Sémantique textométrique

Bénédicte Pincemin

Sémantique textométrique

1. Introduction

La textométrie n'est pas en soi une théorie sémantique, mais plutôt une méthodologie pour l'analyse de corpus de textes[1]. Cela étant, elle apporte un éclairage expérimental important sur le fonctionnement de la langue et des textes, en particulier sur ses dimensions sémantiques. La *sémantique textométrique* s'entend donc ici comme une approche du sens linguistique à travers une certaine forme d'exploration des corpus. Cette approche exploratoire du sens, apportant des éclairages concrets mais partiels, est potentiellement compatible avec diverses modélisations sémantiques visant une représentation intégrée, intelligible et complète du fonctionnement linguistique.

Nous avons conçu ce chapitre en trois temps. Tout d'abord, il s'agit de préciser le champ de la textométrie, et d'expliquer son positionnement spécifique par rapport à d'autres approches d'analyse de textes relevant des Humanités numériques telles que la linguistique de corpus et le *text mining*. C'est en effet en prenant appui sur les choix propres à la textométrie que nous pourrons développer le propos sur la conception sémantique opérationnalisée par cette approche. Ainsi, nous recenserons une demi-douzaine de principes opératoires de la textométrie, comme autant de caractéristiques qui sont postulées sur la langue, qui servent de points d'appui aux procédures d'analyse, et qui heuristiquement en retour, par l'intérêt des observations permises, voient confirmée leur justesse. Avec la même démarche type « retour d'expérience », la troisième partie de notre propos aborde la question des unités linguistiques pour l'analyse sémantique. En pratique, les unités de l'analyse textométrique sont des découpages automatisés qui s'éloignent plus ou moins d'un idéal linguistique ; pourtant, cet écart entre les unités de départ de l'analyse et les unités de description pertinentes pour le linguiste peut être assumé, les unités sémantiques étant non pas données initialement mais construites par un parcours interprétatif au sein d'un ensemble global de textes contextualisant.

2. Présentation de la textométrie

2.1. Histoire : quelques jalons décennaux, de la lexicométrie à la textométrie

De toute évidence, les approches quantitatives de la langue existent de très longue date. Mais l'invention et la diffusion de l'ordinateur dans les dernières décennies du XX^{e} siècle ouvrent de nouveaux terrains de recherche en rendant possible la mise en œuvre de calculs complexes sur des corpus représentatifs de contextes intertextuels larges plutôt que sur des textes isolés.

Précurseurs immédiats, dans les années 1960, les travaux de statistique lexicale de Pierre Guiraud (1960) et de Charles Muller (1973, 1977) préparent la voie à la textométrie, en réfléchissant à l'adéquation de certains calculs statistiques avec certaines questions linguistiques. Muller (1977) souligne en particulier l'importance d'une norme de définition des unités d'analyse, la régularité de ces unités étant une condition fondamentale pour la cohérence des décomptes quantitatifs.

C'est au début des années 1970 que Jean-Paul Benzécri et son équipe développent une branche des mathématiques dénommée *analyse des données,* avec notamment le calcul d'analyse des correspondances et son articulation aux procédures de classification automatique (1973), calcul dont l'application aux données textuelles (1981) sera un des fondements de la lexicométrie.

1 Il s'agit ici de texte au sens large : écrit plus ou moins rédigé mais aussi oral transcrit.

La plupart des autres calculs fondamentaux de l'approche lexicométrique, tels que spécificités (Lafon, 1980), cooccurrences (Lafon, 1981), segments répétés (Salem, 1987), analyse arborée (Luong, 1988), méthode ALCESTE (Reinert, 1990), sont mis au point dans les années 1980. La revue *MOTS – Mots Ordinateurs Textes Sociétés* – est créée, qui accueille et rassemble les idées fondatrices et les résultats de ces explorations informatisées.

Les années 1990 correspondent à une phase d'expansion et de diffusion. Les programmes informatiques développés et exploités au sein des laboratoires de recherche se transforment en logiciels pour ordinateur personnel, expérimentables par une communauté d'utilisateurs élargie, tels qu'Hyperbase (Étienne Brunet, à Nice), Lexico (André Salem, à Saint-Cloud puis Paris 3), Lexploreur (Serge Heiden, à Saint-Cloud), Alceste (Max Reinert, à Toulouse). Un colloque est créé, qui rassemble la communauté de recherche en lexicométrie : les Journées internationales d'Analyse statistique des Données Textuelles (JADT), avec une édition tous les deux ans depuis 1991. L'ouvrage de référence de Ludovic Lebart et André Salem (1988), rapidement épuisé, fait l'objet d'une nouvelle édition enrichie en 1994, et une version anglophone est publiée en 1998 (Lebart *et al.*, 1998).

Dans les années 2000, les techniques de codage de corpus et le traitement automatique des langues (TAL) évoluent fortement. Le texte « brut » typique des cartes perforées puis du Minitel est devenu le texte mis en forme et balisé, à commencer par les pages Internet ; la norme SGML initiale, qui pose les fondements théoriques et pratiques des langages de balisage, a donné naissance au format HTML utilisé pour la présentation des pages Web et la navigation hypertexte, et évolue vers le standard XML pour l'enregistrement structuré de données de toutes natures (textes, mesures, catégories, etc.)[2]. Parallèlement, les outils d'analyse morphosyntaxique automatique acquièrent une certaine maturité, et rendent possible l'enrichissement à la volée des textes par des informations de catégorie grammaticale et d'entrée lexicale (lemme[3]) pour chaque mot. L'analyse peut alors davantage porter sur des structures textuelles (titres, paragraphes, vers, etc.) et des informations linguistiques (morphologie, syntaxe, sémantique, etc.), ce qui motive une redénomination de l'approche : de la *lexicométrie*, semblant dévolue aux questions lexicales, on passe à la *textométrie* ou *logométrie*, déployant les traitements à tous les niveaux de composition linguistique du texte.

2.2. Contenu : principaux types de traitement et d'analyse

La textométrie associe certains calculs statistiques, choisis pour leur adéquation à des questionnements sur des données textuelles, à une démarche de *retour au texte* systématique pour lire les mots en contexte, démarche elle aussi dotée de procédures adaptées, comme l'affichage en concordance. C'est ainsi que la textométrie se présente comme une analyse à la fois quantitative et qualitative.

Nous introduisons succinctement ci-après les principaux traitements textométriques, typiques de l'approche, en les illustrant par des résultats[4] de leur application au corpus VOEUX[5] rassemblant les discours de vœux des présidents français de la V^e^ République pour la période du 1^er^ janvier 1960 au 1^er^ janvier 2013.

2 SGML : *Standard Generalized Markup Language* ; HTML : *Hypertext Markup Language* ; XML : *Extensible Markup Language*. Il s'agit donc dans tous les cas de *Markup Language*, donc de langage de balisage, langage formel servant à repérer des structures dans les données à l'aide d'un système de notations conventionnelles, qui s'ajoutent aux données sans se confondre avec elles. Cela enrichit les données pour les traitements automatiques, en explicitant des informations sinon inaccessibles à ces traitements.

3 Le lemme est une forme normalisée permettant de reconnaître une même entrée lexicale sous différentes flexions. On reprend les conventions usuelles des dictionnaires en rapportant le mot à sa forme non fléchie : le lemme d'un verbe est son infinitif, celui d'un nom son singulier, celui d'un adjectif sa forme au masculin singulier, etc.

4 Les calculs et graphiques ont été produits à l'aide des logiciels *open-source* TXM (Heiden *et al.*, 2010) et IraMuTeQ (Ratinaud et Déjean, 2009).

5 Ce corpus a été mis à disposition par Jean-Marc Leblanc (Université Paris-Est Créteil) et nous en avons utilisé l'édition préparée et diffusée par le projet Textométrie (https://sourceforge.net/projects/txm/files/corpora/voeux).

Le calcul des *spécificités* (Lafon, 1980) mesure le caractère attendu ou exceptionnel de la fréquence d'un mot (ou motif complexe, trait linguistique, etc.) dans une partie du corpus, au regard de sa fréquence dans l'ensemble du corpus et de la taille de la partie. La manière de définir une partie est très libre : les textes d'un auteur, d'une période, d'un genre textuel ; les prises de parole d'un personnage, d'un certain type de locuteur (caractéristique sociodémographique) ; etc. Plus puissant qu'un calcul de fréquence relative (pourcentages), il évite de surévaluer les écarts relatifs importants mais qui ne portent que sur quelques occurrences[6], et il est capable également d'évaluer différemment les absences en pointant les plus étonnantes au regard du corpus. Il peut servir tant à établir le vocabulaire ou les traits linguistiques caractéristique(s) d'une partie (tableau 1) qu'à construire le profil de répartition d'un mot ou d'un trait sur le corpus : s'agit-il d'un mot très commun et partagé par les différentes composantes du corpus, ou bien est-il manifestement concentré dans certaines parties ? (figure 1). Autrement dit, c'est un calcul-clé pour faire le lien entre des traits globaux (textuels, intertextuels) et des traits locaux (typiquement lexicaux et infra-lexicaux), par effet contrastif.

Forme nominale	Fréquence dans les discours de De Gaulle	Fréquence en corpus	Indice de spécificité
coopération	12	14	6,3
développement	18	28	6,0
rapports	8	8	5,4
œuvre	11	15	4,6
nation	18	33	4,5
régime	7	8	3,9
ardeur	5	5	3,4
essor	5	5	3,4
détente	6	7	3,3
univers	6	7	3,3
peuples	20	46	3,2
progrès	21	51	3,0
dialogue	0	22	-2,3
courage	0	22	-2,3
emplois	0	23	-2,4
respect	0	24	-2,5
croissance	0	35	-3,7
chômage	0	36	-3,8
solidarité	0	41	-4,3
compatriotes	0	62	-6,5

6 Ainsi, en troisième ligne du tableau 1, le fait que toutes les occurrences du mot *rapports* se concentrent chez De Gaulle est statistiquement moins remarquable que certaines inégalités de répartition apparemment moins marquées pour des mots plus fréquents (*coopération*, *développement*) : autrement dit, il est « plus difficile » (moins probable) d'avoir « au hasard » chez De Gaulle 18 des 28 occurrences de *développement*, que 8 sur 8 occurrences de *rapports*. De même, statistiquement, plus le mot est rare, moins son apparition exclusive dans la partie est remarquable, parce que lorsqu'il n'y a que quelques occurrences il n'est pas si improbable qu'elles apparaissent toutes dans la même partie (l'indice de spécificité pour *ardeur* et *essor*, de fréquence 5, est moindre que celui de *rapports*, de fréquence 8). Le modèle probabiliste nuance et enrichit le jugement que l'on pourrait porter avec de simples mesures de proportion.

Tableau 1. Illustration des spécificités : noms sur-employés (indices positifs) ou sous-employés (indices négatifs) par De Gaulle.[7]

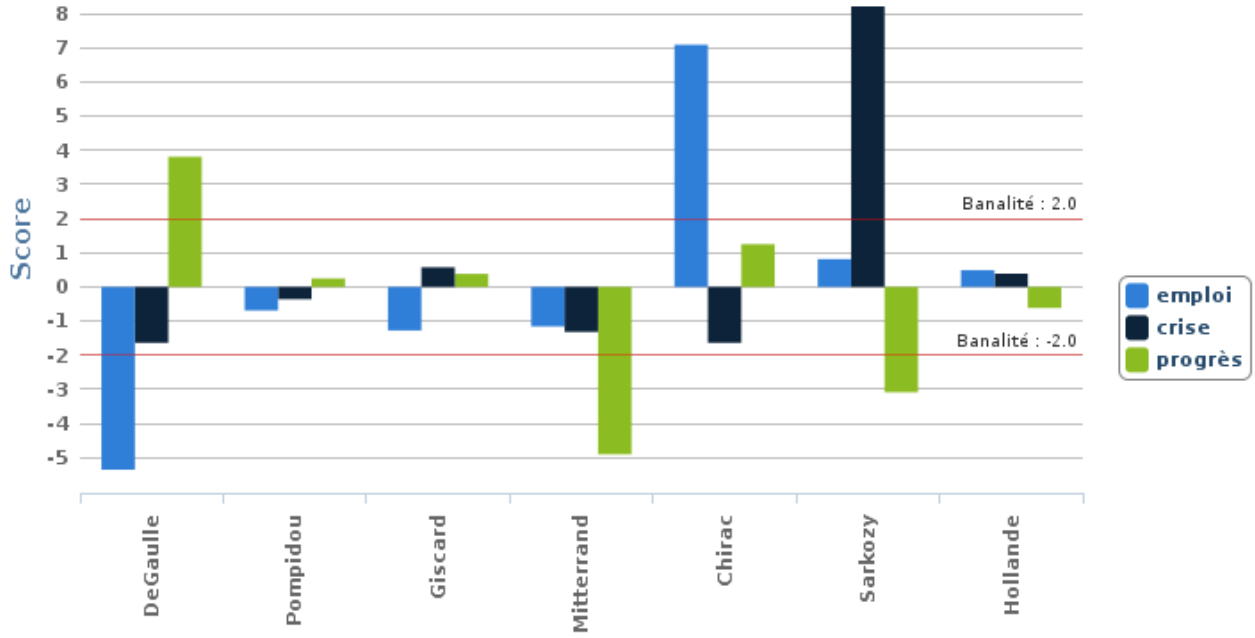

Figure 1. Illustration des spécificités : Profils de répartition de trois noms au fil des différents présidents

Le calcul de *cooccurrence* (Lafon, 1981) vise à repérer les attirances lexicales, de toutes natures, se distinguant plutôt par leur portée : du figement lexical au champ sémantique, en passant par la phraséologie et les contraintes syntaxiques (tableau 2). Techniquement, il est actuellement le plus souvent modélisé comme une variante du calcul de spécificités, la partie à caractériser étant l'ensemble des contextes[8] de l'unité lexicale considérée.

Cooccurrent d'*avenir*	**Fréq.**	**Cofrq.**	**Indice**	**Dist. Moy.**	**Cooccurrent de *crise(s)***	**Fréq.**	**Cofrq.**	**Indice**	**Dist. Moy.**
notre	407	32	7	2,3	financière	8	4	4	1,0
l'	1348	62	4	2,4	sortir	9	4	4	4,8
confiance	59	8	3	3,3	La	135	11	3	2,4
préparer	5	3	3	1,7	grave	7	3	3	3,3
avenir	79	9	3	3,0	décisions	9	3	3	8,3
enfants	35	6	3	3,5	vite	9	3	3	4,3
dépend	14	4	3	2,8	la	1762	55	2	3,1
Conscience	7	3	3	3,0	malgré	11	3	2	2,7
insuffisamment	2	2	3	7,5	sang-froid	3	2	2	9,0

7 Bien noter que ce jugement statistique est limité ici aux noms, et à leur usage dans les discours de vœux présidentiels. Noter également qu'en partie basse du tableau, les mots ont non seulement un indice négatif mais aussi une fréquence nulle (ils sont absents du discours de De Gaulle) : les mots qui ont cette double caractéristique sont appelés *nullax*, le calcul pointe ainsi les mots dont l'absence est la plus étonnante, au regard du reste du corpus,

8 Il s'agit de voisinages textuels, définis en « fenêtres » (proximité en nombre de mots) ou en unités physiques typographiques (la page) ou logiques (la phrase, le paragraphe, le tour de parole). Le terme approprié en linguistique serait ici plutôt *cotexte* que *contexte*, mais pour la clarté de l'exposé nous avons préféré conserver la terminologie en vigueur en textométrie.

Cooccurrent d'*avenir*	Fréq.	Cofrq.	Indice	Dist. Moy.	Cooccurrent de *crise(s)*	Fréq.	Cofrq.	Indice	Dist. Moy.
est	671	32	2	4,5	été	84	7	2	5,6
essentielle	3	2	2	2,0	cette	140	9	2	1,0
atout	3	2	2	2,0	guerre	28	4	2	5,2

Tableau 2. Illustration de la cooccurrence : principaux cooccurrents d'*avenir* et de *crise(s)*.[9]

L'*analyse des correspondances* (Benzécri, 1973) opère une synthèse globale des relations entre mots (ou plus généralement traits linguistiques, motifs) et textes (ou plus généralement parties du corpus). Les mots sont comparés les uns aux autres sur la base des textes qui les emploient ; et réciproquement les ressemblances et écarts entre textes sont évalués par le vocabulaire qu'ils mobilisent. L'analyse des correspondances génère alors mathématiquement une épure faisant ressortir les rapprochements et oppositions les plus forts entre mots, entre textes, et indirectement entre mots et textes. La forme privilégiée de résultat est une représentation cartographique en deux dimensions, positionnant mots et textes dans un espace plan, et dont la lecture est guidée par des indicateurs rendant compte des éléments les plus significatifs pour les configurations observées, sur lesquels peut s'appuyer l'interprétation (figure 2).

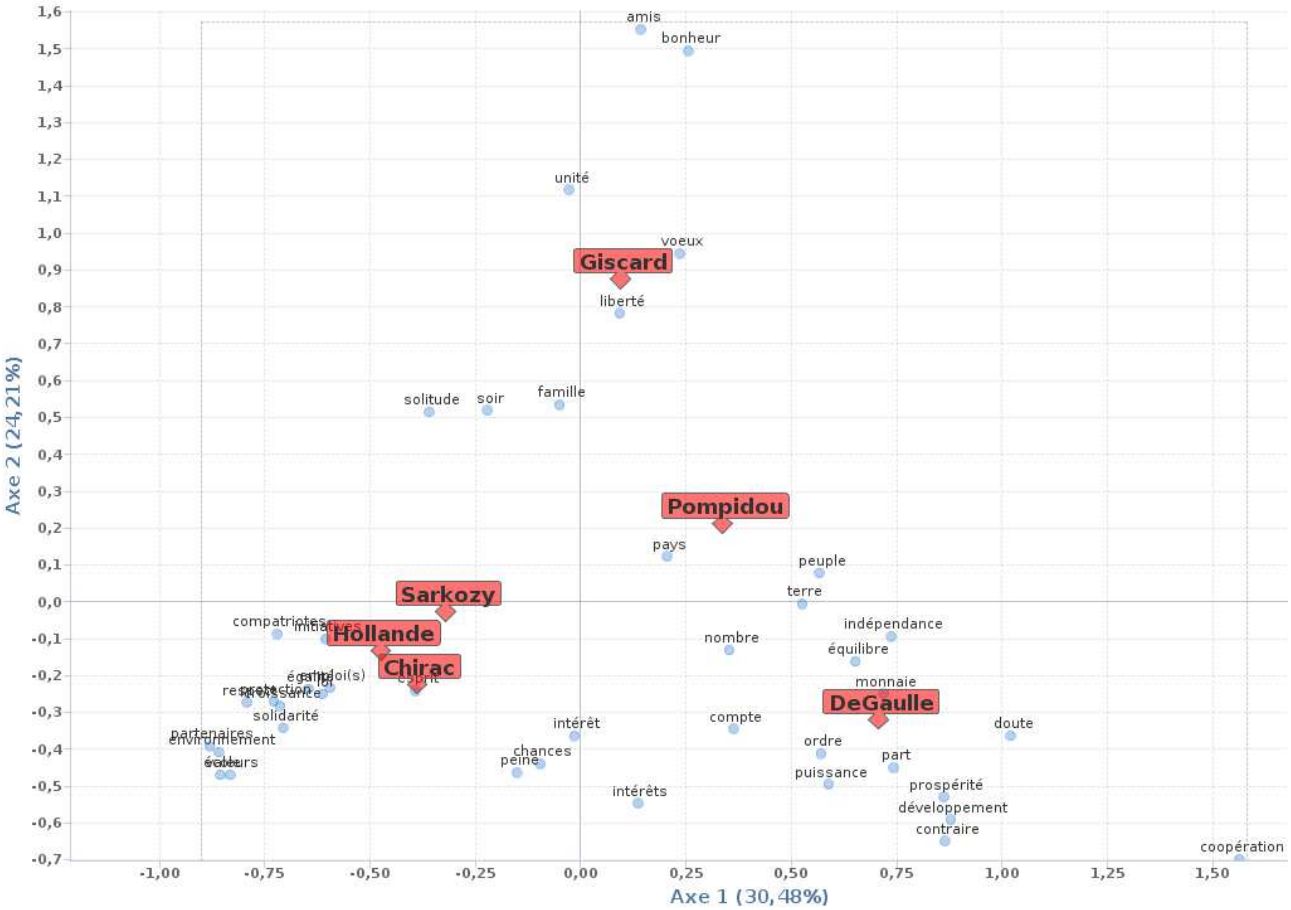

Figure 2. Illustration de l'analyse des correspondances : représentation géométrique des présidents en lien avec une sélection de noms communs[10].

9 Guide de lecture du tableau : le mot *avenir* (volet gauche) compte 79 occurrences dans le corpus, et le mot *crise(s)* (à droite) 58 occurrences au total. Les indicateurs chiffrés sont : la fréquence totale du cooccurrent dans le corpus, le nombre de fois où il est à moins de 10 mots du pivot (*avenir* ou *crise(s)*), l'indice statistique (plus il est élevé plus la cooccurrence est remarquable), la distance moyenne entre le cooccurrent et le pivot quand ils sont dans le même voisinage (ici limité à 10 mots d'écart maximum).

10 Critères de sélection : pour le calcul, on n'a retenu que les noms les plus fréquents (employés au moins 12 fois), en écartant aussi manuellement quelques erreurs d'étiquetage. Pour le graphique, on n'affiche que les points dotés

La *classification automatique* trouve une application privilégiée en portant sur des *unités de contexte* de l'ordre de la phrase ou du paragraphe, avec la méthode ALCESTE, dite aussi méthode Reinert (Reinert, 1990) : les classes construites peuvent être lues comme approchant des univers sémantiques, décrits par les mots caractéristiques que l'on peut automatiquement associer à chaque classe par un calcul de spécificités ou une mesure statistique telle que le chi-2 (figure 3).

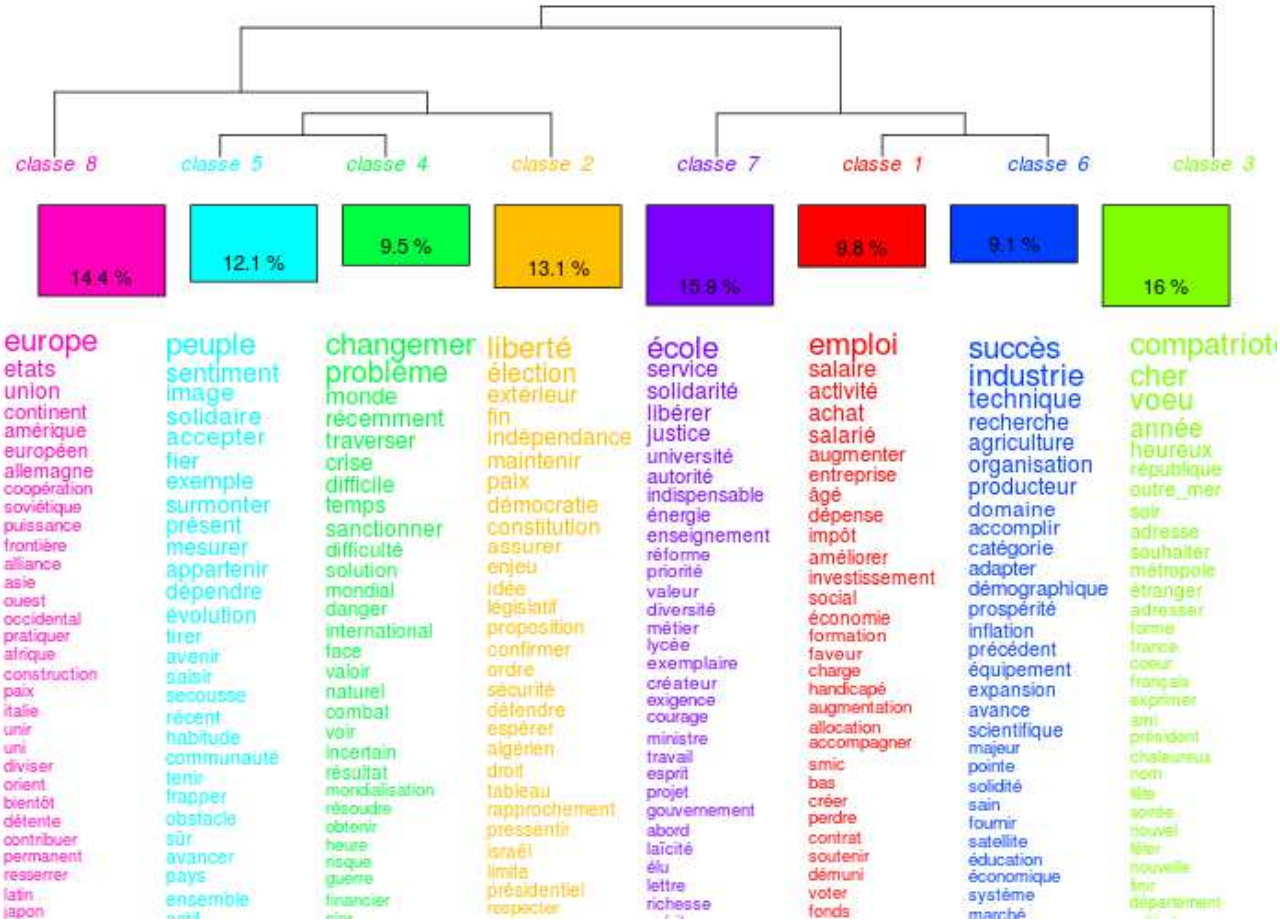

Figure 3. Illustration de la classification automatique selon la méthode ALCESTE sur le corpus VOEUX.

La méthode textométrique ne s'arrête pas à la sélection statistique de mots, elle affirme l'importance de les interpréter en contexte et prévoie plusieurs modes complémentaires de consultation de ceux-ci (Pincemin, 2006). La *concordance* utilise une disposition en tableau (figure 4). Chaque contexte occupe une ligne, qui se déploie sur quatre colonnes : une indication de localisation dans le corpus sous forme de référence pour situer l'extrait, quelques mots qui précèdent l'occurrence (contexte gauche), l'occurrence elle-même (dite « pivot » de la concordance), quelques mots qui suivent l'occurrence (contexte droit). La présentation en tableau assure la superposition des contextes, et le tri des contextes (tri alphabétique des mots en commençant par ceux les plus proches du pivot) fait ressortir les contextes répétitifs, qu'ils soient courts (par exemple pour un mot composé ou une locution non codés comme une seule unité initialement) ou plus longs (phraséologie). La concordance fournit une vue dense des contextes d'emploi, et favorise la mise en évidence des régularités de construction locales, juste autour de l'occurrence, au prix d'un effacement des effets de mise en page et de la disposition dans la

d'une bonne représentation (peu déformée, avec $\cos^2 > 0{,}75$) ou particulièrement impliqués dans la structure mise au jour dans ce plan (contribution à l'un des deux axes > 2 %) (ces critères quantitatifs font que le président Mitterrand n'apparaît pas dans ce graphique).

structure textuelle. Pour ces derniers, un autre mode de retour au texte est disponible, qui met en évidence les occurrences au fil du texte, en conservant la présentation de celui-ci.

Localisation	Contexte gauche	Pivot	Contexte droit
Mitterrand, 1988	d'entrevoir, soit en même	temps	que le meilleur moyen de créer des emplois
Chirac, 1998	pour l'avenir. En même	temps	, hélas, de nombreuses victimes tombaient au
De Gaulle, 1960	les vieilles gens. En même	temps	il nous faut continuer à abaisser les barrières
Giscard, 1979	%. Et dans le même	temps	, nous nous sommes engagés dans une politique
Chirac, 2004	exigences et aux défis de notre	temps	. La loi d'orientation commencera à s'
Giscard, 1976	ses forces aux problèmes de notre	temps	, comme elle est capable de le faire
Giscard, 1980	plus sur les problèmes de notre	temps	que tous ceux qui vous ont précédés.
De Gaulle, 1961	possible, car, en notre	temps	, sont à l'œuvre des forces énormes
Chirac, 1996	d'adapter la France à son	temps	. Cela exige de penser et d'agir
Chirac, 1996	nous adaptons notre nation à son	temps	, si nous nous appuyons sur ce que
Pompidou, 1971	comme ils guettent, en tout	temps	, toute nation, et notre situation ne
Chirac, 2000	fut, en réalité, un	temps	de prise de conscience. Conscience des risques
Chirac, 2001	à chacune et à chacun un	temps	d'adaptation pour trouver ses repères, apprendre
Giscard, 1975	la fraternité des Français en un	temps	où elle est nécessaire. Cette fraternité veut
Chirac, 2001	entre vous, c'est un	temps	de joie, de retrouvailles, de chaleur

Figure 4. Illustration de la concordance : extrait de la concordance du mot *temps* triée à gauche.

Du point de vue sémantique, cette association des opérations statistiques et de la lecture en contexte permet de tirer parti à la fois de la finesse de la lecture méthodique, systématique et organisée, des contextes, et à la fois des suggestions faites par les calculs, qui peuvent tantôt confirmer les attentes, tantôt les compléter, les déplacer, voire les remettre en question.

2.3. Caractéristiques : pourquoi distinguer la textométrie d'autres méthodes d'analyse quantitative ?

L'équilibre entre quantitatif et qualitatif est une des premières originalités de la textométrie à souligner. La démarche typique en *text mining* (ou fouille de texte) consiste à partir d'un corpus de grande taille, à le réduire peu à peu en éliminant les mots trop rares ou trop fréquents et les singularités, et à générer des visualisations suggestives, dont la lecture serait intuitive et autonome : les outils logiciels de *text mining* ne prévoient généralement pas d'interface pour la consultation et le dépouillement méthodique des contextes en lien avec les graphiques produits. À l'inverse, le retour au texte est une fonction au cœur des logiciels de textométrie. Ainsi, dès l'apparition de la notion d'hypertexte, Hyperbase a recouru massivement aux liens pour faciliter le retour au texte depuis tout résultat de calcul. Et dans les logiciels récents, TXM développe la qualité de l'édition numérique des textes du corpus, permettant le cas échéant leur présentation synoptique avec le document source (manuscrit médiéval, images des pages d'une édition de référence, enregistrement audio-visuel dont on analyse la transcription), afin que l'interprétation puisse se baser sur toutes les dimensions de la source linguistique (typographique, matérielle, etc.). Pour la démarche textométrique, l'attention au texte source, le respect de son intégrité, la possibilité de sa consultation précise à tout moment de l'analyse, sont centraux.

En matière de statistique, la textométrie s'intéresse à l'analyse des textes, elle ne cherche pas à modéliser la langue dans l'optique par exemple de prédire l'apparition de mots ou de générer des textes. Ainsi, elle recourt à la partie dite *exploratoire* de la statistique, par opposition aux procédures probatoires qui intéressent d'autres approches.

Ensuite, le choix précis des calculs statistiques utilisés en textométrie n'est pas simplement heuristique, mais motivé par une modélisation claire des données. Ainsi, la plupart des calculs de « mots-clés » (*keywords*) ou de cooccurrence (*collocation*) qui mobilisent des tests statistiques (chi-2, *t-score*, *z-score*, *log-likelihood*)[11] font l'hypothèse d'un certain type de répartition des mots (loi normale) ; alors que le calcul des spécificités correspond précisément à une modélisation de la répartition aléatoire des mots entre les parties (loi hypergéométrique), c'est ce qu'on appelle en statistique un test non paramétrique, à savoir qui ne dépend pas d'une hypothèse de forme de distribution. Plus précisément, le calcul des spécificités est un test exact de Fisher[12], demandant un calcul plus lourd (devenu très facile à mettre en œuvre avec les moyens informatiques actuels), mais plus juste, plus proche du questionnement linguistique que l'on veut exprimer (Gries, 2014).

Le même genre de remarque peut être fait pour l'analyse des correspondances. D'autres calculs existent, permettant de produire des cartographies à partir d'un tableau (matrice) croisant mots et textes : analyse en composantes principales, analyse sémantique latente (LSA[13]), *multidimensional scaling*. L'opération de synthèse repose sur une même opération mathématique de décomposition en valeurs singulières (SVD[14]), mais l'analyse des correspondances prend pleinement en compte le fait que l'on a affaire ici à un tableau de type table de contingence (où le regroupement de mots ou de textes a un sens) et utilise la distance du chi-2 qui est la plus adaptée à cette structure de données (Lebart *et al.*, 1998, p. 63-69). Là encore donc, le choix du calcul prend plus particulièrement en compte la nature et la structure des données textuelles.

Quant à la classification automatique, l'un des parcours mis au point en textométrie, celui de la méthode ALCESTE (Reinert, 1990), montre que l'on construit des classes sémantiques de mots non pas en appliquant directement les algorithmes de classification au vocabulaire d'un corpus, mais indirectement : on classe des segments de texte (de l'ordre de la phrase ou du paragraphe), et on en déduit des ensembles de mots associés. En effet, les algorithmes de classification génèrent des partitions au sens mathématique du terme : chaque élément appartient à une classe et une seule. Or il est évident pour le linguiste qu'une telle structure serait inappropriée pour des regroupements sémantiques de mots, certains mots (représentés par leur forme graphique) relevant de plusieurs classes sémantiques (polysémie), d'autres ayant une valeur thématique trop faible ou singulière pour qu'ils prennent part à la description. Ainsi, la classification proposée s'adapte à la non-univocité sémantique des unités lexicales.

En mettant l'accent sur la motivation, l'adéquation et la clarté des calculs, la textométrie ne prend pas le parti de la technicité statistique, mais vise à rester au plus près de la nature linguistique des données et de ses propriétés spécifiques[15]. Dans l'éventail des traitements proposés, on trouve aussi bien les calculs statistiques élaborés mais motivés qui sont évoqués ci-dessus (spécificités, analyse des correspondances), que des tableaux, sélections et visualisations reposant sur de simples décomptes et tris, ne requérant aucun bagage mathématique spécialisé (concordances, segments répétés, visualisations de la répartition d'un mot au fil du corpus, etc.).

---

11 Tous ces tests, de formulation mathématique plus ou moins complexe, peuvent servir à évaluer la fréquence étonnamment élevée d'un mot dans une partie du corpus : soit cette partie est un texte, auquel cas le mot fréquent est alors un potentiel « mot-clé » du texte, soit cette partie est l'ensemble des voisinages d'un mot, et l'on mesure l'attirance entre deux mots.

12 Ce test statistique a été conçu par Ronald Fisher, biologiste et statisticien du début du XX^{e} siècle ; et c'est Pierre Lafon (1980) qui a proposé son application aux données textuelles sous le nom de calcul des spécificités. Le calcul est équivalent à énumérer toutes les combinaisons possibles de mots dans les différentes parties du corpus, et à évaluer la probabilité de chaque fréquence en fonction de la proportion de fois où on peut théoriquement la rencontrer si toutes les combinaisons étaient également possibles. Le test est donc exact en ce qu'il modélise directement et complètement l'hypothèse de répartition aléatoire des mots entre les parties.

13 LSA est l'acronyme de *Latent Semantic Analysis*.

14 SVD est l'acronyme de *Singular Value Decomposition*.

15 Par exemple, comme évoqué précédemment, en prenant en compte les phénomènes de polysémie (méthode ALCESTE), en se basant sur l'observation des mots en corpus plutôt qu'en postulant une distribution statistique « standard » (calcul des spécificités), en rendant compte des variations possibles de portée des unités lexicales ou textuelles par fusion ou dégroupement (adéquation de l'AFC aux tables de contingence).

Enfin, le chercheur garde une place centrale dans l'analyse, il ne s'agit en rien d'une analyse linguistique ou sémantique automatique. L'ordinateur est mobilisé pour sa capacité de mémoire et sa vitesse de calcul ; en revanche, la conduite de l'analyse relève du chercheur. La composition du corpus, son découpage en unités lexicales et textuelles, les informations pertinentes à considérer, les points d'entrée des investigations, la qualification des résultats des calculs en éléments de réponse au questionnement linguistique, tout cela reste du ressort de l'humain et de son rapport interprétatif à la langue.

3. Principes opératoires : points d'appui et révélateurs sémantiques

Les analyses textométriques ne sont pas nécessairement sémantiques, mais les principes opératoires de la démarche font qu'elle se prête particulièrement bien aux observations sémantiques. Nous considérons aussi que les différentes facettes du langage – morphologie, syntaxe, sémantique, etc. – sont interdépendantes, et que des traits morphologiques ou syntaxiques peuvent prendre un rôle sémantique.

3.1. Le global : de la sémantique du texte, de l'intertexte, du corpus, à la sémantique lexicale

Dans l'analyse textométrique, la composition du corpus est déterminante car c'est le corpus qui sert de référence statistique pour évaluer les usages, les associations lexicales, les sur- ou sous-emplois. Le corpus est la représentation de la langue pour les opérations d'analyse qui pourront être lancées (relevés, tris, calculs statistiques) ; il peut être enrichi d'annotations (morphosyntaxiques, syntaxiques, sémantiques…) et d'indications d'ancrage situationnel via des métadonnées (auteur, caractéristiques d'un locuteur, genre textuel, etc.), mais c'est bien au sein de cet ensemble fini et explicite que les faits et effets sémantiques sont mis en évidence.

Ce rôle de *corpus de référence* est opératoire dans les calculs statistiques : si la composition du corpus change, les résultats varient en conséquence ; les attirances qui ressortent ne sont plus nécessairement les mêmes, les sur- ou sous-effectifs non plus, au regard du nouvel ensemble (Bernard, 2000). Il n'est pas sans influence non plus sur les traitements plus qualitatifs : ainsi, l'étude du sens d'un mot peut passer par le relevé systématique de ses contextes d'apparition, et leur organisation en regroupements (Bourion-Jacquemin, 2001). Les divers sens et acceptions ainsi identifiés sont évidemment relatifs au corpus.

On reconnaît là une forme de détermination du local par le global : la valeur, l'interprétation d'une unité n'est pas intrinsèque, ni même fixée par son contexte phrastique, mais tous les niveaux de contexte jusqu'au niveau intertextuel ont potentiellement leur influence sur la construction du sens. Ainsi, quand bien même chaque unité lexicale serait parfaitement définie et annotée sémantiquement, cette information sémantique resterait modulée par sa contextualisation au sein du corpus. La sémantique lexicale n'est pas indépendante du corpus, elle peut être observée en interaction avec les niveaux de contexte les plus larges dans un corpus dont la construction a été élaborée pour faire sens.

En effet, faut-il le rappeler, et ici donc tout particulièrement, le corpus n'est pas un amas quantitatif de données langagières, dont la valeur tiendrait à l'ampleur ; mais c'est un ensemble construit, respectant les frontières linguistiques et donc, sauf cas particulier, celle des textes, car chaque unité s'interprète naturellement au sein du texte qui la contextualise (Péry-Woodley, 1995 ; Rastier, 2001 et 2011).

Le corpus de référence est non seulement déterminant, mais aussi seul nécessaire : l'annotation sémantique n'est pas un préalable à l'analyse sémantique. L'approche textométrique permet des recherches sémantiques en prenant appui sur les contextualisations, même sans aucune description sémantique des unités lexicales elles-mêmes. La méthodologie ALCESTE, par exemple, trouve

automatiquement des groupements thématiques de mots par la seule considération de la répartition des mots sur l'*ensemble* des contextes locaux d'un corpus.[16]

3.2. Le contexte : une sémantique distributionnelle à tous les paliers de contexte

Les outils de relevé d'occurrences intégrés aux logiciels de textométrie permettent de capter par exemple l'ensemble des adjectifs qui apparaissent en fonction d'épithète d'un nom donné dans le corpus, ou inversement l'ensemble des noms qualifiés par un même adjectif. Ce mécanisme de construction de séries paradigmatiques peut être généralisé à bien d'autres constructions syntaxiques et formes de contextes : verbes utilisés dans des constructions négatives, noms déterminés par des possessifs de première personne, etc. Par ailleurs, les concordances permettent de mettre en évidence des régularités non prévues sur les contextes immédiats, tout en associant des informations de contexte global utiles à l'interprétation. Et les cooccurrences assurent un repérage plus souple, avec des récurrences marquées mais dans des voisinages variables. Le calcul de spécificités permet, quant à lui, d'identifier des affinités marquées entre des usages lexicaux et des caractéristiques textuelles. Les usages lexicaux peuvent ainsi être éclairés par des informations d'associations distributionnelles avec des voisinages plus ou moins stricts, plus ou moins prévus, plus ou moins larges.

Qu'il s'agisse des consultations plus qualitatives des relevés de contextes, ou des procédures statistiques, les contextes locaux ou globaux sont les points d'appui de l'analyse, ce sont les contextes présents dans le corpus qui apportent un éclairage sur les unités et sur les relations entre elles.

3.3. Le contraste : une sémantique différentielle

Plusieurs des calculs statistiques centraux de la textométrie fonctionnent par contraste : les spécificités s'attachent à repérer des sur- ou sous-emplois, par contraste avec une équirépartition sur un ensemble de parties ; et l'analyse des correspondances vise par construction à dégager les dimensions de plus grande variation au sein du corpus, celles sur lesquelles les contrastes seront les plus grands. Ainsi, ce sont les similarités et les contrastes d'usages (usages caractérisés par les voisinages contextuels) qui font l'objet de la plupart des observations.

Comme nous l'avons vu, la sémantique n'est donc plutôt pas ici *référentielle*, puisqu'elle n'a pas besoin de s'appuyer sur une ontologie qui fait le lien avec des concepts et des objets d'un univers sémantique. En revanche elle procède toujours de façon *différentielle*, en suggérant des affinités, des relations, des oppositions. Il s'agit bien ici de rendre compte du mécanisme opératoire interne de la textométrie, du fait des traitements auxquels elle recourt, et non d'un positionnement exclusif. De fait, ce fonctionnement différentiel peut être couplé avec une analyse référentielle, si l'on opère l'analyse textométrique sur un corpus enrichi par des annotations issues d'une analyse référentielle préalable.

Le rapport au monde n'est pas pour autant absent de l'analyse textométrique. D'une part, comme l'analyse est pilotée et validée par le chercheur, celui-ci mobilise toute son expertise et sa connaissance des textes (biographie de l'auteur, contexte de production, etc.) pour enrichir et renforcer tant la conduite de l'analyse que l'interprétation des phénomènes pointés par les calculs. D'autre part, le corpus lui-même porte également les marques d'un certain rapport à un environnement extralinguistique. Chaque texte renvoie en effet à ses pôles extrinsèques (Rastier, 2001), il les reflète comme en creux, par les traces linguistiques de son genre textuel, de sa situation d'énonciation. Par ailleurs, le corpus peut être structuré par des métadonnées (ex. date, source…) qui pourront ainsi faire intervenir directement dans l'analyse des éléments extratextuels et situationnels : ces métadonnées peuvent être indiquées dans les concordances, pour grouper ou situer la lecture des contextes ; les métadonnées peuvent aussi servir à organiser le corpus en lui

16 Avec une vingtaine d'années d'avance, elle anticipait sur les recherches actuelles du *topic modeling*, certaines approches reprenant exactement les mêmes principes.

donnant un ordre ou en le divisant en parties, structure sur laquelle appliquer ensuite des opérations de mise en évidence des évolutions ou des contrastes. Ainsi, ces dimensions non linguistiques peuvent néanmoins contribuer à l'analyse, qui n'est pas purement interne.

3.4. Le factuel : une sémantique matérielle

Comme plus généralement en linguistique de corpus, la textométrie se fonde sur les observations des faits présents dans les données. Très concrètement, les décomptes à la base des calculs traduisent au moins deux seuils linguistiquement significatifs. Le premier seuil est celui qui sépare l'absence (fréquence nulle) de l'attestation (une ou plusieurs occurrences), qui marque la possibilité, l'existence. Le second seuil différencie l'occurrence unique (singleton, *hapax*) de l'apparition répétée, qui marque un emploi confirmé, *a priori* plus régulier (*i.e.* suivant les règles du contexte), par opposition à une apparition singulière, exceptionnelle. Les fréquences suivantes peuvent se lire ensuite par ordre de grandeur – quelques occurrences, quelques dizaines, quelques centaines – et l'on passe ainsi des « basses fréquences » aux « hautes fréquences », des mots à la présence ponctuelle et clairsemée à ceux qui dominent le propos et « occupent le terrain ». Bien que le linguiste puisse être décontenancé par cette traduction très abrupte des mots aux fréquences, aux premiers seuils presque trop nets et aux partages suivants mal définis, certains traitements textométriques exploitent déjà ces simples relevés de fréquences, qui donnent heuristiquement un aperçu des usages.

Les traitements plus statistiques se basent sur ces fréquences discrètes (représentées par des nombres entiers) pour produire des mesures et des représentations continues (où l'on passe progressivement d'un point à un autre, d'un nombre décimal à un autre), susceptibles d'approcher davantage des effets sémantiques plus souples, moins catégoriques, et de nuancer les écarts de fréquence en faisant la part entre ceux qui s'avèrent plus significatifs et marqués, et ceux qui apparaissent plus incertains, plus contingents.

Le fait de se baser sur la matérialité de l'expression linguistique, le décompte de mots ou d'autres types d'unités, et d'appliquer des procédures automatiques, facilite un traitement systématique et formalisé, déterministe et reproductible. Ce comportement clair aide à l'interprétation critique des faits observés, ceux-ci s'expliquant dans le contexte d'un certain corpus, avec un certain paramétrage des traitements.

3.5. L'interprétation : une sémantique interprétative et plurielle

Le caractère déterministe des traitements n'implique pas l'univocité de l'analyse sémantique qui peut être faite, car l'analyse textométrique suppose un parcours construit par le chercheur, qui choisit des points d'entrée, compose des traitements, évalue les résultats. Le calcul produit toujours un *résultat*, mais ce résultat n'est pas en soi une *réponse* à la problématique de recherche, le passage du résultat à la réponse suppose la compétence critique et l'interprétation du chercheur. Ainsi, l'étude d'un même aspect sémantique sur un même corpus avec un même logiciel textométrique peut donner lieu à une pluralité d'interprétations.

Cette pluralité est cependant contrôlée : les interprétations se fondent sur des indices reproductibles, et des mesures chiffrées hiérarchisent l'importance des différents points d'appui.

La conception interprétative qui prévaut en textométrie s'écarte donc de la modélisation sémantique sous-jacente aux campagnes d'évaluation des analyseurs sémantiques en traitement automatique des langues. Pour ces derniers, l'analyse sémantique est une opération qui associe à un élément linguistique (mot, phrase, texte) la représentation de sa signification. Le sens est déterminé et unique, au moins dans le contexte de la tâche que vise à automatiser le logiciel d'analyse sémantique TAL. L'évaluation de la qualité de l'analyse sémantique réalisée par un outil donné se mesure alors par l'écart par rapport à l'analyse correcte, correspondant à *la* réponse attendue. Or pour la textométrie, la qualité d'une analyse n'est pas évaluable au niveau du seul résultat : non

seulement il y a plusieurs analyses sémantiques possibles, mais aussi la valeur de l'analyse intègre la clarté et la cohérence du cheminement qui donne sens aux observations.

La dynamique et la pluralité des interprétations concernent non seulement le niveau global des textes et du corpus sur lesquels porte l'analyse, mais aussi le niveau local des occurrences. Même si on attache une étiquette sémantique à chaque mot, le sens n'est pas contenu dans chaque unité, et l'analyse sémantique du tout n'est pas obtenue par une combinaison des informations sémantiques présentes dans les unités. Ainsi, la variation de la composition et des limites du corpus modifie l'univers de référence pour les statistiques, et l'importance des unités, leur portée, leurs relations, leurs affinités textuelles et intertextuelles se trouvent remodelées par le nouveau contexte.

3.6. L'ouverture : une sémantique ouverte

Une analyse textométrique peut se conclure sur une série de résultats convergents, qui concourent à une interprétation fortement étayée. Elle n'est cependant jamais close (complète, définitive), étant donné la multiplicité des investigations possibles et la pluralité des interprétations.

Par ailleurs, la prise en charge des calculs par le logiciel et la facilité de mise en œuvre des traitements déploie un large éventail heuristique : il n'y a pas d'hésitation à tenter un calcul, à tâtonner pour mieux comprendre ses effets et mieux l'ajuster, à risquer une hypothèse, à explorer un point d'entrée moins connu et plus incertain, à découvrir de nouvelles données. Ces occasions d'investigations plus « aventureuses » nous semblent aller dans le sens d'une analyse sémantique ouverte à la nouveauté, même si, bien sûr, celle-ci suppose la créativité et l'attention experte du chercheur dans ses parcours et ses observations.

4. Mais où est le sens ? La question des unités pour l'analyse

Les développements précédents nous ont permis de souligner que, même dans ses aspects quantitatifs, la textométrie fait des choix pour être au plus proche de la langue et en accord avec des propriétés linguistiques. Reste cependant un autre plan susceptible d'inquiéter le linguiste : la qualité des analyses ne repose-t-elle pas sur la justesse des unités décomptées ? Mais alors, les découpages automatiques en mots, opérés à large échelle sur les corpus numériques, et souvent critiquables dans le détail pour un œil expert, ne compromettent-ils pas dès le départ la valeur linguistique des résultats ?[17] Nous voudrions ici apporter un point de vue moins dubitatif, susceptible au contraire de stimuler la réflexion linguistique : la définition des bonnes unités est-elle possible, déjà au plan théorique ? Plus avant, n'est-elle pas fondamentalement plurielle et dynamique, relative au contexte textuel et intertextuel comme au point de vue de l'interprète ? La démarche textométrique se révélerait alors encore particulièrement particulièrement proche du fonctionnement de la langue en discours.

4.1. Pourquoi et comment définir des unités ?

La partie quantitative de la textométrie suppose de compter des unités en relation d'inclusion : des mots dans des textes, ou dit plus généralement, des éléments dans des parties, des contenus dans des contenants[18]. On peut considérer qu'il s'agit dans les deux cas d'*unités*, qui, selon les calculs, sont mises en relation avec d'autres avec un *rôle*, soit de contenant, soit de contenu : par exemple, des phrases peuvent être des unités de contexte de mots pour une classification (elles jouent le rôle de contenants), ou des éléments des textes pour une caractérisation de ceux-ci en fonction des types de phrases qu'ils contiennent (les phrases ont alors un rôle de contenus). La définition des unités procède de deux opérations formelles : il faut d'une part *délimiter* les unités (à savoir, préciser

17 De façon imagée et lapidaire, ce serait l'idée que les informaticiens ont dénommée le « GIGO » : *Garbage in, garbage out*...

18 Voire, pour une formulation abstraite et euphonique imaginée par André Salem, des *types* dans des *topes* (Söze-Duval, 2008).

quelle(s) portion(s) matérielle(s) du texte en relève), et d'autre part les *identifier* (à savoir, de quel type cette unité est-elle une occurrence, peut-elle être considérée comme la répétition d'une ou plusieurs autres unités ailleurs dans le corpus).

Pour les « contenus », la textométrie s'intéresse d'abord aux unités d'ordre lexical, sans exclure d'autres paliers de description linguistique. C'est autour du palier lexical que l'on trouve aussi les morphèmes (infra-lexicaux) et les syntagmes (supra-lexicaux), éléments potentiellement utiles pour l'analyse sémantique. La plupart des analyses textométriques s'appuient sur les sorties d'un analyseur morphosyntaxique pour définir ces unités lexicales, jouant le rôle de contenu : les délimiter par une opération dite de *tokenisation*, et les identifier en reconnaissant l'occurrence d'un mot désigné par son lemme, et décrit par sa catégorie grammaticale et sa flexion.

4.2. Des unités remodelées et affinées par l'analyse : unités élémentaires et unités descriptives

Le traitement automatique des langues fournit ainsi une *approche* des syntagmes (par le découpage opéré) et des morphèmes (à travers les lemmes et les informations de flexions portées par l'étiquetage). La réalisation effective de ces unités peut comporter des erreurs (dues aux limites du traitement automatique : cas non prévu, etc.), et même la définition des unités dans le modèle de l'outil peut être discutable (désaccords sur la manière de catégoriser les unités, sur le modèle de langue auquel se rapporte l'outil) ; néanmoins, et cela est déjà essentiel, le découpage et la catégorisation de ces unités sont en relation avec des articulations linguistiques, par opposition à un découpage en unités qui serait « orthogonal », sans rapport aucun, avec le fonctionnement linguistique. C'est là une des clés de l'analyse : à partir du moment où elle s'appuie sur des unités rendant compte au moins partiellement de délimitations et de catégorisations linguistiques, l'analyse a les moyens, par recomposition et restructuration des unités, de faire évoluer la description et de l'affiner.

Autrement dit, il s'agit en quelque sorte de passer d'une vision « atomique » de l'analyse à une vision herméneutique. La vision atomique considérerait que l'analyse manipule des unités, conservées du début à la fin, la démarche n'étant qu'une réorganisation, un assemblage, de ces atomes, et le résultat se décodant directement à partir de la valeur de chaque atome. La vision herméneutique fait au contraire évoluer les unités au fil de l'analyse : certaines unités sont déconstruites, d'autres réélaborées au croisement d'unités antérieures. Les unités n'ont pas le caractère stable, définitif, isolé, insécable, de l'atome ; mais, sans renier de premiers points d'appui pris comme unités transitoires ou hypothèses d'unités, les unités effectives de l'analyse se construisent et s'affinent pendant l'analyse elle-même.

Pour comprendre linguistiquement la démarche textométrique, il faudrait ainsi distinguer : des unités *élémentaires*, des unités *descriptives* et des unités *caractérisantes* (Bommier-Pincemin, 1999). Les unités élémentaires sont les points d'appui initiaux de l'analyse, considérés comme une première approche de la réalité à décrire. Les unités descriptives sont des constructions à partir des unités élémentaires, que le corpus révèle comme plus adaptées et plus productives. Les unités caractérisantes seront alors les unités descriptives retenues au terme de l'analyse, et qualifiées dans leur usage en corpus, pour rendre compte d'un résultat obtenu.

Ainsi, par exemple, le calcul textométrique des segments répétés a précisément été conçu comme un moyen d'ajuster la délimitation initiale des unités, avec le repérage puis la prise en compte de figements, de phraséologies (Lafon et Salem, 1983). Ou encore, les classes thématiques ou univers sémantiques mis au jour par la méthode ALCESTE (Reinert, 1990) permettent de restructurer l'ensemble des unités lexicales, en distinguant les homonymes comme en rassemblant les variations de formulation relevant d'un même champ lexical. Dans les deux cas, le découpage élémentaire initial est revu à l'aune d'une analyse d'ensemble sur le corpus : c'est en s'appuyant sur des récurrences globales que de nouvelles unités locales sont définies.

La distinction entre unités élémentaires et unités descriptives, et ainsi le fait de savoir que les unités initiales n'imposent pas une vision définitive du corpus, expliquent que l'analyse textométrique

puisse également quelquefois se fonder uniquement sur un découpage très sommaire du texte sur la base de « caractères délimiteurs » comme l'espace ou la ponctuation, sans que cela empêche la conduite d'analyses linguistiquement pertinentes. Étienne Brunet (2007) montre même que des structures profondes du corpus restent présentes et arrivent à se manifester statistiquement, y compris pour des unités élémentaires apparemment particulièrement dégradées, comme des séquences de trois caractères, voire de simples successions consonnes / voyelles.

4.3. À rebours des évidences

À ses débuts, la communauté de recherche en textométrie fut traversée par le débat sur la lemmatisation (Brunet, 2000) : les partisans, autour de Charles Muller (1977 et 1984), plaidaient pour une préparation du texte avant son analyse, identifiant linguistiquement chaque mot par l'entrée lexicale correspondante, permettant de reconnaître un même mot sous ses différentes flexions ; les adversaires, principalement représentés par l'équipe du laboratoire de Saint-Cloud, soulignaient l'arbitraire et les limites de tout choix de recodage, et préféraient par conséquent une fidélité à la forme graphique originelle du texte (Geffroy *et al.*, 1974 ; Tournier, 1985). « Sur quoi pouvons-nous compter ? » questionnait ainsi malicieusement, mais non sans profondeur, le titre de l'article de Maurice Tournier (1985). Les arguments des uns comme ceux des autres montraient leur pleine conscience de travailler sur des unités jamais totalement satisfaisantes ni définitives ; de même, il était reconnu que ces choix de pré-analyse étaient évidemment à lier aux perspectives d'analyses : une recherche sur les temps verbaux par exemple supposait d'avoir gardé trace de ceux-ci. La querelle scientifique put prendre fin quand il n'y eut plus de choix unique à faire préalablement à l'analyse, et que les deux options (lemmatisation ou non) devinrent souplement et conjointement disponibles, grâce aux analyseurs développés par les TAL (avec lesquels la lemmatisation n'était plus un investissement initial majeur), aux nouveaux formats de codage de corpus (structurés), et à l'adaptation des logiciels de textométrie (capacité à considérer plusieurs points de vue – plusieurs systèmes de types[19] – pour une même unité). La lemmatisation, plus sophistiquée, n'est donc pas d'évidence supérieure, pas plus que la non lemmatisation : la voie la plus pertinente, linguistiquement, est à construire au cas par cas entre ces deux représentations.

L'informatique considère des données : des données textuelles dans notre cas. Mais un point de vue linguistique s'élève contre l'apparente immédiateté présumée par ce terme (Rastier, 2011). Quand on a affaire à la langue, les données (l'objet à considérer dans le traitement informatique) ne sont pas données (déjà définies, livrées, prêtes à l'emploi). La transparence de la langue est illusoire, et les choix de modélisation s'avèrent multiples. Ainsi, la notion même de « texte brut » ou de « texte source » apparaît illusoire : car peut-on séparer ce qui serait l'essence même du texte de sa manifestation matérielle ou numérique, de son support ? N'y a-t-il pas toujours des choix de représentation, d'implémentation, d'« encrage », à travers lesquels se présente le texte et dont ne peut s'affranchir la lecture ? Ainsi les pratiques d'analyse assistée par ordinateur doivent-elles s'ouvrir à la philologie numérique, à savoir la prise en compte des choix de codage et de représentation. Il faut donc renoncer à la quête d'une représentation première, unique, évidente, neutre, à des unités simples et universelles, qu'il suffirait de cueillir à même le texte. Les premiers choix de représentation du texte sont toujours déjà un point de vue sur le texte. Les unités les plus utiles pour décrire et comprendre le texte seront au terme de l'analyse plutôt qu'à son amorce : comme nous l'avons vu dans la section précédente, et comme le confirme l'herméneutique, les unités sont construites et s'affinent avec la connaissance d'ensemble du texte.

En pratique aussi, les unités sur lesquelles vont porter les traitements textométriques ne sont pas imposées par la méthode. Même si telle ou telle option est parfois l'unique adoptée par tel ou tel

19 Par exemple dans le système des catégories grammaticales, chaque mot en contexte peut être rapporté à un type qui est sa catégorie ; dans le système des lemmes (en quelque sorte le dictionnaire du corpus), chaque mot en contexte peut être rapporté à un type qui est le lemme lui correspondant. Les types sont une manière de formuler que différentes occurrences au fil du texte sont la répétition d'une même entité plus générale (un même type) sous un certain angle (un système de types).

logiciel, l'approche textométrique en soi ne fixe pas la nature des unités. Ainsi, il faut combattre certains préjugés qui perdurent : la textométrie *peut* être mais n'est pas obligatoirement, une analyse portant sur un simple découpage du texte en chaînes de caractères ; la lemmatisation *peut* être un choix de représentation efficace, mais n'est pas toujours en soi meilleure, et ne s'impose pas non plus pour toute analyse. Plus encore, la textométrie n'est pas obligatoirement dépendante de tel ou tel logiciel d'analyse morphosyntaxique (TreeTagger, Cordial Analyseur, etc.) : le chercheur peut préparer son corpus et définir les unités qui l'intéressent avec n'importe quel logiciel, ou l'annoter lui-même, ou même combiner une analyse automatique et une correction ou un enrichissement manuels. Il peut prédéfinir certaines unités qui l'intéressent, et pour le reste laisser l'application de textométrie définir les autres unités[20]. Autrement dit, on observe en pratique certains usages courants pour la définition des unités initiales sur lesquelles vont se baser les analyses textométriques, typiquement des lemmes issus de TreeTagger ; mais en théorie le choix et la définition des unités sont laissés à l'appréciation du chercheur. Il n'y a pas de limitation de l'approche à un modèle linguistique ni d'incompatibilité de principe, dans la mesure où le modèle peut se traduire par des unités séquentielles, éventuellement sur plusieurs niveaux emboîtés.

4.4. Unités et sémantique

Puisque les unités initiales que l'on peut chercher et décompter sont au libre choix du chercheur, une analyse textométrique sémantique pourrait considérer comme unités des informations sémantiques liées aux mots. Une première voie serait un étiquetage codant le concept correspondant, défini dans une ontologie. Cela n'est pas sans soulever des questions théoriques et pratiques. D'un point de vue théorique, seules certaines conceptions de la sémantique, plus référentielles, sont compatibles avec cette modélisation. Et, en pratique, les systèmes d'étiquetage sémantique automatique sont moins avancés, moins mûrs, que les outils d'étiquetage morphosyntaxique. Ce qui est davantage envisageable est l'identification manuelle des sens pour certains mots choisis : pour ces mots, chaque occurrence est dotée d'une étiquette désignant le sens réalisé pour cette occurrence. On peut alors décompter ensemble les occurrences relevant d'un même sens plutôt que d'une même forme graphique, et par exemple comparer la répartition des différents sens d'une même unité graphique dans le corpus.

Ce faisant, on observe l'effet en corpus d'une information codée localement, de façon manuelle. Il peut être plus stimulant de chercher à produire automatiquement, par l'observation du corpus lui-même, une caractérisation sémantique des occurrences dans l'idée de gagner en objectivité (l'observation ne repose pas directement sur l'information qu'on a introduite soi-même dans le corpus) et en dynamisme (le travail manuel d'attribution des étiquettes est essentiellement opéré en amont de l'analyse, dans la préparation du corpus, alors qu'une caractérisation automatique en corpus peut être recalculée à chaque redéfinition du corpus ou s'enrichir au fur et à mesure des analyses). Ainsi, les classes thématiques générées par une démarche ALCESTE peuvent être utilisées pour caractériser sémantiquement les occurrences.

Une autre approche consiste à projeter au niveau des unités lexicales non pas une information référentielle, mais des traits sémantiques, selon une vision de la sémantique qui intègre les effets de diffusion contextuelle et d'isotopie (récurrence d'un trait sémantique dans un texte ou un passage textuel). Cependant, le nombre variable des traits affectés à chaque occurrence appellerait des évolutions du modèle textométrique ; mais aussi, pas plus que l'affectation d'étiquettes conceptuelles, cela ne résout la difficile question de l'adéquation du référentiel sémantique utilisé par rapport au corpus (pertinence, couverture, mise à jour, etc.), ou du moins il faudrait le

20 Toutes ces possibilités ne sont pas toujours implémentées, mais elles sont disponibles, par exemple dans le logiciel TXM (Heiden *et al.*, 2010), logiciel récent pour lequel les modalités d'import de corpus ont été particulièrement développées. La plupart de ces modalités reposent sur l'usage du standard de représentation structurée et d'annotation XML.

considérer aussi comme un point de vue sémantique, contextualisé et non pas universel, partiel et pas toujours ajusté.[21]

Là encore donc, une voie plus naturelle en textométrie consiste à considérer que les unités sémantiques seront principalement construites en corpus à partir de l'observation des voisinages, et que la projection d'informations liées à des ressources sémantiques externes est secondaire (non déterminante, ne devant pas occulter ce qui ressort du corpus) et facultative (s'il est bien conçu, le corpus est déjà susceptible d'apporter de lui-même une information enrichissante). Ainsi, même à partir d'unités non directement codées d'un point de vue sémantique, on peut aller vers des observations et une analyse d'ordre sémantique. Du point de vue des théories linguistiques, l'affinité de la textométrie a ainsi été soulignée avec la sémantique interprétative de François Rastier[22], et en particulier le concept d'isotopie (Mayaffre, 2008 ; Pincemin, 2012). Plutôt que de voir les unités lexicales comme porteuses de sèmes, il s'agit de reconnaître la présence d'un sème dans l'association contextuelle d'unités lexicales, comme le pointe Ludovic Tanguy (1997) qui, dans le contexte de la conception d'une application informatique, perçoit l'importance de cette remarque de Rastier :

> En général, on considère l'isotopie comme une forme remarquable de combinatoire sémique, un effet de la combinaison des sèmes. Ici au contraire, où l'on procède paradoxalement à partir du texte pour aller vers ses éléments, l'isotopie apparaît comme un principe régulateur fondamental. Ce n'est pas la récurrence de sèmes déjà donnés qui constitue l'isotopie, mais à l'inverse la présomption d'isotopie qui permet d'actualiser des sèmes, voire *les* sèmes. (Rastier, 1987, p. 12)

Ainsi, l'information sémantique ne provient pas des mots, mais des contextes. Les observations sémantiques reposent alors sur des associations de plusieurs mots, comme si le sens n'était pas *dans chaque* mot, mais circulait *entre les* mots (Bommier-Pincemin, 1999 ; Pincemin, 1999). La cooccurrence peut alors s'interpréter comme un contexte minimal (Mayaffre, 2008), et constituer une unité sémantique mobilisable dans les calculs (Brunet, 2012).

Sur cette base, les différents traitements textométriques multiplient les accès au sens (voir par exemple les études thématiques réunies dans Brunet, 2016). La concordance, qui présente un relevé de contextes disposé de façon à repérer facilement les ressemblances de contexte d'emploi, apparaît comme un outil fondamental pour l'étude sémantique. C'est non seulement un point d'entrée (par sa simplicité et sa clarté), mais aussi un point de contrôle et d'affinement de l'interprétation des traitements quantitatifs. La cooccurrence peut être vue comme une synthèse statistique de la concordance. Rastier (1995) a expérimenté son usage pour l'analyse thématique : l'interprétation de la liste des cooccurrents (associations quantitatives, au niveau du signifiant) permet de repérer des corrélats sémantiques (associations qualitatives, au niveau du signifié). Cette méthode s'accorde bien avec le caractère contextuel des thématiques et leur manifestation diffuse, parfois partiellement lexicalisée. Les variations de sens d'un mot dans différents contextes (par exemple, pour différentes sources, ou à différentes périodes) peuvent être appréhendées par les variations des ensembles de cooccurrents calculés pour chacun des différents contextes ; ou même les différentes composantes du sens d'un mot dans un même discours (nuances, acceptions) peuvent être appréciées par le repérage de différents sous-ensembles dans un réseau de cooccurrents (Mayaffre, 2008). De façon comparable, les résultats de spécificités se dépouillent en identifiant des groupements de mots spécifiques manifestant un même sème.

Il est possible d'appréhender certains écarts de manifestation entre signifiant et signifié. *L'implicite* est à capter à travers les mots voisins : les calculs peuvent faire ressortir des associations de termes témoignant d'une isotopie, et repérer qu'un *sème* est activé, sans pour autant qu'il y ait nécessairement d'occurrence d'un *mot* formulant directement le sème sous-jacent[23]. *Le non-dit*, l'évitement dans une partie du corpus (typiquement un auteur, un texte, une période par rapport à

21 Par exemple, (Reutenauer, 2012) projette des traits sémantiques à partir d'un dictionnaire (le TLF, *Trésor de la Langue Française*), comme repère pour capter des évolutions de sens.

22 Voir chapitre XXX.

23 Voir l'exemple du thème de *l'ennui* dans *Madame Bovary* présenté par Rastier (2001, p.200-201).

d'autres) peut être pointé par le calcul de spécificités : les *nullax*, mots de fréquence nulle à fort indice de spécificité négatif, sont des mots dont l'absence dans la partie est statistiquement étonnante, compte-tenu de sa fréquence en corpus et de la taille de la partie (Bernard et Bohet, 2017) (voir en exemple la partie basse du tableau 1). Il s'agit alors d'interpréter cette absence : préférence lexicale, évolution des sujets d'intérêt, rejet de conventions, etc. Quant aux *faits singuliers*, signaux faibles au plan quantitatif alors que parfois majeurs au plan sémantique, cas particuliers à la fois isolés et nombreux qui à première vue pourraient échapper à l'approche quantitative de larges corpus, ils peuvent notamment participer à l'analyse en étant regroupés en faisceaux en fonction de leur interprétation, de sorte que l'observation qualitative d'un fait singulier puisse être mise en perspective par des indicateurs quantitatifs rendant compte de contextualisations élargies.

5. Conclusion

Du point de vue des théories sémantiques, il ressort de cette synthèse des affinités multiples entre la textométrie et la sémantique interprétative et différentielle de François Rastier, comme nous l'avons développé ailleurs (Pincemin, 2012). Ici l'objectif est de pointer des propriétés et fonctionnements sémantiques généraux mis en valeur par la textométrie, qui puissent être mis en relation avec les modélisations de différentes théories. Les terrains de rencontre et de connivence potentielle sont multiples : l'importance du niveau textuel ; une approche distributionnelle avec le caractère opératoire des contextes (cotextes) ; la primauté des observations endogènes au corpus et le rôle relatif et secondaire des référentiels non textuels ; la représentation de l'articulation non univoque entre la matérialité du signifiant et la construction de signifiés ; la compatibilité de la subjectivité des interprétations et de l'objectivité de la démarche scientifique. De plus, les avancées technologiques actuelles vont dans le sens d'une articulation plus large et plus complète entre la modélisation sémantique des données et leurs représentation et traitement textométriques : codage riche TEI, caractère unifiant des représentations numériques et corpus multimédias intersémiotiques, maturité des TAL, performances des ordinateurs permettant d'embrasser des données plus vastes et plus précises. Ainsi, théories sémantiques et explorations textométriques peuvent s'éclairer mutuellement, comme ouvrir de nouveaux terrains de recherche, avec un enrichissement réciproque.

Bénédicte PINCEMIN
Université de Lyon, CNRS
Laboratoire IHRIM, UMR 5317

Références

Benzécri Jean-Paul *et al.*, 1973, *L'Analyse des données*, Tome 1 : *La taxinomie*, Tome 2 : *L'analyse des correspondances*, Paris, Dunod.

Benzécri Jean-Paul *et al.*, 1981, *Pratique de l'Analyse des données*, Tome 3 : *Linguistique et lexicologie*, Paris, Dunod.

Bernard Michel, 2000, « Le vocabulaire spécifique d'une œuvre », *Colloque Corpus littéraires – Recueil et numérisation, analyses assistées, didactique*, Université Paris VII, 20-21 octobre 2000. Archives sonores publiées en ligne sur le site *Texto! Textes & Cultures*, http://www.revue-texto.net/Archives/Corpus_litteraires/Corpus_litteraires.html

Bernard Michel et Bohet Baptiste, 2017, *Littérométrie. Outils numériques pour l'analyse des textes littéraires*, Paris, Presses de la Sorbonne nouvelle.

Bommier-Pincemin Bénédicte, 1999, *Diffusion ciblée automatique d'informations : conception et mise en œuvre d'une linguistique textuelle pour la caractérisation des destinataires et des documents*, Thèse de Doctorat en Linguistique, Université de Paris IV, dir. François Rastier, 6 avril 1999.

Bourion-Jacquemin Évelyne, 2001, *L'aide à l'interprétation des textes électroniques*, Thèse de Doctorat en Sciences du langage, Université de Nancy II, dir. François Rastier, 14 décembre 2001.

Brunet Étienne, 2000, « Qui lemmatise dilemme attise », Actes des 11e rencontres linguistiques en pays rhénan, *Scolia*, n°13, p. 7-32. Réédité dans Brunet (2011), chapitre 9, p. 165-184.

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

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

Brunet Étienne, 2012, « Nouveau traitement des cooccurrences dans Hyperbase », *Corpus*, n°11, p. 219-246. Réédité dans Brunet (2016), chapitre 17, p. 287-311.

Brunet Étienne, 2016, *Tous comptes faits. Écrits choisis, tome III : Questions linguistiques*, Pincemin Bénédicte (éd.), Paris, Champion.

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

Gries Stefan Th., 2014, « Quantitative corpus approaches to linguistic analysis: seven or eight levels of resolution and the lessons they teach us », *Developments in English: expanding electronic evidence*, Taavitsainen Irma, Kytö Merja, Claridge Claudia & Smith Jeremy (eds.), Cambridge University Press, p. 29-47.

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

Heiden Serge, Magué Jean-Philippe, Pincemin Bénédicte, 2010, « TXM : Une plateforme logicielle open-source pour la textométrie – conception et développement », *Statistical Analysis of Textual Data. Proc. of JADT 2010*, Bolasco Sergio, Chiari Isabella, Giuliano Luca (eds), Rome, Edizioni Universitarie di Lettere Economia Diritto, p. 1021-1032.

Lafon Pierre, 1980, « Sur la variabilité de la fréquence des formes dans un corpus », *MOTS*, n°1, p. 127-165.

Lafon Pierre, 1981, « Analyse lexicométrique et recherche des cooccurrences », *MOTS*, n°3, p. 95-148.

Lafon Pierre, Salem André, 1983, « L'inventaire des segments répétés d'un texte », MOTS, n°6, p. 161-177.

Lebart Ludovic, Salem André, 1988, *Analyse statistique des données textuelles*, Paris, Dunod.

Lebart Ludovic, Salem André, 1994, *Statistique textuelle*, Paris, Dunod.

Lebart Ludovic, Salem André, Berry Lisette, 1998, *Exploring Textual Data*, Dordrecht, Kuwer Academic Publishers.

Luong Nhuan Xuan, 1988, *Méthodes d'analyse arborée. Algorithmes. Applications,* Thèse de Doctorat en Sciences appliquées, Université de Paris V, dir. Jean-Pierre Barthélémy.

Mayaffre Damon, 2008, « De l'occurrence à l'isotopie. Les co-occurrences en lexicométrie », Textes, documents numériques, corpus. Pour une science des textes instrumentée, Mathieu Valette (dir.), *Syntaxe & Sémantique*, n°9, p. 53-72.

Muller Charles, [1973] 1993, *Initiation aux méthodes de la statistique linguistique*, Paris, Champion.

Muller Charles, [1977] 1993, *Principes et méthodes de la statistique lexicale*, Paris, Champion.

Muller Charles, 1984, « De la lemmatisation », Préface à Lafon Pierre, *Dépouillements et statistiques en lexicométrie*, Paris, Slatkine-Champion, p. I-XII.

Péry-Woodley Marie-Paule, 1995, « Quels corpus pour quels traitements automatiques ? », *Traitement Automatique des Langues*, 36 (1-2), p. 213-232.

Pincemin Bénédicte, 1999, « Sémantique interprétative et analyses automatiques de textes : que deviennent les sèmes ? », Dépasser les sens iniques dans l'accès automatisé aux textes, Benoît Habert (dir.), *Sémiotiques*, n°17, p. 71-120.

Pincemin Bénédicte, 2006, « Concordances et concordanciers - De l'art du bon KWAC », *Corpus en Lettres et Sciences sociales. Des documents numériques à l'interprétation. Actes du XXVIIe Colloque d'Albi Langages et Signification*, Rastier François et Ballabriga Michel (dir.), Duteil-Mougel Carine et Foulquié Baptiste (éds), Presses universitaires de Toulouse. Actes également publiés en ligne sur le site *Texto! Textes & Cultures*, http://www.revue-texto.net/1996-2007/Parutions/Livres-E/Albi-2006/Sommaire.html.

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

Rastier François, 1987, *Sémantique interprétative*, Paris, Presses universitaires de France.

Rastier François (dir.), 1995, *L'analyse thématique des données textuelles. L'exemple des sentiments*, Martin Éveline (éd.), collection Études de sémantique lexicale, Paris, Didier Érudition.

Rastier François, 2001, *Arts et sciences du texte*, Paris, Presses universitaires de France.

Rastier François, 2011, *La mesure et le grain. Sémantique de corpus*, Paris, Champion.

Ratinaud Pierre, Déjean Sébastien, 2009, IraMuTeQ : implémentation de la méthode ALCESTE d'analyse de texte dans un logiciel libre, *Colloque Modélisation Appliquée aux Sciences Humaines et Sociales (MASHS2009)*, Toulouse, http://repere.no-ip.org/Members/pratinaud/mes-documents/articles-et-presentations/presentation_mashs2009.pdf/view

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

Reutenauer Coralie, 2012, *Vers un traitement automatique de la néosémie : approche textuelle et statistique*, Thèse de Doctorat en Sciences du langage de l'Université de Lorraine, dir. Jean-Marie Pierrel, Evelyne Jacquey et Mathieu Valette, 20 janvier 2012.

Salem André, 1987, *Pratique des Segments Répétés. Essai de statistique textuelle*, Paris, Klincksieck.

Söze-Duval Keyser, 2008, « Pour une textométrie opérationnelle », Inédit, http://www.tal.univ-paris3.fr/trameur/RTI6provisoire.doc

Tanguy Ludovic, 1997, *Traitement automatique de la langue naturelle et Interprétation : Contribution à l'élaboration d'un modèle informatique de la Sémantique Interprétative*, Thèse de Doctorat en Informatique de l'Université de Rennes 1, dir. Jean-Pierre Barthélémy et Ioannis Kanellos, 7 mai 1997.

Tournier Maurice, 1985, « Sur quoi pouvons-nous compter ? », *Verbum*, vol. 8, Hommage à Hélène Nais, p. 481-492.