Sémantique interprétative et textométrie– Version abrégée

Bénédicte Pincemin

Édition électronique

URL : <http://journals.openedition.org/corpus/2121>

DOI : 10.4000/corpus.2121

ISSN : 1765-3126

Éditeur

Bases ; corpus et langage - UMR 6039

Édition imprimée

Date de publication : 1 novembre 2011

Pagination : 259-269

ISSN : 1638-9808

Référence électronique

Bénédicte Pincemin, « Sémantique interprétative et textométrie– Version abrégée », *Corpus* [En ligne], 10 | 2011, mis en ligne le 15 juin 2012, consulté le 10 décembre 2020. URL : <http://journals.openedition.org/corpus/2121> ; DOI : <https://doi.org/10.4000/corpus.2121>

Sémantique interprétative et textométrie – Version abrégée¹

Bénédicte PINCEMIN

CNRS « Interactions, Corpus, Apprentissages,
Représentations » (Lyon)²

1. Introduction

Cet article fait le point d'une réflexion, de discussions et *d'expériences* sur la pertinence de l'approche logicielle textométrique au regard des principes de la sémantique interprétative.

La *textométrie* s'appelle aussi logométrie ou statistique textuelle, c'est la forme actuelle de la lexicométrie (Lebart & Salem 1994). Elle propose des procédures de tris et de calculs statistiques pour l'étude d'un corpus de textes numérisés. A ces procédures quantitatives la textométrie articule fortement des moyens de parcours et d'interprétation qualitatifs, déterminants quant aux affinités possibles avec une théorie linguistique telle que la sémantique interprétative.

La textométrie ne se confond pas avec la linguistique de corpus. Toutes deux fondent leurs investigations sur un corpus numérique, dont la constitution est déterminante. Comme son nom l'indique, la linguistique de corpus poursuit un objectif de description et de modélisation de la langue. La textométrie, centrée sur le texte, a pu être mobilisée par diverses sciences humaines (histoire, littérature, sciences politiques...). Développée au sein d'une communauté scientifique s'intéressant à

---

¹ Pincemin B. (sous presse). « Sémantique interprétative et textométrie », *Texto!* [<http://www.revue-texto.net/> ISSN 1773-0120]. Version en espagnol : Pincemin B. (2010). « Semántica interpretativa y textometría », *Tópicos del Seminario* 23 : 15-55.

² Laboratoire mixte CNRS / Université de Lyon, UMR 5191.

l'analyse des données textuelles³, elle se caractérise notamment par certains calculs fondateurs, statistiques (les spécificités, les cooccurrences) ou non (les segments répétés, les concordances), et accorde une place fondamentale au « retour au texte » (bien outillé dans les logiciels) pour interpréter les unités (généralement des mots) sélectionnées par les calculs. Une étude qui utilise une approche et des outils textométriques, et qui vise à observer et décrire des phénomènes linguistiques en corpus, peut donc relever à la fois de la linguistique de corpus et de la textométrie, sans pour autant qu'aucun de ces deux courants ne subsume l'autre.

La *sémantique interprétative* renvoie à la théorie linguistique développée par François Rastier (1987, 2001). Cet article nous amènera tout naturellement à en rappeler les principes fondamentaux, pour les confronter aux caractéristiques de la textométrie.

Notre démarche est en quatre temps : (i) invalider les *a priori* négatifs sur la textométrie quant à sa compatibilité avec une approche linguistique, (ii) observer et analyser le développement des pratiques textométriques en sémantique interprétative, (iii) expliquer théoriquement les affinités de la théorie sémantique et de la textométrie par la mise en évidence de liens au niveau de leurs principes fondamentaux, (iv) en tirer des perspectives de recherche. Pour les besoins de cette version abrégée, nous avons choisi de privilégier ici la partie la plus fondamentale, à savoir la mise en évidence des affinités théoriques profondes entre textométrie et sémantique interprétative. Les autres parties sont maintenues mais leur contenu est simplement évoqué sans développement.

---

3 Les actes des Journées internationales d'Analyse statistique des Données Textuelles (JADT) publient de nombreuses communications sur des questions de théorie textométrique et sur des exemples d'application variés, ainsi que des communications relevant d'autres approches d'analyse textuelle outillée. Une édition en ligne de ces actes est accessible depuis le site *Lexicometrica* (<http://www.cavi.univ-paris3.fr/lexicometrica/>), plus spécialement dédié à la textométrie.

2. Discussion de réductions à première vue compromettantes

[On répond ici aux limites communément reprochées à l'approche textométrique quand elle est mal connue : la modélisation du texte s'apparentant à un sac de mots, la conception plus informatique que linguistique des mots en chaînes de caractères, la négligence des basses fréquences alors qu'il peut s'agir de singularités hautement significatives, et plus généralement l'inadéquation d'un traitement quantitatif à un objet et une problématique de nature qualitative.]

3. Des expériences positives révélatrices

[On rappelle une expérience initiale révélatrice : Rastier voit le calcul de l'écart-réduit mettre en évidence, à même le texte de la nouvelle *Toine* de Maupassant, des formes clés pour son interprétation (Rastier 2001 : 96, note 1). Une méthodologie fructueuse est alors développée : un calcul de cooccurrence tel que celui implémenté par la fonction *Thème* du logiciel Hyperbase sert à repérer des corrélats sémantiques pour une description thématique des isotopies et des molécules sémiques (Rastier 2001 : 213). Les apports de la textométrie sont également explorés pour le repérage de passages et la caractérisation des textes et des genres textuels.]

4. Connivences de fond

4.1 Linguistique et sémantique

La sémantique interprétative s'intéresse au sens. Selon elle, le sens peut s'élaborer à partir d'indices morphologiques, syntaxiques, phonétiques, etc. Il peut être transversal aux catégories grammaticales, mobiliser au même titre une ponctuation, un aspect verbal, un morphème, un profil prosodique et rythmique, une typographie, une mise en page (Bourion 2001, Malrieu & Rastier 2001, Beaudouin 2002, Valette 2004, Loiseau 2006). La textométrie est en mesure de prendre en compte des descriptions du texte de toutes natures, pour peu qu'elles soient explicitées par un codage du corpus.

La démarche sémantique proposée par la textométrie est tout à fait en accord avec la demande de « dé-ontologie »

formulée par Rastier. Il s'agit bien d'éviter toute préconception réductrice, on veut surtout rester au plus proche du texte et ne pas commencer par l'étudier à travers le prisme d'une ontologie. Ainsi, dès ses débuts, la textométrie s'est fait une spécialité du traitement des questions ouvertes dans les enquêtes, pour éviter le postcodage des réponses (entre l'enquête et l'analyse) qui efface des variations d'expressions potentiellement significatives (Lebart & Salem 1994).

Le souci de fidélité au texte s'est également vivement exprimé dans le débat traversant la communauté textométrique et concernant la lemmatisation (Brunet 2011) : l'expérience avait révélé que bien souvent les différentes flexions sont porteuses d'une sémantique différente, typiquement les pluriels sont plus concrets que les singuliers (ex. « le travail » vs « les travaux ») (Geoffroy, Lafon & Tournier 1974). La sémantique interprétative partage cette sensibilité à ne pas écraser ces distinctions, comme le montre l'étude de Bourion (2001 : 62) sur « au pied de » vs « aux pieds de » dans un corpus littéraire.

4.2 Sémantique des textes

4.2.1 La place centrale des textes à toutes les étapes de l'analyse textométrique

Dans les termes de la sémantique interprétative, l'objet empirique de la linguistique n'est pas d'abord le mot, la phrase ou la proposition, mais le texte.

L'évolution de désignation de la « lexicométrie » en « textométrie » veut exprimer que l'analyse menée ne se cantonne pas à l'étude du lexique, mais investit l'ensemble du texte. La textométrie compte ainsi des utilisateurs non seulement en linguistique mais aussi dans diverses sciences humaines (littérature, politique, histoire, philologie etc.), dont l'objectif est de se donner des moyens de renouveler la lecture de leur corpus de documents textuels.

Le corpus, comme ensemble construit de textes, est au cœur de l'analyse ; le retour au texte est une fonctionnalité centrale des logiciels de textométrie. L'approche textométrique est celle de la curiosité d'une lecture approfondie et renouvelée par la mise en évidence de régularités non encore perçues. On

est loin ici d'un corpus « réservoir », dont la connaissance globale importe peu, et dont le rôle n'est que de pouvoir livrer quelques extraits ciblés, souvent sans considération particulière pour la formulation employée et sa contextualisation complète.

4.2.2 La contextualisation comme principe d'analyse et le rôle déterminant du corpus de référence

La textométrie compte, situe, caractérise, des unités dans et par des contextes. Or justement « la compréhension du texte [...] procède par contextualisation et intertextualisation » (Rastier 2001 : 93). Cela se déploie à tous les paliers, se reformulant en autant de principes (Rastier 2001 : 92) : le principe de *contextualité* (« deux signes ou deux passages d'un même texte mis côte à côte sélectionnent réciproquement des éléments de signification (sèmes) [...] »...), le principe d'*intertextualité* (analogue pour deux passages de textes différents), et le principe d'*architextualité*, selon lequel tout texte plongé dans un corpus en reçoit des déterminations sémantiques, et modifie potentiellement le sens de chacun des textes qui le composent.

En outillant une lecture non linéaire (par l'extraction de contextes, la génération de concordances), la textométrie joue fondamentalement sur les principes de contextualité et d'intertextualité. Les cooccurrences ont pu être comprises également comme une forme de contextualisation minimale, dans l'esprit de la sémantique interprétative (Mayaffre 2008). Quant au principe d'architextualité, le fonctionnement statistique du corpus de référence en est une concrétisation : en effet, tout texte inclus dans le corpus apporte sa contribution aux fréquences globales, et réciproquement, se trouve caractérisé par rapport à ces fréquences globales. Le choix du corpus de référence conditionne l'interprétation, et un texte reçoit autant d'éclairages possibles que de contextualisations pertinentes en corpus. Par le biais du corpus de référence, le global détermine le local, et l'analyse est la mise en évidence de formes qui se détachent sur un fond (Rastier 2001 : 42 *sq.*).

4.3 Sémantique interprétative

La sémantique interprétative propose une conception dynamique du sens, une praxéologie, relative à des pratiques contextualisées.

Le sens se construit au fil de la lecture, qui s'apparente à une reconnaissance de formes, peu à peu cernées, et même évolutives. Il s'agit d'une activité perceptive, le geste interprétatif s'ajuste en fonction des contraintes linguistiques reçues du texte.

L'analyse textométrique procède également d'une démarche construite : on ne peut pas fournir un corpus, « faire tourner » le logiciel, et récupérer le résultat comme produit fini. Chaque étape suppose des choix et implique l'utilisateur, et bien souvent la dynamique de l'interprétation procède par ajustement progressif des données et des calculs : on retrouve très concrètement l'interprétation comme action et comme geste qui affine dynamiquement sa trajectoire.

On peut reconnaître aussi dans certains processus textométriques des principes herméneutiques traditionnels, comme celui des « passages parallèles » (Bourion 2001 : 116, Pincemin 2006).

4.4 Sémantique différentielle

La sémantique interprétative se définit comme une sémantique différentielle, par opposition à des sémantiques inférentielle ou référentielle.

4.4.1 Le fonctionnement différentiel des représentations et calculs textométriques

Formellement, au plan des décomptes de fréquences et autres calculs qui s'ensuivent, la textométrie suppose de convenir d'un typage des unités, qui fixe ce qui est reprise du même (et se cumule en termes de fréquences) et ce qui est différent (et participe donc au nombre de types). On a donc déjà fondamentalement une modélisation qui, à un moment donné, choisit d'assimiler certaines unités et d'en dissimiler d'autres.

Ensuite, les statistiques contrastives proposées en textométrie ont bien été comprises comme un mécanisme différentiel : car ces statistiques mettent en évidence ce qui s'écarte de la norme (définie par le corpus de référence), elles repèrent les contrastes dans un tout qui fait système (Bourion 2001 : 18, Rastier 2001 : 86). Et même un simple tri alphabétique dans une liste ou une concordance fait entrer en jeu heuristiquement une perception différentielle : l'œil repère les motifs récurrents,

rapprochés par le tri, et les variations à l'intérieur ou aux frontières de ces regroupements (Pincemin 2006).

4.4.2 L'attention à la structure et la dominance du qualitatif

La sémantique interprétative relève d'une approche structurale de la langue. Elle recherche les contraintes posées par la langue et par son usage dans le texte considéré pour hiérarchiser des lectures possibles.

De même, les mesures textométriques (fréquences, scores statistiques) sont utilisées pour le classement qu'elles induisent, dit « tri hiérarchique ». La pratique textométrique consiste alors à travailler sur des zones de la liste triée (notamment les extrêmes), et c'est un examen qualitatif (avec retour au texte, etc.), qui détermine l'interprétation.

4.5 Sémantique unifiée

La sémantique interprétative est unifiée, au sens où les principes différentiels et interprétatifs s'observent à différents paliers, typiquement ceux de la lexie, de la période, et du texte. La textométrie s'accommode très naturellement de la multiplicité des paliers et de leurs analogies de fonctionnement. De fait, formellement, la technique n'exige que de lui fournir des unités ou « contenus », réparties dans des « contenant ». Si l'option la plus courante consiste à étudier le corpus à travers son lexique, on peut tout aussi bien le voir par le biais d'autres descriptions (ex. catégories grammaticales) et d'autres paliers (le morphème ou son approximation par des tri-grams, le tour de parole,...). De même, la variation du choix de la partition (découpage du corpus en parties) est un moyen de rendre compte de catégories méta-textuelles et de caractérisations philologiques (comme le genre, l'auteur, la période historique), de variables situationnelles significatives de la production des textes, comme de structurations intratextuelles (ex. chapitres).

Les unités linguistiques finalement construites au fil de l'analyse peuvent relever d'un autre palier que les unités ayant servi au calcul. Un calcul basé sur des mots peut ainsi conduire à repérer des sèmes (infralexicaux, au sens où un mot est *a priori* porteur de plusieurs sèmes) et à construire des molécules sémiques représentant un thème (supra-lexical, au sens où sa

manifestation peut être diffuse sur tout un passage voire tout un texte, et où il se prête à des lexicalisations multiples) (Rastier 2001 : 206). Les techniques statistiques permettent également de capter des régularités significatives bien qu'imperceptibles pour une lecture non outillée. La textométrie ferait donc partie des techniques capables de plonger dans des dimensions profondes du matériau textuel (Rastier 2005).

5. Perspectives d'apports mutuels

[La textométrie suscite actuellement de nouvelles propositions et de nouveaux développements, tout particulièrement autour de la réalisation collaborative d'une plateforme logicielle ouverte, fédérant les recherches et les développements informatiques des principales équipes du domaine⁴. L'approche sémantique interprétative vient alimenter les réflexions sur (i) la modélisation des textes et des corpus (notamment la pluralité et la dynamique des vues rendues possibles par les corpus structurés et enrichis), (ii) la conception de fonctionnalités (compréhension linguistique des fonctionnalités existantes, proposition de nouvelles fonctionnalités, avec en particulier l'annotation dynamique des corpus), et (iii) l'élaboration de repères méthodologiques (pouvant intégrer d'autres aspects de la sémantique interprétative).]

6. Conclusion : enjeux pratiques et théoriques

D'autres techniques, d'autres calculs, pourront certainement montrer leur pertinence pour l'approche de la sémantique interprétative : si l'argumentaire de cet article est enthousiaste, il ne veut néanmoins prétendre ni à l'exclusivité de la textométrie comme proposition de réponse logicielle à la théorie rastiérienne, ni au caractère idéal de la textométrie actuelle, qui est de fait encore en pleine évolution. Ceci étant, il reste frappant de trouver tant de connivences de fond entre les deux approches : texte, contextualisation, intertextualité et corpus, sémantique différentielle, interprétation et dynamique de la construction d'un sens. L'expérimentation des nouvelles formes de lecture et

---

4 Projet *Textométrie* et plateforme logicielle *TXM* (<http://textometrie.ens-lyon.fr/>).

d'interprétation instrumentées logiciellement n'est pas une simple application, qui plus est inévitablement réductrice, de la réflexion théorique sémantique : elle est en mesure de la relancer et de la renouveler en lui dévoilant des réalités textuelles ou herméneutiques pas encore perçues ou oubliées.

Ce texte participe à la réflexion menée dans le projet Textométrie ANR-06-CORP-029.

Je remercie beaucoup Evelyne Bourion, Carine Duteil-Mougel, Serge Heiden, Sylvain Loiseau, Damon Mayaffre, Céline Poudat, et Mathieu Valette, pour leurs relectures attentives et constructives, et pour les nuances importantes et les précisions qu'ils m'ont permis d'apporter à la version source de l'article.

Bibliographie

- Beaudouin V. (2002). *Mètre et rythme du vers classique – Corneille et Racine*. Paris : Champion.
- Bourion E. (2001). *L'aide à l'interprétation des textes électroniques*. Thèse de doctorat en Sciences du langage, Université de Nancy II, 14 décembre 2001.
- Brunet E. (2011). *Ce qui compte. Ecrits choisis, tome II, Méthodes statistiques*. C. Poudat (éd.), Paris : Champion.
- Geoffroy A., Lafon P. & Tournier M. (1974). « L'indexation minimale – Plaidoyer pour une non-lemmatisation », ENS de Saint-Cloud, 30 pages – Communication au Colloque sur l'Analyse des corpus linguistiques : Problèmes et méthodes de l'indexation maximale, Strasbourg, 21-23 mai 1973.
- Lebart L. & Salem A. (1994). *Statistique textuelle*. Paris : Dunod.
- Loiseau S. (2006). *Sémantique du discours philosophique : du corpus aux normes. Autour de G. Deleuze et des années 60*. Thèse de doctorat en Sciences du langage, Université de Paris X Nanterre, 1^{er} décembre 2006.
- Loiseau S. (2007). « CorpusReader : un dispositif de codage pour articuler une pluralité d'interprétations », *Corpus* 6 : 153-186.

- Malrieu D. & Rastier F. (2001). « Genres et variations morphosyntaxiques », *Traitement automatique des langues* 42, 2 : 547-577.
- Mayaffre D. (2005). « De la lexicométrie à la logométrie », *L'Astrolabe. Recherche littéraire et informatique*. Revue en ligne, article consultable à l'adresse : <http://www.uottawa.ca/academic/arts/astrolabe/articles/art0048/Logometrie.htm>.
- Mayaffre D. (2008). « De l'occurrence à l'isotopie. Les co-occurrences en lexicométrie », *Syntaxe & Sémantique* 9 : 53-72.
- Pincemin B. (2006). « Concordances et concordanciers – De l'art du bon KWAC », in C. Duteil-Mougel & B. Foulquié (éd.) *Corpus en Lettres et Sciences sociales : des documents numériques à l'interprétation, Actes du XVII^e Colloque d'Albi Langages et Signification*, Albi, 10-14 juillet 2006, 33-42. Et *Texto!* XI, 2.
- Poudat C. (2006). *Etude contrastive de l'article scientifique de revue linguistique dans une perspective d'analyse des genres*. Thèse de doctorat en Sciences du langage, Université d'Orléans, 20 juin 2006.
- Prié Y. (1995). *Contribution à une clarification des rapports entre Sémantique Interprétative et Informatique*. Mémoire de DEA en Informatique, Université de Rennes 1.
- Rastier F. (1987). *Sémantique interprétative*. Paris : Presses Universitaires de France.
- Rastier F. (2001). *Arts et sciences du texte*. Paris : Presses Universitaires de France.
- Rastier F. (2005). « Enjeux épistémologiques de la linguistique de corpus », in G. Williams (éd.) *La linguistique de corpus*. Rennes : Presses Universitaires de Rennes, 31-46.
- Rastier F. (2007). « Passages », *Corpus* 6 : 25-54.
- Rastier F. (2008). « Que cachent les “données textuelles” ? », in S. Heiden & B. Pincemin (éd.) *Actes des 9^{es} Journées internationales d'Analyse statistique des Données*

- Textuelles (JADT 2008)*, Lyon, 12-14 mars 2008. Lyon : Presses Universitaires de Lyon, vol. I, 13-26.
- Valceschini-Deza N. (1999). *Accès sémantique aux bases de données textuelles*. Thèse de doctorat en Linguistique, Université de Nancy 2, 29 juin 1999.
- Valette M. (2004). « Sémantique interprétative appliquée à la détection automatique de documents racistes et xénophobes sur Internet », in P. Enjalbert & M. Gaio (éd.) *Approches Sémantiques du Document Numérique. Actes du 7^e Colloque International sur le Document Electronique*, 22-25 juin 2004, 215-230.