Quelques questions à l'attention des utilisateurs de statistique textuelle pour l'analyse des discours

Guillaume Carbou
(Université de Bordeaux / SPH)

Résumé : Cet article, volet d'un travail en deux parties, propose une liste de questions critiques à l'attention des utilisateurs de logiciels de statistique textuelle. Ces techniques d'analyse des discours deviennent de plus en plus courantes dans de nombreux champs de recherche. Or, si la statistique textuelle offre un moyen intéressant d'aborder les corpus, il est nécessaire de prendre un certain nombre de précautions théoriques et méthodologiques pour faire un usage éclairé de cet outil. Les questions posées dans le présent article invitent à ce recul critique. Elles interrogent les hypothèses sur les fonctionnements textuels que porte en elle la statistique textuelle (lexicocentrisme, compositionnalisme, typification lexicale…) ainsi que les difficultés d'interprétabilité des sorties-machine.

Mots-clefs : statistique textuelle, sciences du texte, méthodologie, lexicométrie, textualité.

Cet article s'inscrit dans un diptyque dont le second volet est publié dans la revue *Les cahiers du numérique* (Carbou, 2017). Ces deux articles visent à lister les principales questions que doivent se poser les utilisateurs d'outils d'analyse informatisée des données textuelles (ou encore de lexicométrie, textométrie, statistique textuelle, de *topic modeling*, … nous ne faisons pas ici de distinction entre ces diverses appellations) afin de faire un usage éclairé de leurs instruments. Il ne s'agit en aucun cas de nier les intérêts épistémiques et la puissance heuristique de l'analyse des textes assistée par ordinateur mais simplement de désigner les écueils qui guettent les utilisateurs les moins avertis. Le volet publié dans *Les cahiers du numérique* s'attache à des questions d'ordre épistémologique : il s'agit de se demander quel rôle peut jouer l'outil informatique dans l'étude des grands corpus numériques. Nous y suggérons que la statistique textuelle doit se garder de soutenir les tendances objectivistes de certaines franges des humanités numériques. Au contraire, elle doit participer à affirmer la dimension herméneutique de l'analyse des textes, plaider pour une approche prudente des (trop) grands corpus, ou encore distinguer entre usage probatoire et exploratoire des calculs statistiques.

Dans le présent volet, nous nous intéressons à l'analyse des textes assistée par ordinateur d'un point de vue plus linguistique. D'une part nous adoptons le regard des sciences du texte pour mettre en garde contre certaines inadéquations entre les modes d'analyse des logiciels de lexicométrie et les fonctionnements textuels effectifs. D'autre part, nous soulevons certains des risques liés à la difficile interprétabilité des sorties-machine : les calculs qui déstructurent les textes et les recomposent sous forme graphique peuvent conduire à des « mirages lexicométriques » (Tournier, 1985).

Notre démarche dans ce travail part du constat que l'utilisation de logiciels de statistique textuelle pour analyser les discours traverse aujourd'hui tous les champs de recherche. Psychologie, sociologie, linguistique, sciences de gestion, économiques, politiques, de l'information et de la communication, etc., voient fleurir les approches outillées de leurs données textuelles. À l'heure où les grands corpus numérisés sont aisément accessibles, la lexicométrie semble offrir la possibilité de traitements rapides et économiques.

Toutefois, on peut craindre que l'expansion de cet outil hors de son champ de spécialité (à savoir l'analyse du discours, dûment informée par les sciences du langage et la statistique) mène à des utilisations maladroites. La simplicité apparente d'usage, la possibilité de traiter en quelques clics des grandes masses de données numérisées, la dimension objectivante du traitement statistique, la force rhétorique de résultats présentés sous forme graphique sont autant d'éléments qui favorisent des appropriations

discutables des outils lexicométriques.

Les usagers avertis pourront certes trouver la mise en garde triviale, mais la diffusion massive des logiciels de lexicométrie et l'observation des utilisations cavalières dont ils font parfois l'objet laissent penser qu'elle reste nécessaire. C'est donc dans cette perspective que se place ce travail : son objectif est d'inviter à un positionnement réflexif et critique vis-à-vis de l'usage des logiciels de lexicométrie pour analyser les discours. Précisons que les pratiques que nous interrogeons ici sont celles qui consistent à utiliser un logiciel dans le but de connaître le « contenu », généralement thématique, des textes d'un corpus. Les usages informés en sémantique de corpus ou en traitement automatique des langues par exemple ne font donc pas l'objet de notre propos. Nous nous concentrons au contraire sur les outils qui réalisent des classifications ou des recherches de mots-clefs (projection de dictionnaire) : on pourra citer sans exhaustivité aucune Alceste, Iramuteq, Sphinx, WordStat, Prospero, Diction, Trope, Hyperbase, Proxem, et les algorithmes du *topic modeling*.

Nous proposons donc ici une recension des grandes interrogations qui peuvent s'exprimer, dans la littérature (cf. par exemple Pincemin, 1999 ; 2012 ; Jenny, 1999 ; Dalud-Vincent, 2011) ou dans les couloirs de laboratoires et de colloques, face aux analyses lexicométriques. Un certain nombre de ces remarques sont récurrentes, et il nous a semblé utile de les regrouper dans un document unique (certes en deux parties). Certains points développent des objections précises, d'autres relèvent de réflexions ouvertes. Conformément à la vocation du document, ils cherchent moins à apporter des réponses qu'à susciter une réflexion épistémologique. Le texte peut être lu de manière linéaire mais, en tant qu'il constitue avant tout un recueil de mises en garde, les sections peuvent être consultées indépendamment.

Le plan de cet article est le suivant[1]. Après avoir précisé le sens très large que nous donnons ici aux expressions « lexicométrie » et « analyse des discours », nous réunirons les questions à poser aux approches outillées des textes en deux parties. La première présente la suspicion des sciences du texte face à la déstructuration profonde que fait subir la lexicométrie à son objet. Elle se divise en trois questions. Nous soulevons tout d'abord les problèmes que pose le lexicocentrisme de la méthode puis nous interrogeons la vocation d'analyse de « contenu » que revendiquent nombre de logiciels : comment déterminer la présence ou l'absence d'un thème à partir des mots d'un texte ? Et comment être sûr que les méthodes classificatoires dégagent bien des thèmes dans un corpus ? La seconde partie de l'article expose les interrogations du point de vue des statistiques : quelles transformations – ou déformations – le traitement statistique des textes peut-il induire ? Les sorties-machines, sous forme de graphes ou de classes, peuvent-elles être aussi aisément interprétées que le sens commun y invite ? Et de solides compétences en statistiques ne sont-elles pas nécessaires pour prendre la mesure de ces enjeux ? Enfin, avant de conclure, nous proposons une synthèse de l'ensemble de nos remarques.

En préambule : « lexicométrie » et « analyse des discours »

Précisons le sens que nous donnons ici à ces deux expressions, la lexicométrie et l'analyse des discours[2].

Nous appelons lexicométrie (textométrie, logométrie, statistique textuelle, analyse informatisée des données textuelles, etc., sans distinction) toute approche outillée d'un corpus de textes opérant sur celui-

[1] Le plan du second article est le suivant : dans la première partie nous adoptons le point de vue des « sciences de la culture » et nous demandons si les méthodes informatisées d'analyse des textes ne risquent pas d'entretenir des liaisons dangereuses avec les franges les plus réductionnistes des humanités numériques ; si la centration sur les phénomènes de haute fréquence qui leur est inhérente est bien pertinente pour étudier le contenu des textes ; et si le traitement des (très) grands corpus numériques doit réellement être vu comme un impératif ? Dans la seconde partie, nous faisons état de questionnements épistémologiques plus généraux qui visent à déterminer ce que l'on « fait » lorsque l'on utilise un logiciel de statistique textuelle : met-on réellement à distance la subjectivité et l'interprétation ? Peut-on espérer que les constats statistiques jouent le rôle de preuves ou seulement d'indices dans la démarche scientifique ? Et qu'implique la distinction entre visée pratique et visée épistémique des logiciels pour les procédures de recherche ?

[2]Nous employons volontairement le déterminant « des » plutôt que « du » pour ne pas renvoyer au syntagme figé « analyse du discours » et ainsi au champ disciplinaire auquel il réfère.

ci des traitements quantitatifs et statistiques. Nous excluons donc de notre discussion les fonctions d'exploration textuelle qui consistent à effectuer des requêtes spécifiques pour atteindre des passages précis d'un corpus. Les parangons des usages dont nous parlons sont le calcul de spécificités, l'analyse factorielle de correspondances (AFC), la classification hiérarchique[3] ou encore le calcul de cooccurrence. Le premier permet d'établir les formes spécifiques ou sous-spécifiques à telle ou telle partie d'un corpus. Le calcul de spécificité peut s'effectuer sur des chaînes de caractères brutes, à partir d'un étiquetage morpho-syntaxique ou encore après la projection d'un dictionnaire sur le corpus. L'AFC est une méthode statistique qui calcule la proximité distributionnelle de différentes unités d'un corpus pour en offrir une représentation graphique. Elle permet ainsi de mesurer la « proximité » lexicale de sous-parties d'un corpus ou de visualiser la distribution globale du lexique (quelles formes apparaissent fréquemment dans des contextes similaires). La troisième grande méthode lexicométrique est la classification hiérarchique (généralement descendante). Le traitement du corpus aboutit à la constitution de classes lexicales correspondant à la distribution des formes dans le corpus. La méthode Alceste (Max Reinert) est sans doute la plus diffusée dans les divers champs de recherche. Le calcul de cooccurrence enfin, sous forme de recherche des segments répétés (de N formes) ou de recherche de co-présence dans des fenêtres contextuelles dont on délimite la taille, permet de faire apparaître les tendances des formes à apparaître fréquemment l'une avec l'autre.

Nous nous concentrons par ailleurs sur les usages de la lexicométrie dans une perspective d'analyse des discours, c'est-à-dire lorsque l'utilisation des logiciels a pour but de dire quelque chose du contenu du discours (et non seulement de sa structure expressive) éventuellement en fonction de variables : auteur, chronologie, genre, etc. Nous prenons donc ici le syntagme « analyse des discours » dans une définition très large : il s'agit d'*interpréter* des discours, et de tirer de ces interprétations des considérations d'ordre psychologique, social, politique, historique, etc. L'analyse d'entretiens, de discours politiques, de communiqués de presse ou d'ensemble de « tweets » entre par exemple dans cette catégorie.

1. Les interrogations des sciences du texte

Les sciences du texte, dans lesquelles nous pourrons ranger les linguistiques textuelles, la philologie, l'herméneutique et l'analyse du discours[4], cherchent à délimiter avec précision les conditions d'interprétation des textes. Les différentes traditions de ce vaste champ, d'ancienneté et d'homogénéité variables, développent un ensemble de réflexions sur ce qui *fait texte*. Toutes admettent par exemple que la textualité relève de phénomènes qui dépassent largement le niveau de la phrase (et bien sûr du mot). Autrement dit, un texte n'est pas une suite de mots articulés par des règles de syntaxe.

À cette conception « logico-grammaticale » largement dépassée s'oppose une conception « rhétorico-herméneutique » (Rastier, 2003a:3-5). Dans cette perspective, il est admis que les textes sont des entités dont l'interprétation dépend des interactions entre divers paliers de complexité (du trait phonétique ou du graphème au plus bas, à la culture au plus haut, en passant par le morphème, le mot, la phrase, texte, le corpus, etc. ; les paliers locaux étant déterminés par les paliers globaux) sous la rection d'un genre et dans une pratique sociale déterminée.

On comprend que de ce point de vue, une démarche prétendant accéder au « contenu » (le terme même est discutable) de textes à partir de calculs effectués sur les seuls mots graphiques qu'il contient soulève immédiatement un certain nombre de suspicions. Nous traiterons ci-après les trois suivantes : nier à ce point le fonctionnement de la textualité ne revient-il pas à se priver d'un trop grand nombre d'informations ? La recherche de thèmes peut-elle véritablement s'effectuer en comptant des mots ? Et comment être sûr que les calculs que l'on effectue mesurent bien ce que l'on cherche ?

[3]Pour une illustration rapide de chacune de ces méthodes, mises en perspective par rapport à un autre type de calcul courant (l'analyse de similitudes), cf. Marchand et Ratinaud (2012).

[4] Celle-ci ayant réglé son « déficit herméneutique » (Guilhaumou, 2006).

1.1. Le lexicocentrisme de la méthode n'est-il pas rédhibitoire ?

L'une des critiques les plus classiques entendues face à la lexicométrie est sa négation de la textualité : en prenant le « mot graphique » – une simple chaîne de caractères entre deux blancs – comme unité d'analyse, elle fait fond sur une théorie du texte de sens commun qui considère que les mots ont un sens fixe et indépendant du contexte et que le sens des énoncés est une composition du sens des mots qui les constituent.

En effet, le sens n'est pas dans les mots mais dans les énoncés voire plus encore dans les textes (eux-mêmes insérés dans des pratiques sociales et des contextes de production/réception). De leur côté, les textes sont structurés par des phénomènes qui vont bien au-delà de la seule syntagmatique lexicale : instructions génériques, séquences narratives, descriptives, argumentatives, etc., (Adam, 2005), cohérence et cohésion transphrastique (Charolles, 1995), composantes thématiques, dialectiques, dialogiques et tactique (Rastier, 1989)... Enfin, le sens des énoncés, ou plutôt des textes, ne provient pas de la composition de ses mots mais d'interactions multiples à divers paliers, des plus locaux (ponctuation) aux plus globaux (culture).

Rastier (2007) propose ainsi par exemple de substituer à la problématique du *signe* la problématique du *passage*, ouvert sur des cotextes gauche et droit et inscrit dans un corpus d'interprétation. Les modèles dynamiques de la sémiose pensent quant à eux le signe (lexical ou autre) comme un simple moment de parcours interprétatif, compris dans un cercle herméneutique : le sens se perçoit dans un aller-retour entre saisie perceptive du tout et décompositions en unités[5] (Rastier, 2003b). Le sens n'est pas compositionnel : l'interprétabilité des tautologies (« un sou est un sou »), les rétrolectures (« il est républicain mais honnête »), ou simplement les syntagmes figés (« pomme de terre ») montrent bien qu'un parcours interprétatif n'a rien de linéaire ni de simplement additif.

On peut aussi noter que chez les lexicologues eux-mêmes, peu suspects d'anti-lexicalisme primaire, l'unité lexicale reste un objet à manipuler avec précaution : « si lexique il y a, l'unité lexicale, en tant qu'elle est informée par son statut, n'existe pas autrement que comme artefact théorique et heuristique. Non pas construction de la langue, mais des modèles de description, l'unité lexicale relève d'une heuristique que son impensé a fini par rendre transparente au lexicologue » (Petit, 1999).

Par ailleurs, quitte à s'intéresser à une unité minimale de description, n'oublions pas que le mot tel qu'il est considéré par les logiciels est une unité principalement graphique, une simple chaîne de caractères entre deux blancs (avec les problèmes que posent les expressions figées comme « pomme de terre » qui comptera donc pour 3 mots pour certains logiciels). D'un point de vue sémantique comme syntaxique, ce sont les morphèmes[6] qui sont à la base de la combinatoire linguistique, éléments que peu de logiciels prennent en compte. Certaines analyses les négligent même ouvertement lorsqu'elles lemmatisent leurs corpus[7]. La méthode Alceste par exemple supprime tous les morphèmes grammaticaux (pluriels, genre, temps...). Or, ceux-ci ont évidemment un rôle sémantique. La multiplication des pluriels dans un texte par exemple peut participer à marquer une forme d'/intensité/. Autre illustration, les temps verbaux peuvent se corréler, selon les genres, à des positionnements thymiques (émotions positives ou négatives) : Eensoo & Valette (2014) montrent par exemple que dans un corpus de discussion sur le forum médical « Doctissimo », l'emploi du futur est caractéristique des interventions positives (« tu verras, ça se passera bien »…). L'opposition singulier vs pluriel est également un opérateur important de polysémie. Ainsi, sur des corpus politiques, pour des termes tels que « société(s) », « exploitation(s) » ou encore « propriété(s) » le singulier s'homologue à une acception /abstraite/ et le pluriel à une acception /concrète/ (Pincemin,

[5]Dans la modalité visuelle, ce phénomène est bien illustré par les perceptions duales : l'image du canard/lapin est d'abord saisie comme un tout avant que certaines formes ne puissent être saisies comme des oreilles ou comme un bec.
[6]Unité minimale de sens combinable : « rétro-pro-puls-eur-s » résulte de la combinaison de 5 morphèmes.
[7] Le débat à propos de la lemmatisation est classique au sein de la communauté lexicométrique (Brunet, 2000), il ne s'agit pas là de le réactiver car chaque camp possède ses bonnes raisons. Il faut en revanche avoir conscience des implications de ses choix de modification des données.

2012).

Sur un autre plan, certaines analyses comptent les mots qui sont sous- ou sur-utilisés dans telle ou telle partie d'un corpus (calculs de spécificités par exemple). Or celles-ci ne tiennent pas compte du biais introduit par les reprises anaphoriques. Un mot peut n'apparaître qu'une fois dans un texte et pourtant être au cœur de celui-ci, constamment repris par des synonymes ou des pronoms (« celui-ci », « elle », « ce dernier »...). Le discours journalistique, souvent étudié quantitativement, fait d'ailleurs de l'« anaphore infidèle » une règle de bon usage que l'on retrouve dans tous les manuels (on trouvera ainsi dans un court article sportif les substitutions suivantes pour « Roger Federer » : « Le Suisse » ; « L'Helvète » ; « L'homme aux 19 titres du Grand Chelem » ; « Le Maestro »). On voit la faiblesse des « analyses » grand public comme celle que propose par exemple le journal Ouest-France après le débat d'entre-deux tours de la présidentielle de 2012[8] et consistant à ne comparer que l'emploi de mots seuls entre candidats (qui utilise le plus le mot « travail », « Allemagne », etc.). Citons également cet expert politique invité du 20h de TF1 qui déclare que François Hollande n'a employé le mot « France » qu'à quatre reprises dans son dernier discours et en déduit le peu de cas que fait le Président de... « notre pays ». On pourra rétorquer que ces exemples relèvent de pratiques lexicométriques non-scientifiques mais le problème posé par les anaphores reste réel même dans le cas de travaux sérieux. Par ailleurs, cet usage populaire de l'outil révèle un autre des dangers de la lexicométrie : elle permet de donner à des analyses de sens commun une illusion de scientificité et d'objectivité qu'elles ne méritent pas.

Le dernier problème, et non le moins important, que pose le lexicocentrisme est sans doute la typification lexicale. Les occurrences sont réduites à une forme typique sur laquelle se basent les calculs. Cette opération est extrêmement délicate dans la mesure où la pertinence même du concept de type lexical peut être remise en cause[9]. Puisqu'une forme n'a de valeur que par son occurrence dans un contexte spécifique (nous avons vu l'exemple du « travail » qui n'était pas le même chez Sarkozy ou Laguiller), il est possible de se demander si ramener des singularités à une catégorie unique sur la base du fait qu'elles ont la même forme physique ne revient pas à travailler sur de purs artefacts. Cela revient à tout le moins à appliquer une vision documentaire à des données langagières qui ne sont pas prévues pour être traitées comme des étiquettes. Cette vision « réussit dans [son] cadre d'origine, puisque les mots-clés utilisés par les documentalistes sont choisis dans un référentiel (thesaurus, liste d'autorité) et sont intrinsèquement dotés d'un contexte par leur positionnement dans ce référentiel. Elle échoue pour les moteurs de recherche basés sur le modèle de l'espace vectoriel, dès lors que le texte est représenté par ses mots extraits de leur contexte et isolés les uns des autres » (Pincemin, 1999).

À cette critique répond parfois un argument « pragmatique » : l'observation tiendrait sur le plan théorique, mais dans les faits, et vis-à-vis des problématiques posées, la typification n'aurait qu'une influence négligeable sur les calculs. Au fond, la récurrence du mot « travail » nous indiquerait bien que l'on parle de « travail » dans notre corpus. Selon le corpus que l'on a constitué (politique ou obstétrique), l'ambiguïté ne serait de plus pas permise. Cependant, une telle affirmation pose plusieurs problèmes de taille. Le premier est évidemment qu'elle ne peut se supporter de sa seule évidence apparente : il faudrait la prouver.

[8]http://presidentielle2012.ouest-france.fr/actualite/les-mots-des-candidats-passes-au-crible-04-05-2012-1547

[9]Précisons ce point dont la formulation peut être ambiguë. Le fait les types soient des unités pertinentes de description au niveau *sémiotique* de la langue n'est pas à remettre en cause : la systématique lexicale en est un exemple clair. En revanche, que cette pertinence se transfère au niveau *sémantique* est beaucoup plus discutable. De plus, les modèles de la sémantique lexicale qui admettent cette pertinence ne font pas pour autant des types lexicaux des déterminants du sens. Au contraire, ils s'interrogent sur leur format (noyau sémique ? Schème ?) et sur leur fonction (complètement déterminante ? Partiellement ?) dans un modèle de la « construction du sens en contexte ». Aussi, il est important de comprendre que le fait que l'on puisse reconnaître des types lexicaux n'implique pas automatiquement que ces types lexicaux ont un rôle central dans la sémiosis. Comme le notait justement le Saussure des *Écrits de linguistique générale* : « une forme est une figure vocale qui est pour la conscience des sujets parlants *déterminée,* c'est-à-dire à la fois existante et délimitée. Elle n'est rien de plus comme elle n'est rien de moins. Elle n'a pas nécessairement "un sens" précis [...] » (2002 : 37). Benveniste faisait la même remarque : « Dans la langue organisée en signes, le sens d'une unité est le fait qu'elle a un sens, qu'elle est signifiante. Ce qui équivaut à l'identifier par sa capacité de remplir une « fonction propositionnelle ». C'est la condition nécessaire et suffisante pour que nous reconnaissions cette unité comme signifiante. (...) Un tout autre problème serait de se demander : quel est ce sens ? » (Benveniste, 1966 : 127).

Le second est qu'elle fait intervenir un dangereux postulat de sens commun : la pré-notion de « ce dont on parle » passe difficilement le test de la réflexion critique. Le troisième relève des conséquences qu'elle amène à tirer : si l'horizon maximal de l'analyse des textes par ordinateur est le repérage (possiblement maladroit) de simples thèmes très généraux, l'outil ne perd-il pas de son intérêt ?

1.2. Compter des mots suffit-il à déterminer un thème ?

1.2.1. Il existe des thèmes sans termes

De nombreuses utilisations de la lexicométrie visent à détecter des thèmes dans un corpus. C'est le cas par exemple dans l'analyse d'entretiens ou de questionnaires ou encore dans l'étude de corpus de discours politiques par exemple. Signalons tout d'abord que cette utilisation de la notion de « thème », de « contenu » ou de « ce dont il est question » est une utilisation de sens commun. Les recensions de travaux sur la notion de thème (cf. Missire, 2015, Rastier, 1996) font bien voir la profonde complexité de la question, or celle-ci n'est que très rarement abordée en lexicométrie. Bien que ce problème ne nous paraisse pas mineur, admettons que l'on puisse désigner par « thèmes » l'appréhension naïve de « ce dont il est question » dans un texte. Il se pose encore quelques difficultés quant à la manière de les capter automatiquement.

Tout d'abord, on sait qu'un thème peut ne jamais être lexicalisé explicitement : le mot « ennui » n'apparaît que 7 fois dans Madame Bovary alors que l'« idée » d'ennui, elle, y est omniprésente (Rastier, 1996). Ce sont les traits de sens caractéristiques de l'ennui (/longueur/, /lenteur/, /récurrence/, /négativité/...) qui sont présents de manière dispersée dans plusieurs morphèmes voire structures rhétoriques (rythmes, registre de langue, dispositions tactiques…) sans être forcément compactés dans le lexème « ennui » lui-même. Dans cette conception de l'analyse thématique, la ponctuation même doit être prise en compte. En effet, alors qu'elle est ignorée par les logiciels, elles contribuent fortement à la sémantique du texte (Bourion, 1998). Pour reprendre l'exemple précédent, de nombreux points de suspension peuvent participer à thématiser l'ennui. Comme le note justement Mayaffre (2013), l'analyste doit porter « son attention sur toutes les unités du discours, de la lettre aux isotopies, et pas seulement aux lexies ou mots graphiques : en effet, parfois, l'idéologie transpire d'un enchaînement syntaxique surutilisé ou d'un code grammatical suremployé autant que d'un lexique ramené à une graphie ».

Aussi, chercher à déduire de la présence de certaines formes graphiques la présence de certains thèmes, idées, émotions, etc. apparaît délicat.

Certains logiciels qui étiquettent les textes à partir de listes de mots pré-établies nous donnent l'exemple le plus discutable de cette pratique. Citons par exemple le cas des études réalisées à l'aide du logiciel *Diction*[10]. Celui-ci est généralement utilisé pour projeter un dictionnaire de termes considérés comme « positifs » ou « négatifs » *a priori* sur un corpus. Outre le fait que la notion de « positif » mériterait un développement conséquent, plusieurs problèmes se posent. Tout d'abord, les termes ne sont pas « positifs » dans l'absolu mais en fonction du genre de discours (honorable ne l'est pas vraiment pour une thèse !). Ensuite, une telle démarche oublie que les seuls mots ne sont pas porteurs de sens et double son lexicocentrisme d'une théorie naïve de la communication : dans certains cas, ce sont justement les textes comportant les termes les plus « positifs » qui le sont le moins (le *face work* quotidien repose d'ailleurs sur ce principe : « tu es très gentil, je t'aime beaucoup, mais... »). Sur ce principe, Missire (2013:6) montre que dans un corpus d'évaluation de fonctionnaires par leurs supérieurs hiérarchiques, les évaluations négatives sont moins repérables par l'utilisation de mots négatifs que par l'atténuation adverbiale avec « parfois » et « certains ». Enfin, et plus largement, il suffit de s'intéresser aux travaux sur la textualisation des émotions (cf. Micheli (2014) et les notions d'émotion *dite*, *montrée* et *étayée* par exemple) pour mesurer la difficulté d'aborder les textes par leurs mots graphiques.

[10]Une liste des travaux réalisés à l'aide du logiciel est disponible à l'adresse suivante : http://www.dictionsoftware.com/published-studies/#peerarticles.

Ces éléments invitent donc à la prudence lorsqu'il s'agit de tirer des conclusions sur le contenu sémantique des textes à partir de leur vocabulaire seul. Ainsi, quand Emilie Née (2012) observe que la forme « insécurité » dans le journal *Le Monde* est sur-employée dans les rubriques traitant de l'actualité nationale et sous-employée dans celles traitant d'actualité internationale, elle se garde bien de déduire que le vaste monde est un havre de paix. Le constat n'est ainsi pas utilisé comme expression transparente des thèmes traités dans le journal, mais sert de base heuristique à un ensemble d'hypothèses qui pourront être vérifiées, notamment par le retour au texte et la confrontation à des cadres théoriques politiques, sociologiques, communicationnels, etc.

1.2.2. Il existe des termes sans thèmes

Si un thème peut être présent sans être explicitement lexicalisé, inversement, une même lexicalisation peut participer de thèmes tout à fait différents. Les phénomènes de polysémie sont les plus évidents. Salton, Allan, et Buckley (1994) relèvent par exemple « le cas d'un rapprochement entre une phrase sur le football américain, et une phrase sur la théorie des jeux (mathématiques probabilistes), en raison de mots comme 'games', 'play' et 'team(s)' » (Pincemin, 1999).

Au-delà de la polysémie qui peut en partie être contrôlée par la constitution du corpus[11], les habitudes discursives, ou la contamination de domaines sémantiques par d'autres peuvent entraîner des ambiguïtés : il ne serait par exemple pas étonnant aujourd'hui qu'un logiciel regroupe des appels à projets scientifiques et industriels sur la base des liens créés par la diffusion du vocabulaire managérial. Une telle mise en évidence statistique serait bien évidemment intéressante, mais pose la question de l'interprétabilité des sorties machines : les regroupements ne reposeraient pas dans un tel cas sur des similarités thématiques mais sur des similarités formelles (elles-mêmes corrélées à des phénomènes autres que thématiques).

Ensuite, il est aisément possible de dire des choses différentes en utilisant des mots semblables : une simple négation peut inverser la tonalité d'un propos (« je (ne) t'aime (pas) »). Tout aussi trivialement, on doit prendre en compte que le « sens » des mots est dépendant de leur contexte. Mayaffre (2008) montre par exemple que dans les discours de campagne présidentielle de 2007, le mot « travail » est spécifique à deux candidats, Nicolas Sarkozy et Arlette Laguiller. Certes les deux candidats parlent bien de « travail », mais ce constat rend-il véritablement justice au sens que revêt ce terme chez chaque locuteur ? En effet, l'étude montre par ailleurs que les principaux cooccurrents du travail chez les deux candidats sont respectivement « valeur, fruit, revenu, mérite… » et « comédie, marionnette, dupe, criminel… ». Attention encore une fois à l'interprétation rapide, mais il semble bien que les deux locuteurs ne parlent pas du même travail.

Inversement, des mots différents peuvent dire des choses très similaires (la donne est certes un peu plus compliquée puisque, entre autres en raison de phénomènes dialogiques, au sens de Bakhtine, la manière de dire n'est jamais neutre). C'est la raison pour laquelle les discours politiques d'un même locuteur se répartissent différemment sur une AFC selon qu'il parle à la radio, à la télé, en meeting, à l'assemblée, etc. Le contenu ne joue donc pas de rôle dans de tels résultats : c'est le genre de discours qui impose ses effets.

Nous voyons par ailleurs émerger là une question centrale pour les analyses qui regroupent les textes en classes à partir de leur contenu lexical : quelle variable a présidé au regroupement des textes ? Si les

[11]Certains utilisateurs de lexicométrie tendent à prendre pour acquise l'idée qu'il n'y a pas de polysémie dans un genre de discours unique. Dans un corpus de discours politiques, on ne risque pas de confondre le travail de l'ouvrier et celui de la femme qui accouche. La conclusion est néanmoins un peu rapide dans la mesure où ce n'est pas tant le genre que le domaine sémantique (le « champ lexical ») qui exclue la polysémie. Or les corpus politiques, de presse ou littéraires sont rarement mono-thématiques (on les analyse d'ailleurs justement souvent à l'aide de logiciels pour en extraire les thématiques). Dans cette perspective, le risque que la polysémie entraîne des rapprochements statistiques douteux n'est pas à exclure. De plus, nous ne parlons ici que de polysémie en langue, la polysémie entre idiolectes pose de bien plus grands problèmes : ma « liberté » n'est sans doute pas votre « liberté ».

logiciels de classification promettent généralement que c'est la thématique[12], nous pouvons voir avec l'exemple précédent que ce n'est pas forcément le cas. Nous développons ce problème majeur dans la section suivante.

Ainsi, espérer saisir des contenus, des thèmes, à partir du repérage de mots isolés apparaît particulièrement complexe. Pour nuancer cette critique, on peut considérer que si elle s'applique relativement bien aux simples « comptages de mots » (calcul de spécificités par exemple), elle paraît moins pertinente lorsque l'on observe des ensembles de mots partageant des proximités distributionnelles. En effet, « une notion n'est pas toujours mentionnée de la même façon dans les textes, ce qui se traduit par la difficulté de trouver "le bon mot-clé" dans l'interrogation d'un moteur de recherche. En revanche, des regroupements de mots dans le contexte d'un texte introduisent une redondance sém(ant)ique, et l'évocation d'un thème reste dans le cadre global du vocabulaire du domaine : les recouvrements de vocabulaire permettent de trouver les textes en relation, par-delà les variantes de leur expression linguistique » (Pincemin, 1999). L'idée est donc qu'associer un terme à ses cooccurrents participe généralement à sa désambiguïsation.

Attention toutefois car le fait que des termes apparaissent ensemble ne dit rien des relations sémantiques qu'ils entretiennent : cela révèle une proximité physique récurrente et non une mise en texte. Pour pallier ce problème, certains logiciels proposent des fonctionnalités comme celle des segments caractéristiques (Iramuteq, Hyperbase) qui offrent des visions contextualisées de certains constats statistiques afin de pouvoir revenir à lettre du texte. Ainsi, pour reprendre l'exemple du mot « travail » dans les discours de Nicolas Sarkozy étudiés par Mayaffre (2008), le fait que le terme cooccurre significativement avec « valeur », « mérite », et « revenu » ne dit rien en soi. Si l'on considère en revanche le segment caractéristique suivant : « la valeur travail permet d'obtenir les revenus que l'on mérite », on obtient alors des bases plus solides pour faire une hypothèse sur la conception spécifique du « travail » qui est proposée dans le corpus. Mais là encore, rien n'empêche qu'ailleurs dans le corpus se trouve un segment tout aussi caractéristique (statistiquement s'entend) mais très différent : « personne ne mérite d'obtenir un revenu pour un travail sans valeur, avis aux boursicoteurs ! ». L'interprétation de la cooccurrence qui semblait évidente jusque-là devient alors plus problématique. La représentativité des « segments représentatifs » est donc toujours sujette à caution. Mentionnons sur ce point une anecdote (cf. Missire, 2004) qui donne à réfléchir : le logiciel Hyperbase propose comme poème statistiquement représentatif du recueil Baudelairien *Les fleurs du mal*, une œuvre, Tristesses de la Lune, considérée comme mineure par la critique littéraire. Qu'est-ce donc que la « représentativité » statistique en ce cas ?

Une cooccurrence peut donc être ambiguë : on connaît le célèbre exemple de Brunet (2003) sur « amour » chez Rousseau/Flaubert/Proust. Cooccurrent de mots positifs chez les deux premiers (« volupté, joie, tendre, vertu… ») et de « souffrance » ou « angoisse » chez le troisième. Le risque d'interprétation de sens commun est grand, mais on ne peut dire à partir de ce simple constat si, chez Proust par exemple, l'amour est la source de l'angoisse ou au contraire son remède. Missire (2015) montre de son côté que dans le corpus Frantext (textes littéraires du XVI^{ème} au XX^{ème} siècle), le « sommeil » cooccurre significativement avec des locatifs liquides (« plonger dans », « se couler dans », « sombrer », etc.). On croit ici reconnaître la classique métaphore : les individus plongent dans le sommeil ou sombrent dans les bras de Morphée. Or l'analyse des textualisations effectives montre qu'un renversement actanciel important se passe à partir du milieu du XIX^{ème} siècle : avant cette date, ce ne sont pas les individus qui se coulent dans le sommeil mais le sommeil qui coule, plonge, ou se fond en eux. Nous laissons ici de côté les interprétations à donner à cette observation pour souligner que ce phénomène historiquement remarquable aurait été totalement manqué si l'on en été resté à l'interprétation intuitive à laquelle nous invitait le constat statistique détaché de son contexte syntaxique.

Il faut donc se garder de croire que cooccurrence équivaut à contextualisation. En ce sens, les

[12]Le site commercial du logiciel Alceste annonce par exemple : « les classes obtenues représentent les idées et les thèmes dominants du corpus ». Alors même que, nous le verrons, Max Reinert, concepteur et principal théoricien de la méthode possède – à juste titre – une position très différente.

formulations du type « en passant de l'occurrence (le mot seul) à la cooccurrence (la paire de mots), nous effectuons un saut qualitatif décisif et passons de la forme au sens » (Mayaffre, 2014), prises trop littéralement, peuvent apparaître trompeuses. Autre exemple, l'affirmation suivante est très ambiguë :

« Saussure écrivait "Avant tout on ne doit pas se départir de ce principe que la valeur d'une forme est tout entière dans le texte où on la puise, c'est-à-dire dans l'ensemble des circonstances morphologiques, phonétiques, orthographiques qui l'entourent et l'éclairent." L'ensemble des circonstances qui entourent et éclairent le mot, dont parle Saussure, pourrait se définir, strictement, comme l'ensemble de ses cooccurrences. » (*Ibid.*)

Si l'on prête au mot « cooccurrence » un sens très extensif, il est effectivement vrai de dire que le contexte d'une forme est l'ensemble de ce qui « occurre » avec elle. Toutefois, dans le domaine de la lexicométrie (domaine dont traite le texte cité), le terme « cooccurrence » possède une extension bien plus restreinte : les cooccurrences sont simplement les formes lexicales qui apparaissent de manière significative à proximité géographique les unes des autres, l'empan de cette proximité étant déterminé a priori par le logiciel (ou mieux, mais plus rare, par l'utilisateur). En ce sens, assimiler contexte à cooccurrence est faux, car le contexte d'une forme singulière n'est absolument pas la ou les forme(s) qui apparaissent souvent avec elle dans un corpus donné.

L'objectif de ces remarques n'est absolument pas de nier les intérêts nombreux des études de cooccurrences mais simplement de signaler que celles-ci ne doivent pas être tenues pour un idéal de contextualisation.

Dans ces conditions, il faut conclure que seul le recours systématique à un concordancier permet de passer de constats statistiques à des interprétations assurées. C'est-à-dire que le constat statistique n'est qu'heuristique : il propose une direction à la lecture, mais le retour à la lettre du texte apparaît indispensable. Nous voyons encore poindre ici le problème des (très) gros corpus aujourd'hui analysés informatiquement pour lesquels même la lecture de contextes de concordancier devient humainement impossible. L'interdiction catégorique de revenir au texte est dans ce cas-là très invalidante.

1.3. Les analyses classificatoires permettent-elles vraiment de mettre au jour des thèmes ?

Cette section pointe un problème majeur des usages lexicométriques non spécialistes (sociologie, sciences politiques, sciences de gestion, psychologie, sciences de l'information et de la communication…). Dans ces disciplines, les utilisateurs présupposent que les traitements statistiques (analyses de similitudes, méthodes Alceste, *topic modeling*) permettent de mettre au jour des thèmes[13].

Nous avons pointé précédemment les difficultés que soulevaient cette entreprise dès lors qu'on tentait de ne la mener qu'en s'intéressants aux mots isolés. Mais admettons ici, que ces éléments ne sont pas déterminants, et cherchons à poser le problème d'une autre manière. Admettons donc que les analyses de cooccurrences, les analyses de similitudes, les analyses classificatoires *peuvent* effectivement mettre au jour des thèmes : est-on bien sûrs que c'est ce qu'elles font ?

Une classe, un *topic*, une analyse cooccurrentielle quelconque nous dit que tel mot, dans tel discours, cooccurre, ou voisine, significativement avec tel autre. Mais que veulent dire ces cooccurrences et ces proximités ? Le simple constat que telle chaîne de caractères est, dans tel corpus, souvent présente à proximité géographique de telle autre permet-il réellement de porter une conclusion ?

La question de ce que l'on mesure lorsque l'on cherche des cooccurrences ou lorsque l'on opère des

[13] À la décharge des profanes, il faut admettre que les outils avancent maquillés : le *topic modeling* doit bien chercher des *topics* et le site d'Alceste affirme que « les classes obtenues représentent les idées et les thèmes dominants du corpus ». À y regarder de plus près, on s'aperçoit pourtant que la littérature sur le *topic modeling* enregistre l'ambiguïté du terme et s'arrête à définir un *topic* comme « *a recurring pattern of co-occurring words* » (Brett, 2012) ; et que le concepteur et théoricien d'Alceste, dans une approche plus scientifique que commerciale du logiciel, se contente de parler de « mondes lexicaux ».

classifications est d'autant plus sensible que l'on sait que les constats statistiques portant sur les hautes fréquences ne mesurent généralement pas les mêmes phénomènes textuels que ceux portant sur les basses fréquences[14] : « la conclusion, confirmée par d'autres monographies (par exemple Balzac, Verne, Zola, Proust, Anatole France), est que le choix du genre et du sujet impose davantage sa loi dans les basses fréquences. […] Mais les fréquences hautes n'en sont pas moins animées de mouvements qui paraissent plus lents mais plus profonds et qui décrivent sourdement l'évolution de l'écriture. Ces mouvements de fond sont sans doute moins conscients ou moins volontaires que les choix clairs que l'écrivain fait parmi les genres et les sujets. Plus stylistiques que thématiques, ils sont davantage le reflet de la structure que du contenu » (Brunet, 2009).

On notera à profit que le concepteur et principal théoricien d'Alceste, Max Reinert, s'oppose justement à une conception des classes lexicales comme des classes de contenu (Reinert, 1990 ; 1993 ; 2007)[15]. Son approche, marquée par la psychanalyse, traite de « mondes lexicaux » qui renvoient à des traces de positions énonciatives. Cette proposition laisse la place à des qualifications très diverses, et non pas seulement thématiques, des classes lexicales[16]. Prise au sérieux, cette approche impose donc à l'analyste une démarche inductive : l'outil met en lumière des régularités qui sont *peut-être* intéressantes (nous abordons le cas des artefacts statistiques et autres déformations dans les sections suivantes), et l'analyste cherche alors à mobiliser des théories diverses (dont, inévitablement, des théories du texte), pour leur donner du sens. Dans cette optique, il faut bien noter que la problématique de recherche est *postérieure* à l'application de l'analyse statistique. Cette démarche est pleinement valide et fort intéressante, mais n'est pas, il faut le regretter, la plus répandue. Au contraire, l'utilisation des classifications est le plus souvent postérieure à la formulation de la problématique : on cherche à extraire des textes ce que l'on suppose déjà être des thèmes.

Aussi, lorsque l'on analyse des résultats comme la projection d'auteurs sur des graphes ou des classes, on ne peut affirmer que la connexion est thématique : « les paramètres en cause dans le contenu lexical sont multiples : le sujet, les thèmes, le genre, l'auteur, l'époque, la longueur du texte, sa dynamique propre, etc., et il est impossible de les isoler pour mesurer précisément l'influence de chacun d'entre eux. » (Bernet, 2009).

Typiquement, lorsque l'on opère une classification sur un corpus de polémique sur des réseaux sociaux, il est fréquent d'obtenir une classe qui réunit les insultes et les grossièretés : il ne s'agit pas là d'une classe thématique mais bien stylistique. Certes, il arrive que certains styles se corrèlent avec certains thèmes ; cela est fort intéressant en soi (nous retrouvons là en partie le concept de formation discursive (Maingueneau, 2011)), mais ne change rien à l'affaire car il reste que sans investigations plus profondes on peut craindre que moins que des thèmes, ce sont des genres que repèrent les classifications. Voici par exemple la conclusion d'un article commentant les classes obtenues après analyse statistique d'un corpus de discours sur le climat : « D'un point de vue sociologique, les analyses nous ont permis de mettre en

[14] Rappelons que nombre d'analyses lexicométrique, notamment sur les très gros corpus, ne travaillent que sur les hautes fréquences (jusqu'à seulement les 10% les plus hautes).

[15] On trouve d'ailleurs sur la page de membre de son laboratoire la remarque suivante dans la section « activités de recherche » : « un fait d'expérience est que cette méthode Alceste est souvent utilisée pour préparer une analyse de contenu, alors que son algorithme ne recouvre qu'une analyse automatique purement formelle du texte. Un thème de cette recherche est justement de travailler les rapports entre le sens où s'engage un chercheur relativement au "contenu" qu'il croit percevoir dans les résultats d'une analyse, et la structuration d'une lecture possible du texte étudié… »

[16] Cf. par exemple les articles où il propose de considérer des « classes lexicales stabilisées », c'est-à-dire des classes lexicales similaires dans des corpus pourtant très différents (récits de cauchemars, entretiens sur la guerre d'Algérie, œuvres de Nerval…). Reinert (2007) voit là une stabilisation culturelle et donc transdiscursive de grands modes d'énonciation qui sont celui du témoin (prépondérance du vocabulaire des sensations et des images concrètes), de l'acteur (prépondérance du vocabulaire des affects et des actions) et du patient (prépondérance du vocabulaire de l'abstrait et du distal) qu'il relie aux modes lacaniens Imaginaire, Réel et Symbolique ou encore aux peirciennes priméïté, secondéïté et tercéïté. Il est intéressant de noter que le créateur d'Alceste est bien loin de considérer ses classes comme des extractions de thématiques mais au contraire qu'il cherche à mobiliser une théorie complexe du texte comme du psychologique pour interpréter les observations statistiques.

évidence une tendance significative à une forme de typification des discours en fonction des catégories d'acteurs [institutionnels, politiques, médiatiques, etc] » (D'Apollonia, Luxardo, Piet, 2014). Il est clair qu'ici, ce qui a été capté par les statistiques n'est pas le « sens » des discours ni même leur « contenu », mais bien leur forme typifiée par une pratique sociale particulière.

À partir de ce constat, il est encore possible de pousser plus loin la suspicion. En effet, dès lors que l'on remarque que certaines classes proposées par une classification ne sont pas thématiques mais stylistiques, il devient légitime de se demander si celles qui *paraissent intuitivement* thématiques le sont réellement[17]. En effet, une telle conclusion ne peut se soutenir du seul fait que l'analyste pense reconnaître un champ lexical pertinent dans la suite de mots qui lui est présentée. Donner du sens est compulsif chez l'être humain (Rastier, 1999)[18], et il est toujours possible de voir une cohérence dans les signes les plus aléatoires. Comment donc ne pas craindre la simple projection des attentes (ce qui est grave) ou des savoirs déjà constitués (ce qui l'est moins) de l'analyste ? La réponse à ce problème se trouve bien évidemment et encore une fois dans le retour au texte (et non simplement à ses « segments représentatifs », nous avons déjà discuté ce point) : l'outil sert alors à mettre en avant des phénomènes que l'analyste investigue plus avant, au ras du texte, pour valider leur intérêt.

On pourra arguer que la difficulté que nous pointons là n'est donc pas un problème de la méthode (donc de l'outil) mais de l'usage (donc de l'analyste). Autrement dit, il pourra être rétorqué que le logiciel nous met face à des données et que leur interprétation relève tout simplement du travail d'analyse[19]. À l'analyste donc d'étayer les raisons pour lesquelles il considère que telle classe de mot peut être corrélée à telle thématique dans le corpus analysé. Toutefois, nous pensons que le problème est plus complexe que cela : d'une part parce que le logiciel porte avec lui ses présupposés et incite à un usage a-critique ; et d'autre part parce qu'il n'est pas certain que les sorties machines soient réellement interprétables. Ces deux points sont traités dans les sections suivantes.

2. Les interrogations des statistiques

Nous avons vu dans la section précédente les différentes inquiétudes que pouvaient exprimer les sciences du texte quant à la pertinence d'analyses quantitatives lexicocentrées pour étudier le contenu de textes. Une autre grande catégorie de griefs adressés à la lexicométrie relève de considérations plus générales sur les effets des traitements statistiques sur l'objet étudié. En effet, l'analyste transforme des données langagières cohérentes en un ensemble de données statistiques sous forme chiffrées ou graphiques. Dans cette mesure, il est possible de se demander si cette déformation qualitative des données ne pose pas de lourds problèmes de lisibilité ; si les données statistiques conservent bien un lien pertinent avec l'objet qu'elles sont censées décrire ; et si la complexité des traitements qui les produisent n'interdit pas leur interprétation.

2.1. Comment « lire » les sorties-machine ?

Un phénomène particulièrement frappant lorsque l'on utilise un logiciel de lexicométrie est la facilité avec laquelle semblent pouvoir s'interpréter les résultats des analyses. Or, l'analyste doit à tout prix se méfier de l'effet « monsieur météo » que les sorties machines sous forme graphique tendent à engendrer. Nous faisons ici référence à la pratique qui consiste à proposer pour toute analyse un commentaire vague et *ad*

[17] Le problème se pose avec encore plus d'acuité lorsque l'on a affaire à des classes dites « inexploitables ». Si le logiciel nous soumet des résultats qui apparaissent dénués de sens, comment affirmer que ceux auxquels *nous* donnons du sens sont réellement pertinents ?

[18] « L'activité interprétative spontanée, compulsive et incoercible des sujets se déploie particulièrement sur les formations sémiotiques. Elle les conduit à interpréter même des non-mots, ce pourquoi nous avons pu dire que l'homme est condamné au sens. »

[19] À ne pas confondre avec l'idée pourtant répandue que le logiciel fournit des *résultats* (donc déjà encadrés par un *cadre théorique*, une *problématique* et une *méthodologie*), qu'il ne reste plus qu'à *discuter*, pour employer la terminologie de la méthode scientifique orthodoxe.

hoc des nuages de mots présentés sous différentes formes.

Maurice Tournier pointait déjà ce risque en parlant de « mirage lexicométrique » pour désigner « ces longs enchaînements suscités par une lecture (trop) spontanément transitive des graphes, et qui simulent des constructions linguistiques en fait non réalisées en corpus. » (Tournier, 1985).

Les logiciels offrent des représentations qui invitent à une interprétation intuitive (« drogue » et « peur » apparaissent à proximité, donc la drogue fait peur...) et la facilité d'usage pousse parfois à l'usage facile. Les évolutions apportées aux différents logiciels qui consistent souvent en de simples ajouts cosmétiques (nouveaux types de graphes, représentations 3D, variations colorimétriques...) tendent à renforcer la dimension « ergonomiste » de l'outil, c'est-à-dire l'invitation à un usage naturalisé et a-critique.

Nous pouvons mettre en parallèle le danger que nous pointons ici avec les critiques des outils de cartographie en analyse des controverses. Comme le montre précisément Yves Jeanneret, la transposition de réalités en cartes, graphes ou tableaux, loin d'être neutre, influence leur perception et invite à des lectures spécifiques qui peuvent éloigner grandement l'observateur de l'objet qu'il souhaitait explorer (Jeanneret, 2013). Ainsi, l'usage d'outils de traitement et de reconfiguration automatique des données « suscite un regard surplombant et a pour effet de présenter un savoir totalisant, sur le mode de la révélation. Au-delà des "prétentions cartographiques" qui se déploient dans les formats de visualisation, et qui posent déjà problème dans la dénomination "cartographie des controverses", il faut considérer que ces outils informatiques, qui voudraient présenter des relations entre des acteurs, procèdent d'un traitement quantitatif de réalités qui sont avant tout des éléments de nature documentaire, c'est-à-dire dont l'existence, en tant que textes, est régie par des formats éditoriaux qui encadrent leur production. Les traitements successifs que ces dispositifs opèrent sur les documents prélevés – agrégations de productions écrites hétérogènes, requalification comme "traces d'usages", calculs statistiques, visualisations homogénéisantes et neutralisantes – leur font subir des transformations souvent importantes, des méta-morphoses. Ces dispositifs suscitent notamment une invisibilisation d'un certain nombre de médiations pourtant cruciales dans l'appréhension et l'interprétation des textes observés. Or, ces informations ont un sens social et communicationnel indispensable à analyser pour comprendre ce qui se joue sur les réseaux entre les acteurs. De fait, en extrayant des éléments de nature documentaire de leur contexte d'emploi, c'est-à-dire en effaçant les médiations éditoriales qui les font exister dans le réseau (médiations prises en charge par les propriétés techno-sémiotiques des dispositifs de réseautage social), on prend le risque de les "trahir", en quelque sorte, en les dotant d'un sens nouveau. » (Bigot, Julliard & Mabi, 2016).

Un cas typique de ce phénomène en lexicométrie est la tendance généralisée à interpréter des distances entre points sur une AFC. Une AFC étant l'écrasement en deux dimensions d'un graphe comportant un très grand nombre d'axes (et donc de dimensions), les distances observables qui en résultent deviennent proprement ininterprétables. La représentation graphique, doublée d'une méconnaissance des méthodes statistiques, peut ainsi induire des parcours interprétatifs douteux.

2.2. La déformation des structures textuelles par les procédures mathématiques ne conduit-elle pas à de purs artefacts ?

Plus problématique encore que la simple « déformation » induite par le changement qualitatif des données étudiées est le risque de travailler sur de purs artefacts, des *chimera topics* (Schmidt, 2012). Puisqu'un outil n'est jamais qu'une théorie matérialisée, il est possible que l'inexactitude des hypothèses sur les fonctionnements textuels et signifiants que le logiciel porte en lui-même (à commencer par le choix du mot comme unité d'analyse pertinente) mène à des résultats inexploitables.

Par exemple « le modèle de l'espace vectoriel est lourdement tributaire de la désarticulation du texte sous forme d'une série de mots, coupés de leur ancrage contextuel et isolés sur des dimensions orthogonales. Seule la pondération introduit un rééquilibrage global, au pouvoir expressif réduit. Le calcul de rapprochement ne fournit qu'un score cumulatif, qui dans le meilleur des cas réunit a posteriori des mots

en relation d'isotopie, mais qui peut tout aussi bien juxtaposer des mots sans relation significative entre eux, issus de contextes différents. Du caractère numérique et additif de la mesure dérivent deux cas pathologiques opposés : un rapprochement effectué sur un seul mot à forte pondération (mais dépourvu de contexte), et un rapprochement résultant d'une accumulation de mots de faible importance et sans lien sémantique consistant. » (Pincemin, 1999).

Aussi, les statistiques proposent une image au moins biaisée de la réalité textuelle. Typiquement, un nuage de mot ou une proximité lexicale revient à présenter de manière rapprochée des formes qui, en texte, ne le sont pas tant que ça. L'analyste est alors invité à prêter à ces groupes de formes un poids indu, comme par exemple lorsqu'il infère la présence d'un noyau thématique qui n'existe en fait pas véritablement dans le texte : « nos pratiques consistent le plus souvent à croiser, dans un tableau rectangulaire (tableau de contingence) des textes d'un côté et des unités textuelles élémentaires que sont les mots (lemmatisés ou non) de l'autre. Sur les nuages ainsi obtenus […], dont il ne s'agit pas de remettre en cause la puissance exploratoire et heuristique, le danger existe que l'on imagine que les mots voisins cooccurrent. Puisqu'ils sont à proximité dans un même quadrant, les mots entretiendraient des relations cooccurrentielles, appartiendraient peut-être aux mêmes thèmes, relèveraient d'une même isotopie. Conclure de la sorte ne constitue pas une erreur à coup sûr, mais est un raccourci méthodologique et une surinterprétation de la méthode » (Mayaffre, 2014).

Dans le même ordre d'idées, le binarisme des calculs cooccurrenciels ou classificatoires pose un problème théorique majeur : « à partir d'un ensemble initial d'items. La procédure s'effectue en de multiples étapes élémentaires, qui à chaque fois considèrent deux entités : deux éléments, un élément et une classe, ou deux classes (deux sous-ensembles à agréger pour une classification ascendante, ou deux parties résultant d'une scission pour une classification descendante). Or, même si « plusieurs » commence à « deux », et que deux occurrences de sèmes soient le minimum pour étayer une isotopie [c'est-à-dire deux membres d'un champ lexical pour étayer le-dit champ], rien ne permet de conclure que les constructions et les interactions linguistiques, notamment sémantiques, se laissent analyser en interactions binaires. Considérons les classifications de mots en fonction de leur cooccurrence dans un voisinage de l'ordre de la phrase : l'idée est de grouper peu à peu des mots à partir de dépendances syntagmatiques d'un mot avec un autre. Or les structures actancielles mobilisent une constellation de rôles, ou encore les prédicats peuvent avoir un nombre variable d'arguments, deux mais aussi bien un ou trois. De même, deux occurrences d'un sème ne sont pas toujours suffisantes pour ancrer une isotopie, c'est plutôt une convergence sémantique d'un ensemble d'occurrences qui confirme la présomption d'isotopie. La validité d'une décomposition binaire des interactions linguistiques suppose la possibilité d'une modélisation compositionnelle de la langue, ce que théorie et observations tendent à infirmer (Nazarenko, 1998). Pour être plus proche des réalités linguistiques, il faudrait donc « repenser les fonctions de similarité, de distance, d'association, pour construire et introduire dans les traitements des fonctions d'évaluation globale de cohésion, de concentration » (Pincemin, 1999).

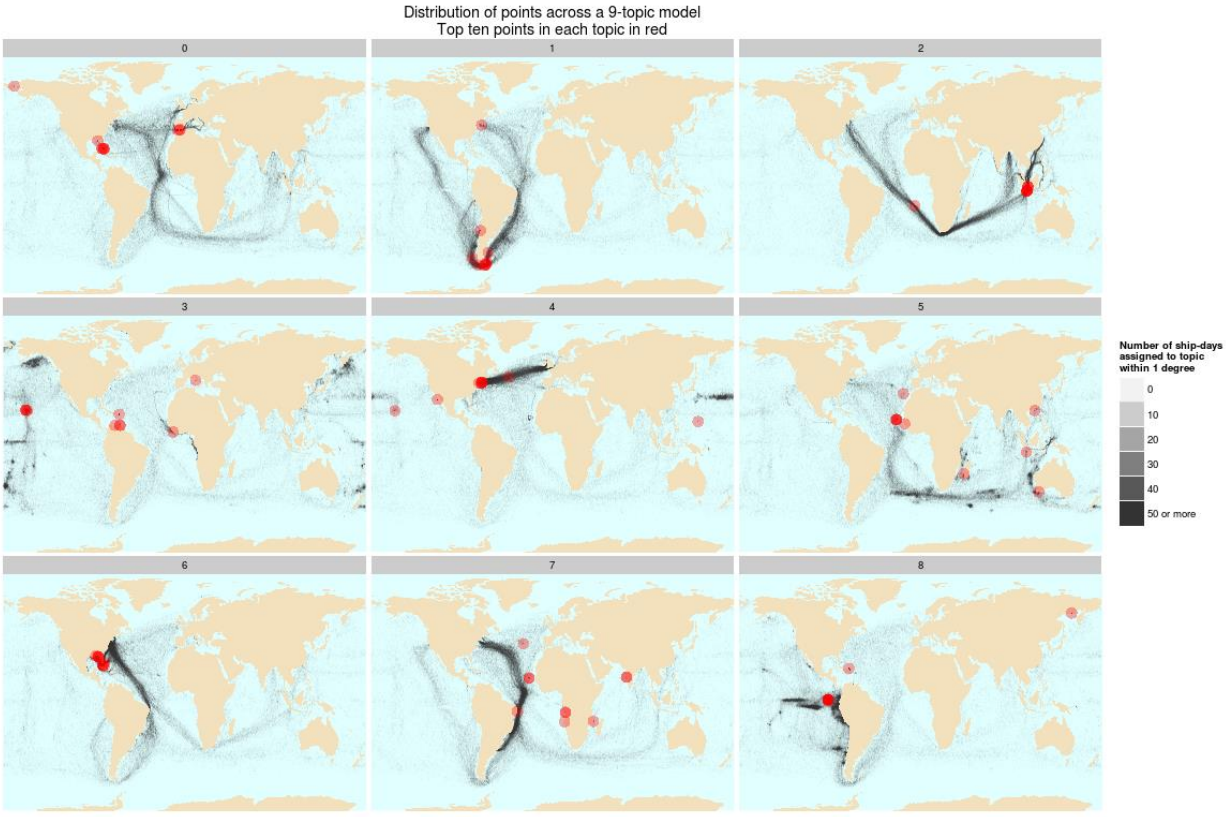
Dit plus clairement encore, « les algorithmes produisent naturellement une représentation déformée des associations sémantiques : les contraintes mathématiques, qui forcent la répartition complète des mots dans des classes disjointes, induisent des distorsions linguistiques » (*Ibid.*).

Les excellents travaux de Schmidt (2012) donnent une ingénieuse illustration de ce phénomène. Afin de questionner les trois grands présupposés des analyses classificatoires dont nous avons déjà parlé (l'idée que les classes sont cohérentes, signifiantes et stables à travers le corpus), Schmidt applique une méthode classificatoire non sur un corpus de textes, mais sur un corpus d'enregistrements de trajets de navigation réalisés par l'US Navy au cours du 19ème siècle. Chaque « texte » du corpus correspond donc à un voyage de bateau, et chaque « texte » est composé de « mots » qui sont les coordonnées géographiques du bateau enregistrées à intervalle régulier. L'intérêt d'un tel corpus est que chaque « texte » possède une référence fixe : il correspond à un trajet de bateau (par exemple, New-York – Le Havre). De ce fait, des classes réunissant des « mots » (coordonnées GPS) appartenant à des « textes » (trajets de bateau) qui n'ont rien à voir apparaissent immédiatement suspectes. En effet, il serait étonnant de trouver des coordonnées

GPS du Pacifique Sud dans une classe contenant des coordonnées GPS de trajets intra-européens. Par ailleurs, les résultats de la classification peuvent être projetés sur une mappemonde pour être visualisés[20]. Ainsi, les résultats attendus de l'analyse sont des classes correspondant aux trajets typiques de bateau au 19ème siècle, à savoir par exemple les trajets commerciaux USA-Europe, Amérique du Nord-Amérique du Sud, les zones de pêche principales, etc. L'analyse montre qu'en effet, les méthodes classificatoires aboutissent à ce type de résultats. Mais deux problèmes sont à relever.

Le premier est que certaines classes qui apparaissent globalement cohérentes mélangent néanmoins des trajets qui ne possèdent strictement aucune caractéristique commune (des trajets commerciaux de 1870 en Inde et de la chasse à la Baleine dans le détroit de Béring en 1810 par exemple). Dans le cas de l'analyse de texte, des interprétations expliquant de tels recouvrements seraient aisées et pourraient mener facilement à des conclusions erronées. On imagine également bien comment l'*insight* fourni par le logiciel ouvrirait la voie à des biais de confirmation divers. Aussi, le présupposé majeur des méthodes classificatoires qui est que les classes produites sont cohérentes et signifiantes doit être considéré avec précaution.

Le second problème que soulève l'expérimentation de Schmidt est plus problématique encore. En effet, il remarque que si l'on ne s'intéresse qu'aux dix ou quinze premiers « mots » de chaque classe, comme il est généralement recommandé de le faire en analyse de textes, on s'aperçoit que l'on obtient une représentation très déformée du « sens global » du cluster. Le schéma suivant permet de visualiser le phénomène.



Projection sur une carte des 9 classes issue de l'analyse des trajets maritimes (Schmidt, 2012)

[20] Pour donner un exemple peut-être plus parlant, on peut imaginer le même type de manipulation avec des trajets de train et le nom des gares traversées à la place des coordonnées GPS. Ainsi, chaque trajet serait un texte de type « Toulouse – Montauban – Castelsarrasin – Agen – Marmande – Bordeaux »… Projeter la localisation des gares sur une carte tracerait ainsi le trajet du train.

Chaque case correspond à une classe. Chaque point noir correspond aux coordonnées géographiques de la classe. On reconnaît globalement les grands trajets de navigation typique dont nous avons parlé. En revanche, si l'on observe les points rouges qui correspondent aux « *top ten points* » de chaque cluster, on se rend compte qu'une interprétation basée uniquement sur ces données serait largement biaisée. Le cluster 6 par exemple qui correspond visiblement aux trajets USA –Amérique du Sud pourrait n'être lu que comme correspondant à la navigation dans le golfe du Mexique. Le cluster 7 présente quant à lui des *top points* éparpillés et qui ne se trouvent même pas sur le trajet principal que semble capter le cluster. Plus généralement, on voit combien les top points induisent une vision du cluster différente de celle induite par l'ensemble des points du cluster (souvent bien plus cohérente). Cette observation nous semble mettre puissamment en cause les pratiques d'analyses de texte les plus lapidaires qui ne se fondent que sur la lecture des premiers mots d'une classe.

Enfin, l'article de Schmidt bouscule un autre présupposé de l'analyse classificatoire : l'idée que les classes désignent des phénomènes généraux dont les occurrences particulières sont *constantes*. C'est-à-dire que la classe telle qu'elle apparaît à l'analyse du corpus est globalement identique à la classe telle qu'elle apparaît dans les différents textes du corpus. Or, Schmidt montre que des divergences très profondes apparaissent lorsque l'on observe l'ordre dans lequel apparaissent les mots d'une classe en fonction des parties du corpus qui participent à former cette classe. En analysant un corpus, puis en le divisant en deux en fonction d'une variable quelconque (temporelle en l'occurrence) on peut obtenir deux classements des mots des classes selon que l'on s'intéresse à l'une ou l'autre des deux parties. Or ce faisant, on s'aperçoit que ce classement peut être très différent, au point de faire apparaître deux « thèmes » distincts. Ainsi, postuler qu'une analyse classificatoire mesure un phénomène constant et stable au sein d'un corpus est une erreur : on ne fait que mesurer un phénomène propre à un artefact massif – le corpus – qui doit être compris comme un tout et non pas comme une collection d'éléments qui possèdent leur propre cohérence[21]. Le problème est le même que lorsque l'on étudie, par exemple, l'âge d'une population. Sa moyenne est une caractéristique statistique de l'ensemble « population », mais ne nous dit rien, surtout en l'absence de mesure de dispersion, de l'âge de ses membres. Ce point d'épistémologie statistique élémentaire prend évidemment une ampleur considérable lorsque l'on considère les calculs autrement plus complexes qu'opèrent les algorithmes et surtout le fait qu'ils n'ont pas lieu sur des chiffres (qui sont des symboles) mais sur des mots isolés (qui n'en sont pas).

Ainsi, les classes dégagées par une analyse classificatoire ne fonctionnent pas sur le modèle type-occurrence avec les textes du corpus, elles ne sont pas des types qui occurrent dans les textes, elles sont des phénomènes uniquement propres au corpus *per se*. La statistique porte donc un point de vue *holistique* et non *analytique* sur un corpus. Or c'est souvent le second qui est présupposé en analyse de contenu assistée par ordinateur.

Nous avons là une illustration de la maxime connue que les traitements statistiques dégagent des phénomènes qui s'appliquent à la masse de données mais pas aux données elles-mêmes. Autrement dit, les classes relevées sur un corpus font sens si l'on considère ce corpus comme un tout mais pas si l'on s'intéresse aux textes eux-mêmes. Lors d'une analyse voulue thématique, cela signifie que l'on saisit des « thèmes » du corpus total – quoi que cela veuille dire – mais que ces « thèmes » n'ont finalement pas de rapport avec les thèmes qui sont présent dans les textes qui constituent ce corpus. Cela signifie également que tenir un propos sur les données à partir de l'analyse de la masse de données, comme c'est le cas lorsque l'on projette des variables (auteur, date, CSP…) sur les résultats d'une classification, est un saut

[21] Pour donner une image plus accessible de ce qu'il se passe, cela revient à distinguer l'étude d'une table *en tant que table* de l'étude d'une table en tant qu'ensemble d'éléments (des pieds, un plateau, des angles, une longueur, une largeur, un matériau, des atomes…). Les considérations qui s'appliquent à l'objet table ne s'appliquent pas pour autant à ses constituants : on peut manger sur une table, pas sur un pied de table. De même, la forme, la taille, le poids de la table, s'ils dépendent de l'agencement et des caractéristiques de ses parties, n'en disent rien. Aussi, de la même manière que le mode d'existence épistémologique de la table-comme-tout est différent de celui de la table-comme-collection (elles ne sont pas justiciables des mêmes interrogations), le mode d'existence du corpus comme tout n'est pas le même que celui des textes qui le composent.

particulièrement risqué.

2.3. Peut-on travailler avec une boîte noire (voire deux) ?

Ces différentes remarques rendent particulièrement saillante une difficulté pourtant souvent occultée : celle de travailler avec des outils de modélisation mathématique dont la maîtrise demande une spécialisation poussée.

Comme le précise la documentation de la plateforme textométrique construite autour du logiciel TXM (ENS-Lyon)[22] : « la bonne compréhension des traitements ne relève pas ici d'une simple curiosité technique, mais des conditions nécessaires pour une juste appréciation et utilisation des résultats proposés. L'un des enjeux d'une plateforme textométrique ouverte est d'expliciter (à tous les niveaux : théoriques, informatiques, méthodologiques) les fonctionnalités disponibles. Une telle maîtrise possible du fonctionnement de l'outil donne accès à une compréhension juste et efficace des résultats des interrogations ».

Cette maîtrise apparaît d'autant plus importante avec la complexification des fonctions statistiques : si des cooccurrences de premier degré par exemple peuvent se représenter assez aisément dans l'esprit d'un analyste, ce qui se passe réellement dans une AFC ou dans une LDA n'est pas à portée de conceptualisation humaine immédiate.

Évidemment, les outils ont été conçus par des spécialistes afin d'être utilisables par des non-spécialistes. Aussi, tout comme le neuroscientifique n'a pas besoin de savoir réparer un scanner pour l'utiliser, l'utilisateur de logiciels de statistique textuelle ne peut être sommé de savoir coder un algorithme de LDA. Néanmoins, l'un comme l'autre doit avoir une conscience au moins générale de ce qui se passe dans sa machine et surtout de ce qu'elle est censée mesurer.

La maîtrise au moins générale des calculs opérés sur les corpus est ainsi une condition importante non seulement de la compréhension des sorties-machines mais aussi de la pertinence ou non des différents outils en fonction des types de problématiques et des types de corpus. L'usage de la lexicométrie n'est pas l'usage d'un bouton magique mais une pratique complexe de tâtonnements mettant en jeu un nombre de connaissances portant sur le corpus comme sur les outils.

Au final, « l'analyse statistique des textes est sans doute, comme l'Académie fondée par Platon, un domaine où nul ne devrait entrer s'il n'est géomètre, mais aussi un domaine d'où nul ne devrait sortir en criant Eureka s'il n'est (pas du tout) statisticien. » (Lebart, 2012).

Par ailleurs et enfin, ajoutons à l'attention du sens commun que l'analyse statistique en question est analyse statistique *textuelle*. Or celui qui fait de la statistique textuelle en n'étant ni statisticien, ni linguiste (du texte !) tend à multiplier les boîtes noires.

3. Synthèse des remarques

Rappelons, pour condenser, l'ensemble des problèmes ci-dessus exposés. Insistons par ailleurs sur le fait que nos questions ici ne sont pas des questions de détail. Les minimiser en les considérant comme des arguties de spécialistes et se contenter de les ignorer ne nous semble pas la voie raisonnable à suivre.

En effet, si l'on rappelle les étapes de détextualisation opérées par une méthode de *topic modeling* ou une classification hiérarchique de type Alceste, un travail de légitimation clair et argumenté apparaît indispensable. Rappelons que ces méthodologies appliquent sur un corpus de textes les opérations suivantes :

[22] http://textometrie.ens-lyon.fr/

- sélection de la forme lexicale comme unité pertinente de description linguistique (postulat de sens commun largement battu en brèche par la longue tradition textualiste) ;

- réduction de la forme lexicale à une chaîne de caractères entre deux blancs (excluant de fait le morphème un degré en-dessous, le figement un degré au-dessus et les unités du plan du contenu comme la narration ou le cadrage par exemple) ;

- réduction des occurrences lexicales à leurs formes lemmatisées (écrasant les différences entre travail et travaux, entre lumière et lumières, etc. ; entre les temps grammaticaux ; entre les genres) ;

- suppression des mots-outils, par opposition aux mots-pleins (gommant par exemple les effets importants de la négation et l'intégralité de la structure logico-argumentative avec la disparition des connecteurs donc, mais, cependant, enfin…) ;

- suppression des fréquences basses (jusqu'à 90% des formes pour l'analyse des très grands corpus ce qui pose tout de même un sérieux problème de précision) ;

- réduction de ces occurrences lexicales (pleines, lemmatisée et de haute fréquence) à des types (opération fort douteuse puisque la signification lexicale est un artefact du grammairien ou du pédagogue : les occurrences du mot « liberté » dans un corpus ont peu à voir avec une définition stabilisée et hors contexte telle qu'on pourrait la trouver dans un dictionnaire) ;

- suppression de la syntaxe ou réduction de celle-ci à une simple topographie : des formes graphiques partageant des contextes graphiques sont appariées, sans que cet appariement ne puisse être caractérisé (cooccurrence de premier ordre ? De second ordre ? Équivalence distributionnelle ? Figement ? Ou simple participation à des contextes vaguement similaires ?) ;

- pour finir, recombinaison spatiale des formes graphiques par des opérations statistiques non représentables pour l'esprit humain et dont le paramétrage, particulièrement complexe, peut influencer lourdement les résultats. Les nouvelles données sont alors livrées sous une nouvelle forme sémiotique (le graphe, la classe de mot, le nuage…) qui reste à interpréter. Les règles de cette interprétation restent par ailleurs nébuleuses.

Ainsi, considérant le nombre important d'étapes et le fait que chacune d'elles puisse être l'objet de contestations solides, l'hypothèse qu'une sortie machine soit en fait un artefact totalement inexploitable ne paraît pas farfelue. Elle mérite en tous les cas un examen sérieux, qui ne peut être remplacé par la contestation molle de chacun des points pris séparément. L'analyse informatisée des données textuelles apporte un nouveau *mode de lecture* des textes, elle permet des enquêtes sémantiques et philologiques de grande ampleur, elle constitue un outil supplémentaire dans l'attirail de l'analyste du discours pour explorer le rôle et le fonctionnement du sémiotique dans la vie humaine : il serait dommage que ces formidables possibilités soient éclipsées par une appropriation négligente. Pour progresser et s'affermir, le champ doit être capable d'entendre, de comprendre, et d'intégrer les observations des spécialistes des textes.

Conclusion

Les différentes questions posées ci-dessus et dans le second volet de ce travail (Carbou, 2017) ne prétendent aucunement nier les intérêts épistémiques et la puissance heuristique de la lexicométrie. Elles entendent simplement souligner que l'outil doit être utilisé avec précaution. Elles indiquent également le type de problèmes que tout analyste doit être en mesure de comprendre (sans pour autant avoir à les résoudre) pour faire une utilisation éclairée des logiciels de lexicométrie.

Dans leur modestie, les questions que nous posons entendent aussi se faire respecter : les considérer comme des subtilités byzantines et se contenter de les ignorer ne suffit pas à les périmer. On sait combien

il est aisé de se cacher derrière un argument « pragmatique » pour perpétuer une pratique établie. Mais le « bon sens » a des limites et il faut admettre qu'il se mue parfois en illusion collective : l'absurdité de certains travaux, pourtant sanctionnés par la reconnaissance des pairs et des institutions, en témoigne[23]. Si nous sommes ici si tranché, c'est que nous pensons que le domaine de l'analyse de discours possède un intérêt public : ses analyses éclairent nos motivations culturelles souvent inconscientes et participent à nous en émanciper. De ce fait, nous nous inquiétons de voir son paysage investi par des approches outillées contestables car mal maîtrisées. Dans un contexte où la recherche (et donc le contenu des enseignements) devient une compétition pour l'audience comme pour les financements, il serait fâcheux que le puissant outil d'*Aufklärung* que constitue l'herméneutique critique, outillée ou non, soit éclipsé par des méthodes qui perpétuent le sens commun objectiviste, quantitativiste et technophile alors même qu'elles sont censées le révéler.

Références

Adam J.-M., 2005. *La linguistique textuelle - Introduction à l'analyse textuelle des discours*. Paris : Armand Colin.

Bigot J.-E., Julliard V., Mabi C., 2016. « Humanités numériques et analyse des controverses au regard des SIC », *Revue française des sciences de l'information et de la communication*, n°8 [En ligne].

Benveniste E., 1966. *Problèmes de linguistique générale*. Paris : Gallimard.

Bernet C., 2009. « La distance intertextuelle et le théâtre du Grand Siècle », *Mélanges offerts à Charles Muller pour son centième anniversaire*. CILF : Paris, pp. 87-97 [En ligne].

Bourion E., 1998. « Ponctuation et accès sémantique aux banques textuelles », *Actes du colloque « à qui appartient la ponctuation ? »*. Paris, Bruxelles : Duculot, pp. 409-435 [En ligne].

Brett M., 2012. « Topic Modeling: A Basic Introduction », *Journal of Digital Humanities*, n°2 (1) [En ligne].

Brunet E, 2000. « Qui lemmatise dilemme attise », *Lexicometrica*, n°2 [En ligne].

Brunet, E., 2003. « Lexicométrie et étude du vocabulaire », *Hubert de Phalèze. A la recherche des Illusions perdues*. Paris : Nizet, pp. 29-47 [En ligne].

Brunet E., 2009. « Muller le lexicomaître », *Mélanges offerts à Charles Muller pour son centième anniversaire*. CILF : Paris, pp. 99-119 [En ligne].

Carbou G., 2017. « Quelques questions pour l'analyse statistique des données textuelles à l'ère des humanités numériques », *Les cahiers du numérique*.

Charolles M., 1995. « Cohésion, cohérence et pertinence du discours », *Travaux de Linguistique*, n°29, pp.125-151.

Chatauraynaud F., Debaz J., 2010. Prodiges et vertiges de la lexicométrie, http://socioargu.hypotheses.org/1963.

Dalud-Vincent M., 2011. « Alceste comme outil de traitement d'entretiens semi-directifs : essai et critiques pour un usage en sociologie », *Langage et Société*, n° 135, pp. 9-28.

[23] Citons par exemple l'engouement inquiétant pour les « *culturomics* », association évocatrice de « *culture* » et « *genomics* », qui envisagent une « *quantitative analysis of culture using millions of digitized books* » (cf. l'article publié dans le numéro 331 de Science en 2010 ; pour une présentation et une discussion voir Chatauraynaud et Debaz, 2010). Autres cas intéressants : les diverses techniques « hédonométriques » (http://hedonometer.org/about.html) ou encore les « *cultural analytics* » de tous bords qui proposent, par exemple, de mesurer la « production culturelle » des pays à partir de l'analyse du nombre de pages Wikipédia de leurs « personnage historiques » (http://pantheon.media.mit.edu/methods).

Scotto d'Apollonia L., Luxardo G., Piet G., 2014. « Approche lexicométrique des controverses climatiques », *JADT 2014 : 12e Journées internationales d'Analyse statistique des Données Textuelles*, pp. 606-616 [En ligne].

Guilhaumou J., 2006. « Science du texte et analyse de discours. Enjeux d'une interdisciplinarité », *Langage et Société*, n°116, pp. 149-151.

Jenny J., 1999. « Pour engager un débat avec Max Reinert, à propos des fondements théoriques et des présupposés des logiciels d'analyse textuelle », *Langage et Société*, n°90, pp. 73-85.

Jeanneret Y., 2013. « Les chimères cartographiques sur l'internet, panoplie représentationnelle de la traçabilité sociale », Galinon-Mélénec B., Zlitni S. (dir.), *Traces numériques : de la production à l'interprétation*, Paris : CNRS éditions, pp. 250-261.

Lebart L., 2012. « L'articulation entre exploration et inférence en analyse statistique de textes », *JADT 2012 : 11èmes Journées Internationales d'Analyse Statistiques des Données Textuelles*, pp. 708-715 [En ligne].

Micheli R., 2014. *Les émotions dans les discours. Modèle d'analyse, perspectives empiriques*. Bruxelles : De Boeck.

Maingueneau D., 2011. « Pertinence de la notion de formation discursive en analyse de discours », *Langage et Société*, n° 135, pp. 87-99.

Marchand P., Ratinaud P., 2012. « L'analyse de similitude appliquée aux corpus textuels : les primaires socialistes pour l'élection présidentielle française (septembre-octobre 2011) », *JADT 2012 : 11èmes Journées Internationales d'Analyse Statistique des Données Textuelles*, pp. 687-699 [En ligne].

Mayaffre D., 2008. « Quand "travail", "famille", "patrie" co-occurrent dans le discours de Nicolas Sarkozy. Étude de cas et réflexion théorique sur la co-occurrence », *JADT 2008 : 9èmes Journées Internationales d'Analyse Statistiques des Données Textuelles*, pp. 811-822 [En ligne].

Mayaffre D., 2013. « Sarkozysme et populisme. Approche logométrique du discours de Nicolas Sarkozy (2007-2012) », *Mots. Les langages du politique*, n°103, pp 73-87 [En ligne].

Mayaffre D., 2014. « Plaidoyer en faveur de l'Analyse de Données co(n)Textuelles. Parcours cooccurrentiels dans le discours présidentiel français (1958-2014) », *JADT 2014 : 12èmes Journées Internationales d'Analyse Statistique des Données Textuelles*, pp 15-32 [En ligne].

Missire R., 2004. « Une larme baudelairienne : essai de description morphosémantique de "Tristesses de la lune" », *Texto !*, [En ligne].

Missire R., 2013. « Perception sémantique et perception sémiotique », *Texto !*, Vol. XVIII, n°2 [En ligne].

Missire R. 2015. « Sémantique des thèmes : de la collocation lexicale à la corrélation sémantique », *Communication orale au séminaire du LILPA ER - Fonctionnements Discursifs et Traduction*.

Nazarenko A., 1998. « Présentation du numéro », *Traitement automatique des langues*, n°39, pp. 3-7.

Née É., 2012. *L'insécurité en campagne électorale*. Paris : Honoré Champion.

Petit G., 1999. « Présentation : le statut de l'unité lexicale », *Linx*, n°40, pp 7-10 [En ligne].

Pincemin B., 1999. « Sémantique interprétative et analyse automatique de textes : que deviennent les sèmes ? », *Sémiotiques*, n°17, pp. 71-120.

Pincemin B., 2012. « Sémantique interprétative et textométrie », *Texto !*, Vol. XVII, n°3 [En ligne].

Rastier F., 1989. *Sens et textualité*. Paris : Hachette.

Rastier F. 1996. « La sémantique des thèmes - ou le voyage sentimental », *Texto !* [En ligne].

Rastier F., 2002. « Sur l'immanentisme en sémantique », *Texto !* [En ligne].

Rastier F., 2003a. « Le langage comme milieu : des pratiques aux œuvres », *Texto !* [En ligne].

Rastier F., 2003b. « Le silence de Saussure ou l'ontologie refusée », Bouquet S. (éd.), *Saussure*, Paris : L'Herne, pp. 23-51.

Rastier F., 2007. « Passages », *Corpus*, n°6, pp. 25-54.

Reinert M., 1990. « ALCESTE, une méthodologie d'analyse des données textuelles et une application : Aurélia de Gérard de Nerval », *Bulletin de méthodologie sociologique*, n°26, pp. 24-54.

Reinert, M., 1993. « Les "mondes lexicaux" et leur "logique" à travers l'analyse statistique d'un corpus de récits de cauchemars », *Langage et Société*, n°66, pp. 5-39.

Reinert, M., 2007. « Postures énonciatives et mondes lexicaux stabilisés en analyse statistique de discours », *Langage et Société*, Vol. 121-122, n° 3, pp. 189-202.

Salton G., Allan J., Buckley C., 1994. « Automatic structuring and retrieval of large textfiles », *Communications of the ACM*, n°37 (2), pp. 97-108.

Saussure F., 2002. *Écrits de linguistique générale*, Texte établi et édité par Simon Bouquet et Rudolf Engler, Paris : Gallimard.

Tournier M., 1985. « Dans l'ombre portée des sigles confédéraux : un mirage lexicométrique (CGT et CFDT 1972) », *Colloque de Nice - Méthodes quantitatives et informatiques dans l'étude des textes*, pp. 842-853.

Valette M., Eensoo E., 2014. « Approche textuelle pour le traitement automatique du discours évaluatif », *Langue française*, n° 184, pp. 109-124.

Eensoo E., Valette M., 2014. « Sémantique textuelle et TAL : un exemple d'application à l'analyse des sentiments », Ablali D., Badir S., Ducard D. (éd.), *Documents, textes, œuvres. Perspectives sémiotiques*, Rennes : Presses Universitaires de Rennes, pp. 75-89

Valette M., 2016. « Analyse statistique des données textuelles et traitement automatique des langues. Une étude comparée », *JADT 2016 : 13èmes Journées d'Analyse Statistique des Données Textuelles*, pp. 697-706 [En ligne].