitre 14, p. 279-292.

BRUNET, Étienne (2011), *Ce qui compte. Écrits choisis, tome II : Méthodes statistiques*, Poudat Céline (éd.), Paris, Champion.

BRUNET, Étienne (2016), *Tous comptes faits. Écrits choisis, tome III : Questions linguistiques*, Bénédicte Pincemin (éd.), Paris, Honoré Champion, coll. « Lettres numériques ».

CARBOU, Guillaume (2017a), « Quelques questions à l'attention des utilisateurs des statistiques textuelles pour l'analyse du discours », *Texto! Textes et cultures*, XXII (4), Coordonné par Créola Thenault.

CARBOU, Guillaume (2017b), « Analyser les textes à l'ère des humanités numériques. Quelques questions pour l'analyse statistique des données textuelles », *Les Cahiers du numérique*, 13 (3), pp. 91-114.

CHATEAURAYNAUD, Francis (2003), *Prospéro. Une technologie littéraire pour les sciences humaines*, Paris, CNRS Éditions.

CHATEAURAYNAUD, Francis, DEBAZ, Josquin (2012), « Prodiges et vertiges de la lexicométrie »,in Pierre Mounier (dir.), *Read/Write Book 2 : Une introduction aux humanités numériques*, Marseille, OpenEdition Press, http://books.openedition.org/oep/279.

DALUD-VINCENT, Monique (2011), « Alceste comme outil de traitement d'entretiens semi-directifs : essai et critiques pour un usage en sociologie », *Langage et société*, 2011 (1), 135, pp. 9-28.

DEMAZIÈRE, Didier, BROSSAUD, Claire, TRABAL, Patrick, VAN METER, Karl (dir.) (2006), *Analyses textuelles en sociologie. Logiciels, méthodes, usages*, Rennes, Presses universitaires de Rennes.

DUPUY, Pierre-Olivier, MARCHAND, Pascal, (2011), « Confrontation et positionnement dans les duels de l'entre-deux-tours : une approche lexicométrique », in Marcel Burger (éd.), *La parole politique en confrontation dans les médias*, Louvain-la-Neuve, De Boeck Supérieur, coll. « Culture & Communication », pp. 129-147.

GEFFROY, Annie, LAFON, Pierre (1982), « L'insécurité dans les grands ensembles. Aperçu critique sur Le vocabulaire français de 1789 à nos jours d'Etienne Brunet », *MOTS*, 5, pp. 129-141.

GEFFROY, Annie, LAFON, Pierre, TOURNIER, Maurice (1974), « L'indexation minimale, Plaidoyer pour une non-lemmatisation », *Colloque sur l'analyse des corpus linguistiques : Problèmes et méthodes de l'indexation minimale*, Strasbourg, 21-23 mai 1973.

GUERREAU, Alain (2012), « Textes anciens en série », *Bulletin du centre d'études médiévales d'Auxerre | BUCEMA*, Collection CBMA, Les outils, http://journals.openedition.org/cem/12177.

GUIRAUD, Pierre (1960), *Problèmes et méthodes de la statistique linguistique*, Paris, Presses universitaires de France.

JENNY, Jacques (1997), « Méthodes et pratiques formalisées d'analyse de contenu et de discours dans la recherche sociologique française contemporaine. État des lieux et essai de classification », *Bulletin de Méthodologie Sociologique*, 54, pp. 64-112.

KERBRAT-ORECCHIONI, Catherine (2017), « Les débats de l'entre-deux-tours des élections présidentielles françaises : constantes et évolutions d'un genre », Paris, l'Harmattan, coll. « Du sens ».

LACOUR, Philippe (2005), « Les sciences cognitives. Entretien avec François Rastier », *Labyrinthe*, 20, pp. 117-134.

LAFON, Pierre, SALEM, André (1983), « L'inventaire des segments répétés d'un texte », *MOTS*, 6, pp. 161-177.

LEBART, Ludovic (2004), « Validité des visualisations de données textuelles », in Gérald Purnelle et al. (éds), Actes des 7es Journées internationales d'analyse

statistique des données textuelles (JADT 2004), Louvain-la-Neuve, Presses universitaires de Louvain, vol. II, pp. 708-715.

LEBART, Ludovic, PINCEMIN, Bénédicte, POUDAT, Céline (2019), *Analyse des données textuelles*, Québec, Presses de l'Université du Québec.

LEBART, Ludovic, SALEM, André (1994), *Statistique textuelle*, Paris, Dunod.

LECA-TSIOMIS, Marie (2014), « Du bon usage de l'informatique dans la recherche littéraire et historique », *Dix-huitième siècle*, 46 (1).

LEJEUNE, Christophe (2010), « Montrer, calculer, explorer, analyser. Ce que l'informatique fait (faire) à l'analyse qualitative », *Recherches Qualitatives*, 9, pp. 15-32.

LEJEUNE, Christophe, BÉNEL, Aurélien (2012), « Lexicométrie pour l'analyse qualitative. Pourquoi et comment résoudre le paradoxe ? », in Anne Dister, Dominique Longrée, Gérald Purnelle (éds), *JADT 2012. Actes des 11es Journées internationales d'analyse statistique des données textuelles*, Université de Liège / Facultés Universitaires Saint-Louis Bruxelles, pp. 591-602.

LEMERCIER, Claire, ZALC, Claire (2008), *Méthodes quantitatives pour l'historien*, Paris, La Découverte, coll. Repères.

MAIREY Aude (2017), *Chemins de traverses*, rapport de synthèse inédit présenté dans le cadre d'une Habilitation à diriger des recherches intitulée « Langues, cultures et société politique en Angleterre à la fin du Moyen Âge », soutenue le 21 janvier 2017 à Université Paris 1.

MARCHAND, Pascal, RATINAUD, Pierre (2016), « Faut-il faire des nuages de mots ? », *Site IRaMuTeQ*, rubrique Études. http://www.iramuteq.org/Members/pmarchand/faut-il-faire-des-nuages-de-mots

MAYAFFRE, Damon (2007), « Analyses logométriques et rhétoriques des discours », in Stéphane Olivési (dir.), *Introduction à la recherche en SIC*, Grenoble, Presses universitaires de Grenoble, 2007, pp. 153-180.

MAYAFFRE, Damon (2010), *Vers une herméneutique matérielle numérique. Corpus textuels, Logométrie et Langage politique*, Mémoire d'Habilitation à diriger des recherches en Histoire, Université Nice Sophia Antipolis.

MULLER, Charles [1977] (1993), *Principes et méthodes de statistique lexicale*, Paris, Champion.

MULLER, Charles (1987), « Introduction », in Étienne Brunet, *Méthodes quantitatives et informatiques dans l'étude des textes*, Genève, Slatkine, et Paris, Champion.

PINCEMIN, Bénédicte (2012a), « Sémantique interprétative et textométrie » [version française complète], Christophe Cusimano (dir.), *Texto! Textes & Cultures*, 17 (3), http://www.revue-texto.net/index.php?id=3049.

PINCEMIN, Bénédicte (2012b), « Hétérogénéité des corpus et textométrie », *Langages*, 187, pp. 13-26.

PRÉVOST, Sophie (2005), « Constitution et exploitation d'un corpus de français médiéval : enjeux, spécificités et apports », in Anne Condamines (éd.), *Sémantique et corpus*, Paris, Lavoisier, Hermès science, série Cognition et traitement de l'information, Traité IC2 Information, Commande, Communication, pp. 147-176.

RASTIER, François (1991), *Sémantique et recherches cognitives*, Paris, Presses universitaires de France, coll. Formes sémiotiques.

RASTIER, François (2001), *Arts et sciences du texte*, Paris, Presses universitaires de France, coll. Formes sémiotiques.

RASTIER, François (2005), « Sémiotique du cognitivisme et sémantique cognitive : Questions d'histoire et d'épistémologie », Texto! mars 2005 [en ligne]. Disponible sur : <http://www.revue-texto.net/Inedits/Rastier/Rastier_Semantique-cognitive.html>. (Consultée le 28 janvier 2019)

RASTIER, François, (2011), *La mesure et le grain. Sémantique de corpus*, Paris : Honoré Champion, coll. Lettres numériques.

RIZKALLAH, Élias (2013), « L'analyse textuelle des discours assistée par ordinateur et les logiciels textométriques : réflexions critiques et prospectives à partir d'une modélisation des procédés analytiques fondamentaux », *Cahiers de recherche sociologique*, 54, pp. 141–160.

TOURNIER, Maurice (2006), « Dans l'ombre portée des sigles confédéraux : un mirage lexicométrique (C.G.T. et C.F.D.T. en 1972) », in Étienne Brunet (dir.), *Méthodes quantitatives et informatiques dans l'étude des textes*, Genève, Slatkine, et Paris, Champion, pp. 841-853.

VALETTE, Mathieu (2016), « Analyse statistique des données textuelles et traitement automatique des langues. Une étude comparée », in Damon Mayaffre, Céline Poudat, Laurent Vanni, Véronique Magri, Peter Follette (éds), *Statistical Analysis of Texual Data. JADT 2016. Proceedings of 13th International Conference on Statistical Analysis of Textual Data*, vol. II, pp. 697-706.

Résumé

Si la textométrie procède essentiellement de comptages (purement quantitatifs) sur des mots (définis à travers leur matérialité graphique), alors n'accuse-t-elle pas d'évidentes limites du point de vue des finesses de la réalité linguistique ? Et, malgré sa mise en œuvre formelle et calculatoire, ne manque-t-elle pas de rigueur concernant l'évaluation des résultats obtenus, en l'absence de procédures comparatives chiffrées, à l'instar des pratiques en traitement automatique des langues ? Vis-à-vis de telles critiques de fond, l'objectif de cet article est d'apporter des éléments de compréhension du modèle textométrique permettant de dépasser un certain nombre de points perçus comme problématiques, et de proposer une conception de la textométrie compatible avec les exigences de la linguistique.

Mots-clés

Analyse statistique des données textuelles (ADT), linguistique, analyse quantitative, discussion critique, évaluation.

Abstract

Considering that textometric statistical analysis of textual data is based on plain counts (which are just numeric values) of words (which are defined as rough character string tokens), this approach shows obvious limits from a linguistic point of view. Moreover, Textometry uses formal and mathematical models to analyse corpora, but results are often delivered in a very informal manner, without any quantified evaluation procedure like the ones that are applied in the Natural Language Processing field. The present paper aims at giving an in-depth understanding of the textometric methodology, so that such critical points may not be relevant anymore. Then, Textometry can meet the requirements of a linguistic-aware and scientific study of textual data.

Keywords

Statistical Analysis of Textual Data, Linguistics, Quantitative Analysis, Critical Discussion, Evaluation.