Travaux neuchâtelois de linguistique, 2011, 55, 77-93

Analyse textométrique et interprétation littéraire – *Hyperbase*, Rousseau et les Lumières

Véronique MAGRI-MOURGUES
Université de Nice Sophia Antipolis, UMR 7320: *Bases, Corpus et Langage*

The purpose of this paper is to examine under which conditions we can establish the value and relevance of quantitative analyses, with a view to interpreting literary textual data. The quantitative analysis was carried out on the empirical corpus of the works of Rousseau, within the group of The Enlightenment Philosophers. The hypertextual software Hyperbase was used for this purpose. The output of the software supplies a tabular presentation of the corpus deconstructed into paradigmatic groups and produces a reticular display by projecting associative networks onto the page. The hermeneutic approach, which leads from quantitative to qualitative data, relies on linking and differentiation operations, which are liable to bring to light significant contrasts. A central issue of this study is to test the functions of the software allowing the extraction of cooccurrences. These functions are assessed at different levels of increasing size fitted into each other, from the sentence to the whole corpus, relying on an automatic contextualisation of the data.

Introduction

L'analyse des données textuelles s'est affirmée depuis plusieurs années maintenant comme l'alliée de la linguistique textuelle. Elle a assis son autorité heuristique grâce à ses compétences dans le traitement de très grands corpus. Le traitement informatisé des données textuelles, dont les enjeux sont résumés dans cette récente dénomination disciplinaire qu'est la textométrie, permet de soumettre de très grands corpus à un questionnement stable, à défaut de pouvoir toutefois se revendiquer comme objectif; cette appellation fait suite à celle de lexicométrie comme étude du vocabulaire ou encore à celle de statistique linguistique ou de linguistique quantitative, selon que l'on choisisse de mettre l'accent sur les outils d'analyse ou sur le type de corpus étudié. Le terme de stylométrie, quant à lui, précède les travaux de linguistique quantitative puisqu'il est attesté dès la fin du XIX^{ème} siècle, avec le sens vague d'étude du style d'une œuvre. Il est tentant de le réinvestir à la lumière du numérique et de le redéfinir comme possible mesure du style. Que faut-il entendre par là?[1] Le style serait-il quantifiable? Cette formulation n'est acceptable que comme un raccourci. Ce qui peut être mesuré, autrement dit, dénombré, ce sont toujours des unités concrètes, discrètes et observables, de sorte qu'un

[1] Voir par exemple Juhan Tuldava: "Stylistics, author identification", 368-387 et Jadwiga Sambor, Adam Pawłowski: "Quantitative linguistics in Poland", 115-129. In: Reinhard Köhler, Gabriel Altmann, Rajmund G. Piotrowski (eds.) (2005): Quantitative Linguistik - Quantitative Linguistics. Ein internationales Handbuch. Berlin/New York (De Gruyter).

repérage automatique est envisageable. Se pose alors la question de la démarche herméneutique qui autorise le passage des observations chiffrées à leur interprétation qualitative, orientée vers le dévoilement du sens du texte. Comment donner valeur et pertinence aux dénombrements pour opérer ce parcours interprétatif depuis le quantitatif jusqu'au qualitatif?

Ce questionnement épistémologique, qui porte sur le fondement même de la discipline qu'est la textométrie, sera ici posé au travers d'un outil d'analyse hypertextuelle particulier, *Hyperbase*, développé au sein de l'UMR 7320, *Bases, Corpus et Langage*[2]*;* certaines fonctions en seront davantage exploitées et appliquées à un corpus numérique d'auteur, Rousseau, et, plus généralement, aux philosophes des Lumières. Des bases de données ont en effet été constituées, consacrées à chacun des philosophes du XVIII[ème] siècle, Montesquieu, Diderot, Voltaire, Rousseau, et regroupant la plupart de leurs œuvres[3].

Une analyse, pour autant qu'elle vise le sens du corpus textuel, impose au chercheur une réflexion sur les données qu'il construit, sur le traitement qu'il leur fait subir et sur les conclusions qu'il peut tirer des résultats obtenus. La particularité de la textométrie est de convertir, pour ainsi dire, les données textuelles en résultats chiffrés, pariant sur le potentiel interprétatif des dénombrements.

La validité de ce processus ne peut toutefois être assurée que si les dénombrements sont appréciés de manière relative et contrastive. Ce n'est que par rapprochement et différenciation que les résultats obtenus sont susceptibles d'être promus éléments significatifs.

Les analyses fondées sur les fréquences des variables décomposent le corpus textuel en ensembles paradigmatiques d'unités de même nature (formes, lemmes, structures syntaxiques, codes grammaticaux); pour retrouver une lecture linéaire, il faut associer à l'évaluation quantitative

---

2 http://www.unice.fr/bcl/spip.php?rubrique38. Possibilité est donnée de télécharger une version d'évaluation du logiciel.

3 La base *Rousseau* qui sert de corpus d'étude à mes observations comporte trente-cinq fragments, soit dix-huit œuvres segmentées le cas échéant en raison de leur longueur: *Le Devin du village* (1752), *Discours sur les sciences et les arts* (1750), *Discours sur l'origine et les fondements de l'inégalité parmi les hommes* (1755), *Discours sur l'économie politique* (1755), *Lettre à d'Alembert sur les spectacles* (1758), *Julie ou La Nouvelle Héloïse* (1761), *Émile ou De l'éducation* (1762), *Lettres à M. de Malesherbes (*1762), *Du Contrat social ou principes du droit politique* (1762), *J.-J. Rousseau, citoyen de Genève, À Christophe de Beaumont, archevêque de Paris* (1762-1763), *Lettres écrites de la montagne* (1764), *Les Confessions* (1765-1770), *Extrait du Projet de paix perpétuelle de Monsieur l'Abbé de Saint-Pierre* (1761), *Polysynodie de l'abbé de Saint-Pierre* (1756-1782), *Les Dialogues ou Rousseau juge de Jean-Jacques* (1772-1775), *Considérations sur le gouvernement de Pologne* (1772), *Les Rêveries du promeneur solitaire* (1776-1778), *Projet de constitution pour la Corse* (1765). L'ordre des titres de ces œuvres suit l'ordre imposé par la structure de la base de données à notre disposition. Pour les autres bases sollicitées dans cette étude, se référer à É. Brunet (2009).

des unités étudiées une étude de leur distribution dans le déroulement phrastique et séquentiel du texte, autrement dit recontextualiser les résultats au sein de la phrase ou d'une séquence plus grande comme le paragraphe. Cette contextualisation intratextuelle peut s'élargir à une contextualisation intertextuelle par confrontation entre plusieurs textes d'auteurs différents par exemple. On pourra alors s'interroger sur le bénéfice interprétatif à attendre d'une telle mise en contraste.

1. Rapprochement et différenciation

1.1 Le dénombrement des variables

Les variables qui peuvent être repérées et dénombrées par le logiciel *Hyperbase* sont de deux ordres, lexical et grammatical. Le travail de préparation du corpus textuel par ce logiciel consiste en premier lieu à transformer le corpus en un ensemble de paradigmes exploitables. Les lexèmes constituent un premier ensemble, qui répertorie tous les mots du corpus textuel, autrement dit, toutes les chaînes graphiques séparées les unes des autres par un séparateur, comme un blanc ou un signe de ponctuation, selon les normes implémentées dans le logiciel. Le dénombrement des lexèmes et leur regroupement en champs lexicaux, voire sémantiques par le chercheur permettent d'esquisser une approche thématique du corpus.

Depuis quelques années déjà, le logiciel intègre un analyseur, *Treetagger*[4], qui procède à l'étiquetage morpho-syntaxique des mots d'un texte, autrement dit qui fournit pour chaque mot la graphie, le lemme de rattachement – qui correspond à l'entrée du dictionnaire – et le codage grammatical qui identifie la nature du mot et lui associe de fines indications dépendantes de la partie de discours à laquelle il appartient: des indications de genre et de nombre pour les substantifs et les adjectifs; une identification de tous les paramètres véhiculés par la forme verbale pour autant qu'ils fassent l'objet d'un marquage en langue: le mode, le tiroir verbal, la personne sont repérables de manière automatique; ce n'est évidemment pas le cas de l'aspect.

Outre le paradigme lexical constitué par l'ensemble des lexèmes d'un corpus, se constituent ainsi des paradigmes grammaticaux organisés autour d'un tiroir verbal ou d'une personne par exemple.

Le corpus textuel se réalise comme corpus paradigmatique, privilégiant soit le versant lexical du texte, soit le versant morpho-syntaxique.

Dans ces ensembles de paradigmes qui dénombrent des unités lexicales ou morpho-syntaxiques et qui sont à solliciter comme nouveaux corpus

[4] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/.

d'études, construits et orientés en fonction d'une recherche précise, reste à décider ce qui peut être déduit comme remarquable et qui peut assurer le passage du donné strictement quantitatif à l'évaluation qualitative. Les fréquences doivent pouvoir être interprétées comme récurrences non aléatoires et convergentes.

1.2 Fréquences, récurrences et convergences

La pertinence des dénombrements ne peut s'établir que par rapport à une norme qui sert d'étalon aux mesures. Seule une démarche différentielle qui fait contraster des corpus peut faire d'une récurrence observée une itération au sens de François Rastier[5], autrement dit une récurrence signifiante parce que non aléatoire. Les observations statistiques n'ont de pertinence que par l'effet de saillance obtenu par des mesures contrastives.

Récurrence et variation sont les paramètres qui règlent l'interprétation, qui permettent d'apprécier la valeur de telle fréquence. Une stratégie contrastive et différentielle associe deux corpus, l'un qui est le corpus d'étude dont on veut observer les fréquences remarquables, l'autre le corpus de référence. La théorie de l'écart stylistique retrouve ici une légitimité, mais la variation s'apprécie non pas en regard d'un usage standard, comme un hypothétique degré zéro stable, mais par rapport à un ensemble englobant, le corpus de référence. Les deux corpus sont construits selon les objectifs de la recherche: on pourra confronter des corpus d'écrivains d'une même tranche temporelle, si on veut mettre en évidence des particularités d'écritures individuelles, ou bien des ensembles génériques distincts si la recherche a pour objectif la caractérisation d'un genre littéraire par exemple.

Par cercles concentriques, l'objet d'étude peut être, pour ainsi dire, monologique lorsqu'il s'agit de l'œuvre entier d'un écrivain, ou simplement d'un fragment de cet œuvre; il peut être transversal et regrouper des œuvres censées illustrer un même genre littéraire ou encore une période de l'histoire littéraire. Corollairement, le corpus de référence varie en proportion à l'extension du corpus d'étude choisi: si on cherche à caractériser l'écriture d'un écrivain, on choisira l'ensemble englobant des œuvres des contemporains; si c'est une caractérisation générique qui intéresse le chercheur, on pourra faire contraster le corpus d'étude avec un autre genre connexe.

Le calcul de l'écart réduit[6] est le fondement de nombreuses évaluations quantitatives et différentielles dans *Hyperbase*, permettant l'appréciation de l'écart entre une fréquence théorique et une fréquence observée dans

5 Voir F. Rastier (1987: 93).

6 Voir Ch. Muller (1979).

un corpus donné. Une fonction dans *Hyperbase* permet de calculer cette variation entre les mots d'un corpus textuel donné et une tranche chronologique contemporaine extraite de la base *Frantext*. On obtient ainsi ce qu'il est convenu d'appeler les spécificités lexicales[7] d'un corpus textuel. Pour la base *Rousseau*, voici un extrait de la liste que l'on obtient (Figure 1). La lecture du corpus textuel se fait tabulaire: les fragments du corpus qui correspondent à chaque œuvre qui le compose sont croisés sur un tableau à deux entrées. Cette liste fournit, pour chaque forme retenue en raison de sa représentativité dans le corpus étudié, sa fréquence dans le corpus de référence externe, *Frantext*, sa fréquence dans le corpus d'étude. Le premier chiffre dans la première colonne donne le résultat du calcul de l'écart réduit, autrement dit l'évaluation de l'emploi qui est fait par Rousseau de telle forme, comparativement à celui de *Frantext*, posé comme représentant une référence fiable de l'emploi des contemporains de l'écrivain. Dans la seconde colonne s'inscrit la fréquence de la forme recherchée dans le corpus *Frantext*; enfin la troisième colonne fournit la fréquence de la forme dans le corpus d'étude, ici la base *Rousseau*. Les premières places sont occupées logiquement par les noms propres tandis que d'autres substantifs par exemple se distinguent comme reflet de l'imaginaire et de la pensée rousseauistes, tels en tout cas qu'ils trouvent leur écho dans les œuvres choisies.

[7] Le calcul des spécificités lexicales externes s'appuie sur une confrontation entre le corpus d'étude et une tranche contemporaine de la base *Frantext*. Dans ce cas, c'est le calcul de la loi normale qui est utilisé: toute fréquence (ou effectif) observée est comparée à la fréquence attendue et convertie en écart réduit selon la formule:

écart réduit (z) = $\frac{k - fp}{\sqrt[2]{fpq}}$

k = fréquence observée dans le texte,

f = fréquence observée dans le corpus

p = étendue du texte *(t)* / étendue du corpus *(T)*

$q = 1 - p$

Voir *Hyperbase*. Manuel de référence (2011: 37-41).

Écart réduit	Fréquence dans *Frantext*	Fréquence dans le corpus d'étude	forme
434.70	13	335	Émile
138.60	129	342	État
71.43	219	236	Wolmar
64.95	139	170	Édouard
63.39	1948	690	Julie
60.35	451	295	assés
55.08	296273	20558	qu'
46.54	418	223	note
44.20	589	257	Genève
40.44	191777	12945	n'
40.07	3187	635	gouvernement
39.54	3109	619	lois
38.37	690	248	sitôt
37.93	266246	16902	ne
37.68	24685	2443	ni
37.64	777	261	maximes
37.47	912	285	autrui
36.95	1323	350	durant
34.66	13758	1525	état
34.39	148871	9947	on
33.90	31571	2794	moins
33.34	724	226	Rousseau
33.25	28413	2554	toujours
32.61	170827	11041	plus
32.19	60192	4562	leur
31.89	32622	2788	jamais
31.49	69167	5073	sans
30.37	2585	453	volonté
29.34	72656	5160	ils
28.98	20025	1832	celui
28.97	611	182	longtemps
28.76	11518	1209	avais
28.40	14245	1404	eux
27.20	53189	3881	point
26.08	8896	950	mêmes
26.04	35295	2731	autre
26.00	204667	12293	pour
25.94	1031	227	constitution
25.85	6001	714	droit
25.79	20177	1746	hommes
25.54	309017	17800	en
25.50	33040	2568	quand
25.41	1899	327	citoyens
25.37	3399	478	sentir
25.35	1663	300	nul
24.44	13087	1222	doit
24.32	478189	26445	l'
24.29	4204	537	auteur
24.02	2717	397	autorité
23.72	10163	996	force
23.67	4561	559	enfants
23.54	24700	1970	autres
23.45	13119	1200	mal
23.34	35135	2613	rien
23.31	8542	868	besoin
23.31	6137	683	savoir
23.16	18103	1530	ceux
22.85	71425	4717	être
22.41	3041	409	selon
22.35	96461	6092	y
22.19	420686	23180	les
22.18	4159	504	contraire
21.88	12435	1115	raison
21.81	1319	232	membres
21.71	12656	1126	nature
21.34	40129	2830	homme
21.21	1103	203	particuliers
21.16	6241	655	propre
21.02	667894	35591	et
20.96	1058	196	vrais
20.87	51479	3471	faire
20.79	14737	1243	mieux
20.44	42081	2906	peut
20.40	1217	210	préjugés
20.34	16190	1326	seul
20.27	15052	1250	celle
20.20	2207	308	devoirs
20.10	853	166	magistrats
19.27	2178	296	humain
19.18	53642	3512	sont
19.04	487197	26123	que
18.95	8042	747	corps
18.92	1570	235	élève
18.88	2858	352	naturel
18.85	2136	288	guère
18.64	4540	485	loi
18.51	1012	174	citoyen
18.41	1889	261	jugement
18.34	5405	546	goût
18.33	4847	504	chacun
18.26	97256	5864	tout
18.09	8716	777	peuple
18.03	3304	379	conseil
18.02	8620	769	étais
17.90	2694	327	juger
17.86	11390	951	comment
17.75	2169	280	musique
17.74	5945	576	objet
17.54	943	160	abus
17.51	3839	415	passions
17.49	1313	198	vices
17.48	10257	868	sens
17.18	1217	186	système
17.15	4293	445	soins
17.09	7962	706	enfant
17.07	1336	197	établir
17.02	427582	22813	il
16.97	5192	509	vivre
16.94	2032	260	publique
16.76	1289	190	ôter
16.70	5680	540	sait
16.66	18916	1393	soit
16.61	24959	1753	faut
16.59	1190	179	leçons
16.44	3221	353	vouloir
16.34	12258	972	gens
16.33	2638	305	montrer
16.33	1169	175	maman
16.16	2049	254	éducation
16.01	119318	6883	par
15.95	2358	278	agit
15.93	258807	14100	des
15.93	5686	528	intérêt
15.91	3252	349	objets
15.89	11653	923	reste

15.84	7721	666	aucun
15.83	1762	226	souverain
15.82	1228	177	simplicité
15.74	1111	165	rapports
15.73	1288	182	écrits
15.65	5749	528	étant
15.65	1138	167	savent
15.63	13842	1054	eût
15.58	1440	195	forcé

15.57	4173	414	espèce
15.56	2408	278	connaître
15.51	2451	281	soi
15.51	1546	204	générale
15.31	7420	636	font
15.31	1992	241	naturelle
15.30	1081	159	songer
15.29	282519	15238	est

Fig. 1 – Extrait des spécificités lexicales dans la base de données *Rousseau*[8].

Aux critères de la répétition et des variations significatives évaluées par rapport à un corpus posé comme norme, s'ajoute celui de la sériation des résultats, pour compléter le processus interprétatif. Les mots se regroupent par réseaux lexicaux et développent telle sphère lexicale, voire sémantique, à même de caractériser la pensée d'un écrivain, par contraste avec un corpus externe.

La distribution d'une variable quelconque dans un segment textuel, autrement dit l'observation de son voisinage, est un autre paramètre à considérer pour mener une interprétation où se croisent analyse quantitative et évaluation qualitative. En accord avec la théorie contextuelle de la signification des linguistes anglo-saxons, selon laquelle les emplois d'un mot permettent d'en appréhender le sens[9], il s'agit de dépasser l'étape des dénombrements et de l'analyse des fréquences par un travail de recontextualisation des résultats. Les fréquences proposent une lecture paradigmatique du corpus textuel; les fonctionnalités du logiciel qui autorisent le traitement non plus seulement des fréquences, mais des unités recontextualisées, rendent au corpus textuel une part de sa linéarité.

2. Contextualisation intratextuelle

> L'activité interprétative procède principalement par contextualisation. Elle rapporte le passage considéré, si bref soit-il (ce peut être un mot), à son voisinage, selon des zones de localité (syntagme, période) de taille croissante; à d'autres passages du même texte, convoqués par des procédures d'assimilation ou de contraste; enfin, à

8 Pour le logiciel, un mot correspond à une suite de caractères délimitée par des signes séparateurs, comme le blanc ou un signe de ponctuation: pour cette raison, les formes élidées comme "n'" ou "l'" sont séparées des formes pleines correspondantes ″ne″ ou ″le″ même si ces formes, élidées ou pleines, ne constituent qu'une unité lexicale en langue.

Le logiciel a fait la différence entre "état" avec majuscule et sans, ce qui justifie deux entrées distinctes dans le tableau pour ce lexème. L'acception du terme varie d'ailleurs selon l'emploi de la majuscule ou non.

9 Voir la différence établie par F. Rastier (juin-septembre 2003) entre sens et signification. Le sens désigne les acceptions ou les emplois en contexte du mot tandis que la signification renvoie au contenu supposé invariant du mot.

> d'autres passages d'autres textes, choisis dans le corpus de référence, et qui entrent dans le corpus de travail[10].

Le traitement des textes par *Hyperbase* problématise la notion de contexte, envisagée ici dans son acception linguistique. L'environnement contextuel se définit à différents paliers, de taille croissante. Sur le plan intratextuel, les unités contextualisées se moulent dans les patrons syntaxiques du syntagme, de la phrase – clairement délimitée par une ponctuation forte – ou se modulent selon des unités de plus grande échelle, comme le paragraphe, voire le texte tout entier.

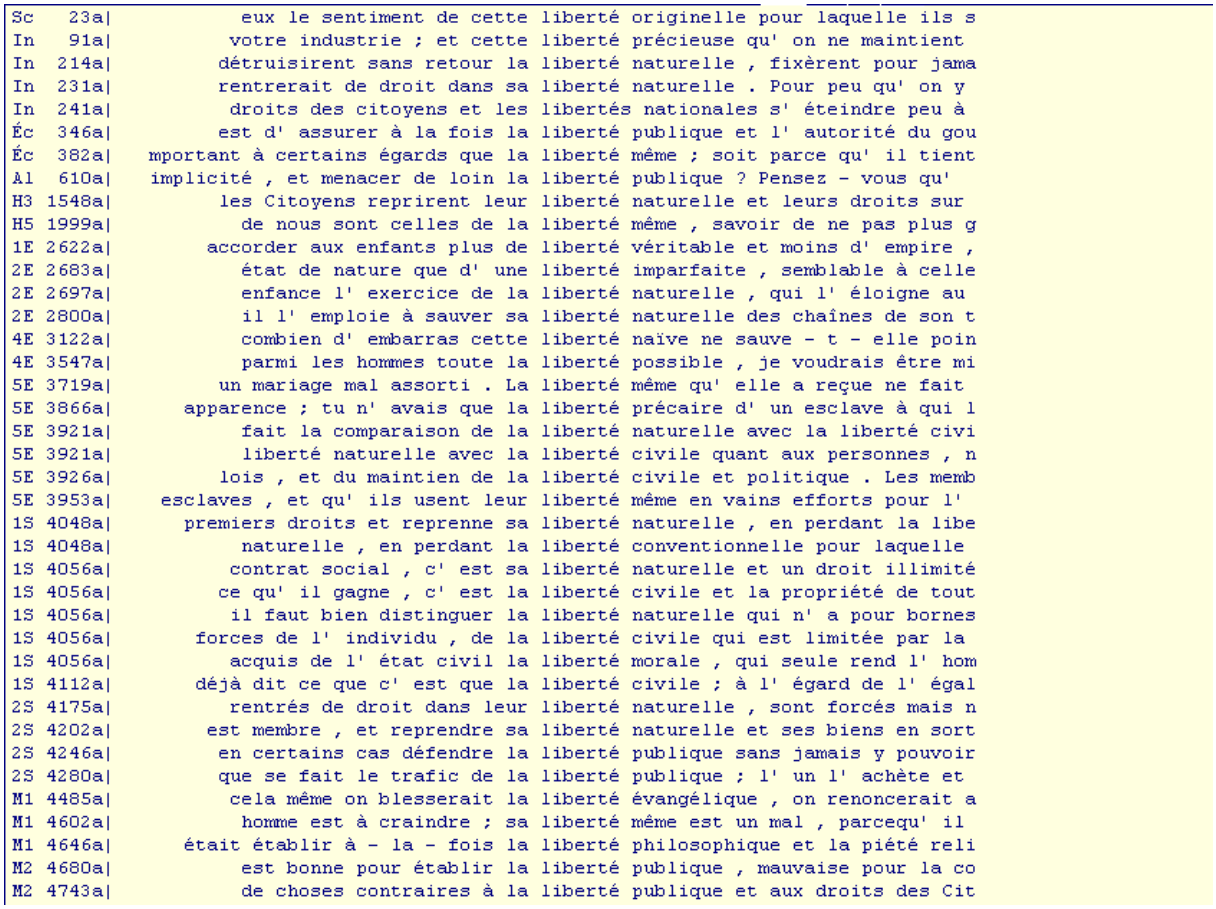
2.1 La mise en contexte phrastique: les cooccurrences

L'analyse des cooccurrences qui se réalisent au niveau du syntagme se maintient au niveau de la phrase. Les cooccurrences s'exercent entre deux unités linguistiques contiguës. Le voisinage étroit, situé à gauche ou à droite d'un mot choisi pour pôle, peut être étudié sous différentes perspectives complémentaires. Aux contraintes de la syntaxe peut se surimposer la force de l'usage. En effet, le chercheur doit composer avec les notions de structures lexicalisées ou phraséologiques, de moindre intérêt pour le rendement interprétatif qu'une association lexicale inédite, si on veut caractériser l'originalité d'une écriture.

Les concordanciers[11] proposent une liste de contextes étroits organisés autour du mot-pôle. La figure suivante propose les concordances du lemme *liberté* caractérisé par un adjectif qualificatif à sa droite dans le corpus *Rousseau.*

[10] F. Rastier (2001: 92).

[11] Cf. *The KWIC concordance* (keyword in context – mot-clé en contexte).



```
Sc   23a|           eux le sentiment de cette liberté originelle pour laquelle ils s
In   91a|          votre industrie ; et cette liberté précieuse qu' on ne maintient
In  214a|         détruisirent sans retour la liberté naturelle , fixèrent pour jama
In  231a|         rentrerait de droit dans sa liberté naturelle . Pour peu qu' on y
In  241a|          droits des citoyens et les libertés nationales s' éteindre peu à
Éc  346a|         est d' assurer à la fois la liberté publique et l' autorité du gou
Éc  382a|   mportant à certains égards que la liberté même ; soit parce qu' il tient
Al  610a|   implicité , et menacer de loin la liberté publique ? Pensez - vous qu'
H3 1548a|         les Citoyens reprirent leur liberté naturelle et leurs droits sur
H5 1999a|           de nous sont celles de la liberté même , savoir de ne pas plus g
1E 2622a|        accorder aux enfants plus de liberté véritable et moins d' empire ,
2E 2683a|           état de nature que d' une liberté imparfaite , semblable à celle
2E 2697a|           enfance l' exercice de la liberté naturelle , qui l' éloigne au
2E 2800a|           il l' emploie à sauver sa liberté naturelle des chaînes de son t
4E 3122a|           combien d' embarras cette liberté naïve ne sauve - t - elle poin
4E 3547a|           parmi les hommes toute la liberté possible , je voudrais être mi
5E 3719a|         un mariage mal assorti . La liberté même qu' elle a reçue ne fait
5E 3866a|      apparence ; tu n' avais que la liberté précaire d' un esclave à qui l
5E 3921a|           fait la comparaison de la liberté naturelle avec la liberté civi
5E 3921a|           liberté naturelle avec la liberté civile quant aux personnes , n
5E 3926a|         lois , et du maintien de la liberté civile et politique . Les memb
5E 3953a|    esclaves , et qu' ils usent leur liberté même en vains efforts pour l'
1S 4048a|      premiers droits et reprenne sa liberté naturelle , en perdant la libe
1S 4048a|           naturelle , en perdant la liberté conventionnelle pour laquelle
1S 4056a|          contrat social , c' est sa liberté naturelle et un droit illimité
1S 4056a|         ce qu' il gagne , c' est la liberté civile et la propriété de tout
1S 4056a|          il faut bien distinguer la liberté naturelle qui n' a pour bornes
1S 4056a|       forces de l' individu , de la liberté civile qui est limitée par la
1S 4056a|          acquis de l' état civil la liberté morale , qui seule rend l' hom
1S 4112a|       déjà dit ce que c' est que la liberté civile ; à l' égard de l' égal
2S 4175a|          rentrés de droit dans leur liberté naturelle , sont forcés mais n
2S 4202a|        est membre , et reprendre sa liberté naturelle et ses biens en sort
2S 4246a|         en certains cas défendre la liberté publique sans jamais y pouvoir
2S 4280a|         que se fait le trafic de la liberté publique ; l' un l' achète et
M1 4485a|          cela même on blesserait la liberté évangélique , on renoncerait a
M1 4602a|           homme est à craindre ; sa liberté même est un mal , parcequ' il
M1 4646a|       était établir à - la - fois la liberté philosophique et la piété reli
M2 4680a|           est bonne pour établir la liberté publique , mauvaise pour la co
M2 4743a|           de choses contraires à la liberté publique et aux droits des Cit
```

Fig. 2 – Concordancier du mot *liberté* + adjectif dans la base de données *Rousseau*.

La figure permet une visualisation verticale des occurrences de ce mot dans une zone de localité étroite. Les éléments en co-présence, substantif et adjectif, interagissent; le sens des mots est modulé selon l'environnement linguistique. On observe ici l'emploi minoritaire du substantif au pluriel, une seule fois avec l'adjectif *nationales*, tandis que l'emploi au singulier du mot *liberté* se décline selon ses applications dans différents domaines, associé à des adjectifs relationnels: *la liberté civile* s'oppose ainsi à *la liberté naturelle* par exemple, en accord avec les convictions de Rousseau. Les adjectifs qualitatifs sont moins nombreux: la liberté est *précieuse, véritable, imparfaite, précaire* ou *honnête* entre autres. Cet exemple montre le rendement interprétatif qui peut être obtenu d'une simple mise en évidence des cooccurrents syntagmatiques, d'une part sur le potentiel sémantique d'un substantif et sur son éventuelle polysémie levée par la contextualisation, d'autre part sur l'imaginaire d'un écrivain qui s'esquisse au travers de structures syntagmatiques relativement privilégiées. Ainsi pourrait être amorcée une étude qualitative du *profil cooccurrentiel* d'un corpus.

2.2 La mise en contexte réflexive

Le niveau phrastique peut être dépassé pour atteindre le palier du texte. Des fonctions implémentées dans *Hyperbase* permettent le calcul des cooccurrents les plus fréquents: la fonction *Thème* et celle qui calcule les corrélats lexicaux ou associations privilégiées. Toutes deux ont pour finalité de construire des réseaux associatifs organisés autour de mots-pôles. La démarche heuristique est cependant différente[12].

La fonction *Thème* observe les fréquences de tous les mots présents dans le voisinage d'une forme choisie comme mot-pôle. Le voisinage correspond en fait à un paragraphe ou à une dizaine de lignes délimitées par le logiciel, quand le découpage en paragraphes n'existe pas dans le corpus textuel ou, au contraire, lorsqu'un paragraphe est trop long pour être exploité. L'ensemble des séquences ainsi constituées compose un nouveau sous-corpus où sont appréciées les spécificités d'emploi de chaque mot par référence au dictionnaire de l'œuvre qui en recense tous les mots. Les fréquences observées dans ce sous-ensemble sont comparées aux fréquences de ces corrélats dans le corpus tout entier. La norme adoptée est donc endogène et ce calcul permet de mettre en évidence les associations lexicales qui s'organisent autour d'un mot, ce qu'il est convenu d'appeler un *thème*.

L'établissement des corrélats lexicaux repose sur le calcul des liens préférentiels qui se tissent entre les quatre-cents mots - substantifs et adjectifs - les plus fréquents du corpus. Les réseaux cooccurrentiels ainsi mis en évidence sur des schémas peuvent être interprétés comme des réseaux thématiques, pour autant qu'on admette que la contiguïté dans l'espace du texte, matérialisée plus clairement sur un schéma-plan par la formation de constellations lexicales, puisse être traductible en termes de proximité sémantique. Ces corrélats sémantiques peuvent enfin être considérés comme lexicalisation partielle d'un thème[13].

Pour le corpus *Rousseau*, on pourrait prendre l'exemple du mot *autorité*, proposé par certains spécialistes de l'écrivain comme mot-clé de son vocabulaire[14]. Le programme élimine les mots-outils de la liste des spécificités obtenue, réalise le tableau général des cooccurrences puis calcule et trie les indices qui permettent d'apprécier la distance entre les mots de la liste pris deux à deux[15].

---

12 Pour des explications plus précises sur la procédure suivie, voir É. Brunet (2007).

13 Voir F. Rastier (1996).

14 Voir par exemple Léo et Michel Launay qui, les premiers ont amorcé une analyse structurale du vocabulaire de Rousseau à l'aide d'outils statistiques.

15 *Hyperbase*. Manuel de référence (2011: 98).

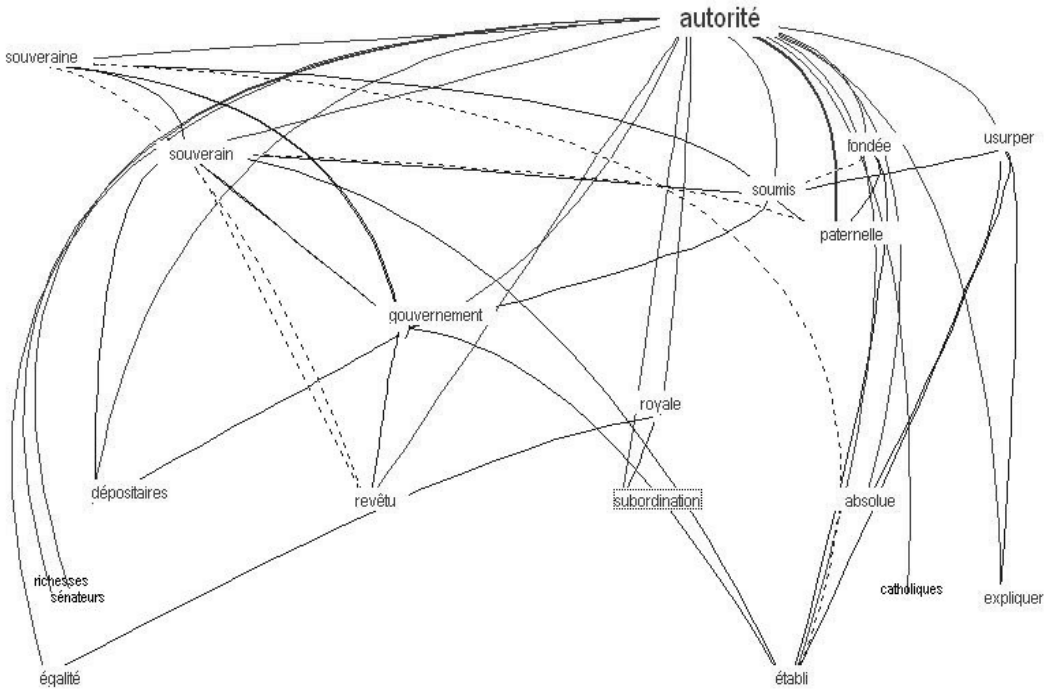


Fig. 3 – Cooccurrents privilégiés du mot *autorité* dans la base de données *Rousseau*. Fonction *Thème*[16]

Ce graphe présente les cooccurrents privilégiées du mot *autorité* – substantifs ou adjectifs – dans le contexte du paragraphe. On y reconnaît des adjectifs qui orientent le sens du mot dans certains domaines; *l'autorité souveraine* voisine avec *l'autorité royale* et *paternelle.* La fonction *Thème,* parce qu'elle limite la recherche à la fenêtre d'un paragraphe, met davantage en valeur les liens syntagmatiques entre les unités. Le graphe suivant, obtenu à partir du calcul des associations privilégiées, fournit des résultats plus complexes qui tissent le réseau lexical, voire sémantique qui s'organise autour du mot-pôle *autorité* en sélectionnant des items déjà représentatifs de l'univers imaginaire de l'écrivain, en termes de fréquence relative.

[16] Le schéma laisse apparaître à la fois les cooccurrents directs du mot-pôle et les relations secondes entre les cooccurrents-mêmes. L'épaisseur du trait (pointillé ou trait plein) est fonction de la densité de la liaison. La taille des caractères est en relation avec l'intégration du mot dans le réseau lexical esquissé autour du mot-pôle.

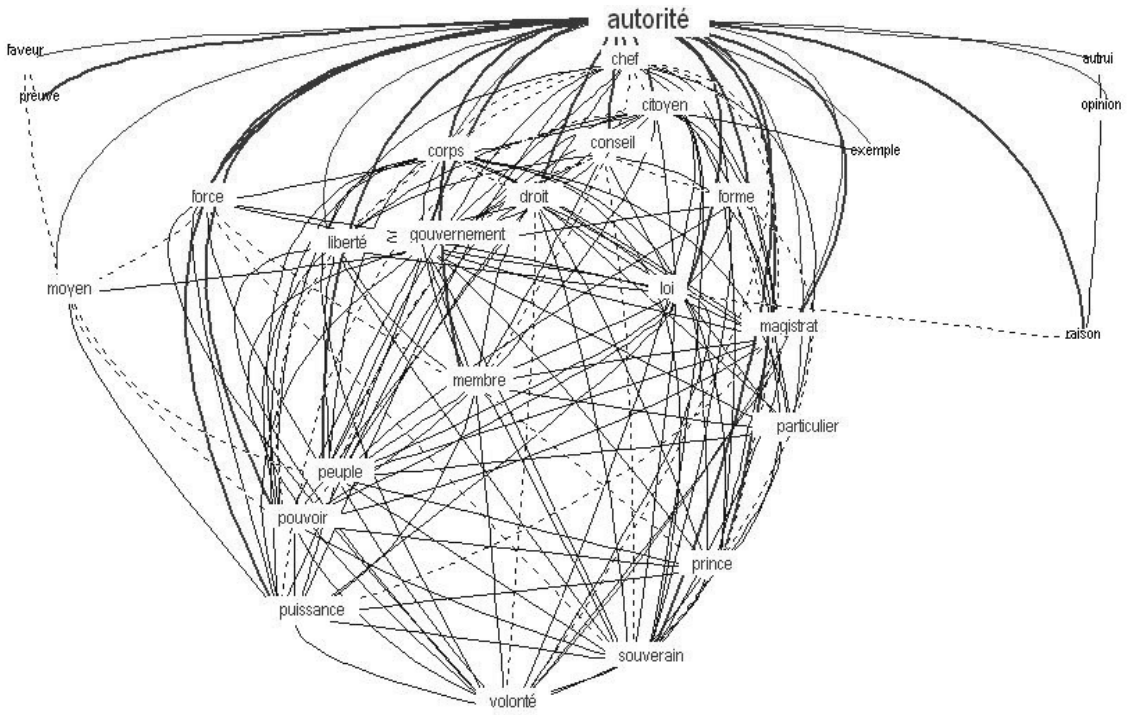


Fig. 4 – Les corrélats lexicaux du mot *autorité* dans la base de données *Rousseau*. Fonction des associations privilégiées.

Ce graphe esquisse les sens politiques ou juridiques du terme – avec des termes comme *citoyen, chef, droit, souverain, loi*. Le mot *liberté*, donné comme antonyme du terme par le *Trésor de la Langue Française*, n'entre cependant pas toujours en conflit avec *autorité*: l'autorité du gouvernement comme l'autorité de la loi garantit la liberté du citoyen. Le recours à l'examen des textes explicite les associations, pointées par les calculs et présentées dans les graphiques. On observe ainsi que l'emploi politique du terme apparaît majoritairement dans *Discours sur l'inégalité* et dans *Le Contrat social*, notamment en association avec l'adjectif *souveraine*. L'autorité est plus morale dans *Émile*, régissant les relations éducatives tandis que la proximité avec *opinion* renvoie au sens étymologique du terme *auctoritas*.

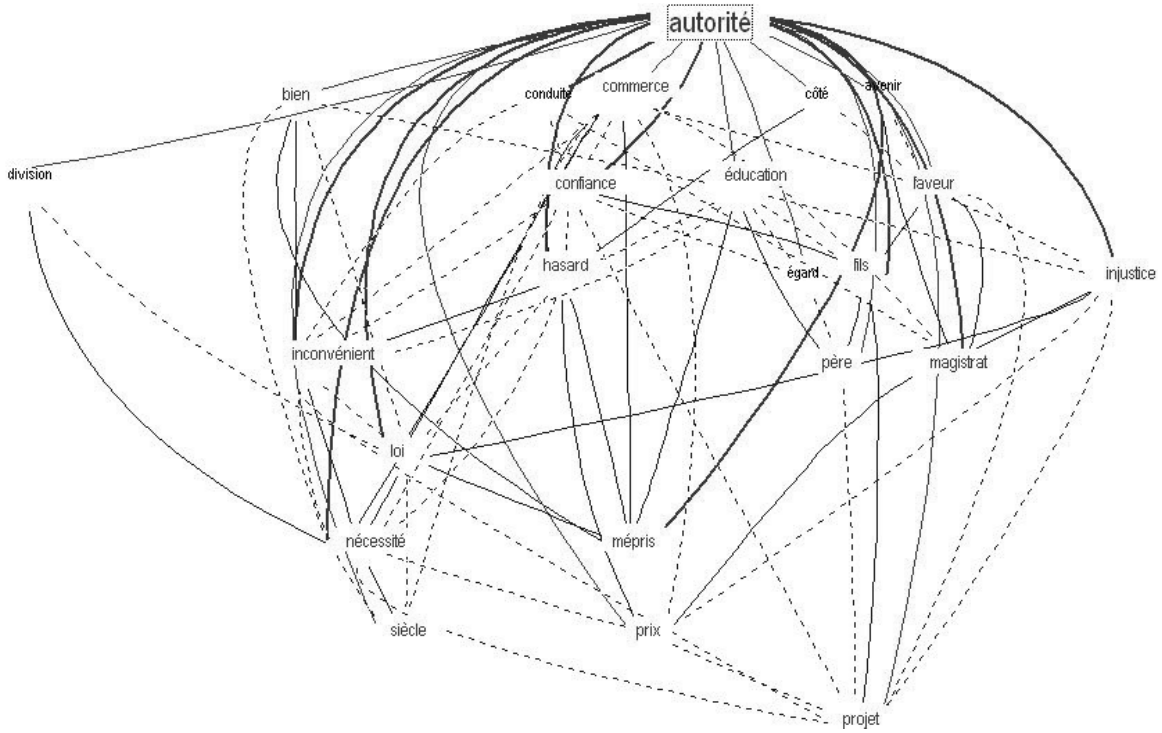


Fig. 5 – Les corrélats lexicaux du mot *autorité* dans la base de données *Diderot*. Fonction des associations privilégiées.

À titre de comparaison et pour introduire à la contextualisation intertextuelle, la figure 5 propose les corrélats lexicaux du même mot *autorité* dans la base *Diderot*. Les champs lexicaux qui s'esquissent ici diffèrent de ceux illustrés par la figure 4. Si le vocabulaire appartient essentiellement au domaine politique chez Rousseau, celui de Diderot emprunte au domaine juridique avec le terme *magistrat* et surtout à la sphère familiale et éducative, tout en associant des mots évaluatifs, absents chez Rousseau, comme *mépris, injustice, confiance*. Une contextualisation intertextuelle plus systématique peut être menée et amorcer une caractérisation de l'imaginaire, de la pensée d'écrivains contemporains.

3. Contextualisation intertextuelle

La contextualisation intertextuelle adopte un point de vue exogène en confrontant plusieurs corpus. Les bases constituées sur chacun des philosophes des Lumières servent de champ d'étude où peut s'exercer la confrontation.

À titre expérimental, nous proposons une mise en contraste des philosophes des Lumières, Montesquieu, Diderot, Voltaire, Rousseau autour du mot *sentiment*.

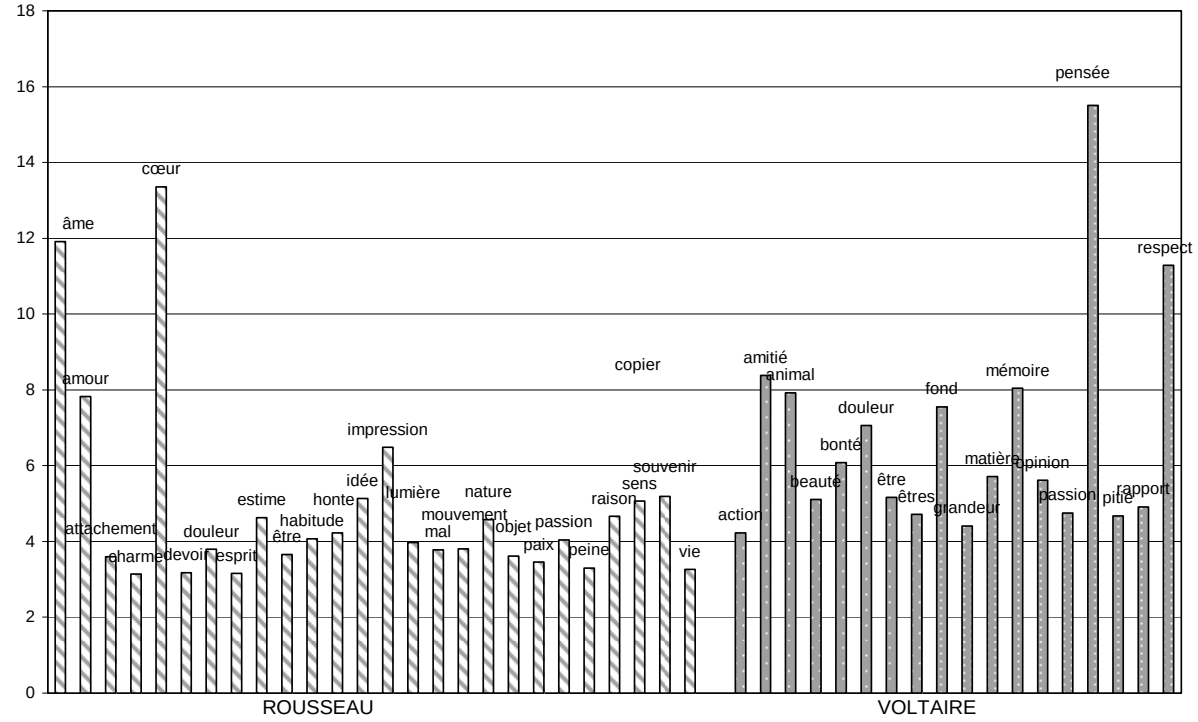


Fig. 6 – Les corrélats lexicaux du mot *sentiment* dans les bases de données *Rousseau* et *Voltaire*[17]

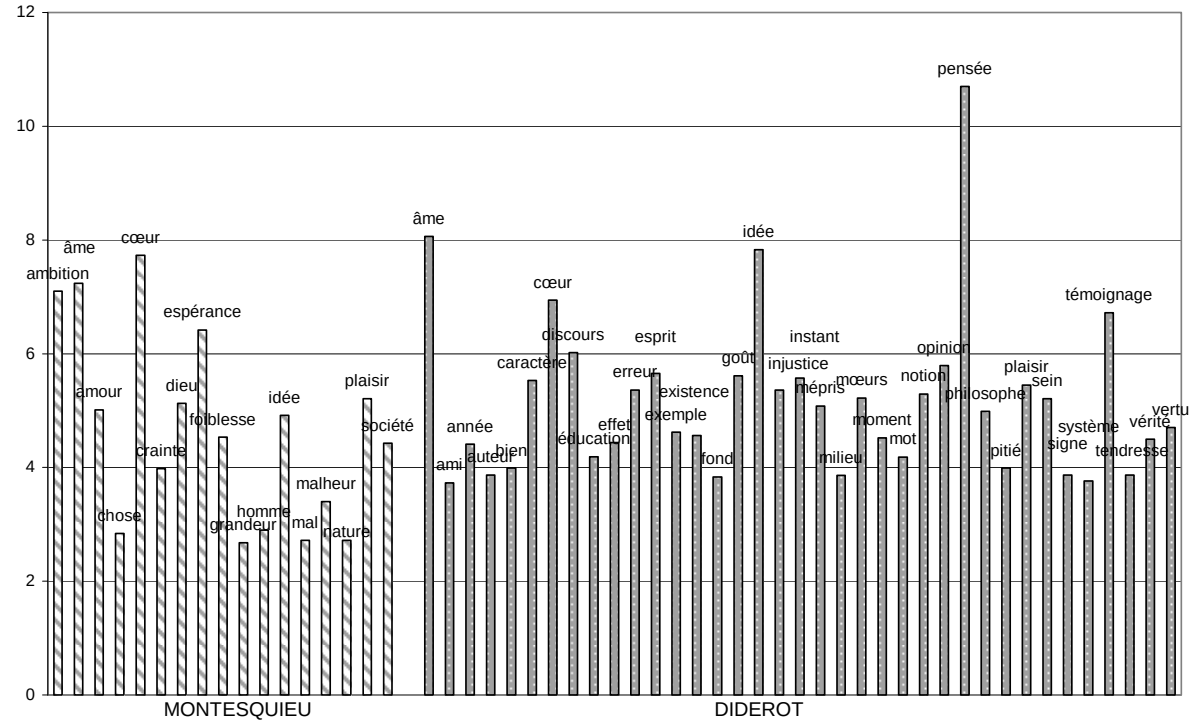


Fig. 7 – Les corrélats lexicaux du mot *sentiment* dans les bases de données *Montesquieu et Diderot*

[17] Ces histogrammes s'interprètent de la façon suivante: La hauteur des bâtons est déterminée par la valeur de l'écart réduit affecté à chaque donnée, après calcul des spécificités. Plus les *bâtons* sont hauts, plus les écarts réduits sont élevés, plus la force d'attraction avec le mot *sentiment* est forte.

Des précisions sur la méthodologie s'imposent. Ces histogrammes proposent les associations lexicales privilégiées qui s'organisent autour du mot *sentiment* chez les quatre auteurs. Sont éludées cependant, de fait, les configurations syntaxiques dans lesquelles s'insère le mot recherché: on ne retient en effet que les substantifs, extraits de l'environnement syntaxique. Mais les configurations syntaxiques où s'insère le mot *sentiment* peuvent être à l'origine de variations sémantiques puisque l'on sait que le sens d'un terme dépend de son entourage contextuel. On risque alors de mêler des formes identiques en apparence, mais aux acceptions cependant divergentes; seul le retour au contexte élargi permet de rétablir les nuances sémantiques qui s'imposent.

Par exemple, l'actualisation par l'article indéfini ou la complémentation par un syntagme nominal prépositionnel orientent vers le sens affectif de *sentiment*, proche d'une *émotion*, tandis que l'emploi absolu, déterminé par un article défini, induit des considérations plus philosophiques. Comme l'a montré l'analyse par les concordanciers, le choix des adjectifs prédicats infléchit le sens du substantif qu'ils qualifient. Ces précautions prises, ce sont les corrélations lexicales contrastives qui peuvent être interprétées.

Dans le voisinage immédiat du terme *sentiment*, deux mots pointent les deux orientations principales du terme, présentes dans les dictionnaires de l'époque, notamment le *Dictionnaire de l'Académie française*, 4^{e} édition (1762) et l'*Encyclopédie, ou Dictionnaire Raisonné des Sciences, des Arts et des Métiers* et sont partagées, comme cooccurrents privilégiés, par Montesquieu, Diderot et Rousseau: les mots *âme* et *cœur*. Le sentiment est en effet, pour une première acception, la "perception que l'ame a des objets, par le moyen des organes des sens" (*Dictionnaire de l'Académie)*.

Le mot *âme* ne se retrouve pas, en revanche, parmi les associations privilégiées mises ainsi en évidence chez Voltaire; cela ne veut par dire que le terme soit absent dans ce corpus, mais cela signifie qu'il n'entre pas dans les corrélations privilégiées de *sentiment* ou que d'autres cooccurrents prévalent.

En observant de plus près les contextes des occurrences du terme – par le biais par exemple d'un concordancier – on se rend compte qu'une trilogie lexicale parcourt le corpus voltairien associant les items *sentiment, animal et mémoire*: elle révèle la réflexion réitérée sur les animaux, leur mémoire ou leurs éventuels sentiments menée par Voltaire.

Chez Montesquieu, Diderot, Rousseau encore, le mot *idée* se retrouve en position de cooccurrent privilégié sur les histogrammes – sans doute à la faveur de la relation antonymique qui lie généralement les deux mots à l'époque.

Le mot *cœur* qui se distingue sur tous les histogrammes, sauf sur celui attribué à Voltaire, manifeste la seconde orientation sémantique essentielle du terme, liée à l'affectivité. Le recours à la contextualisation intratextuelle est encore indispensable pour affiner les commentaires. Les

structures syntaxiques sont en effet spécifiques dans ce cas; un complément déterminatif lie un second substantif par une relation de type attributif – du type *sentiment de pitié* - ou alors le syntagme *un sentiment* désigne un état affectif contingent. L'éventail des sentiments est étroitement lié cette fois au contexte dans lequel s'insère le mot. Chez Montesquieu, ces sentiments ont un emploi politique, comme la *vertu*, définie comme *l'amour de la République*, la *crainte*, posée comme pilier d'un gouvernement despotique dans *L'Esprit des lois* ou encore l'*ambition*, sentiment utile à la société lorsqu'il se dirige bien (*Pensées diverses*).
La gamme de "tous les mouvemens de l'ame" (*Dictionnaire de l'Académie*) est parcourue dans les textes à vocation romanesque, dont on retrouve des représentants au travers des associations remarquables. Diderot privilégie *l'injustice, le mépris, la tendresse* et, comme Voltaire, *la pitié*, que Rousseau définit comme le "premier sentiment relatif qui touche le cœur humain dans l'ordre de la nature" dans *Émile.* Mais dans l'univers rousseauiste, tel qu'il est représenté dans la base, d'autres mots prennent le pas sur la pitié; *l'amour*, "ce sentiment céleste", *la passion, l'attachement, l'estime, le devoir*, autant de termes spécifiques de l'ensemble romanesque tandis que le substantif *raison* se partage à peu près équitablement entre les essais et les romans écrits par Rousseau.
Enfin, un cooccurrent attesté chez Diderot et Voltaire, le mot *opinion*, renvoie au sens de *sentiment* comme *conviction intellectuelle,* conformément à l'étymon *sententia,* et se retrouve dans le syntagme *mon sentiment*.

Conclusion

Le travail préparatoire des données textuelles effectué par le logiciel *Hyperbase* déconstruit le corpus en ensembles paradigmatiques, qui développent soit son versant lexical, soit son versant grammatical.
Ainsi structurées, les données sont soumises à un traitement quantitatif, qui en calcule les fréquences, appréciées de manière relative par rapport à un corpus-norme. La mesure des variations met en évidence des contrastes, seuls susceptibles d'ouvrir aux commentaires interprétatifs. La présentation des résultats adopte une forme tabulaire, sous forme de listes.
L'examen des cooccurrences, qui peut s'exercer à différents paliers du texte, depuis le syntagme jusqu'au texte tout entier, observe les données replacées dans leur environnement linguistique. La projection des résultats se fait alors sur un schéma où sont représentées des constellations lexicales, qui mettent en évidence des réseaux associatifs plus ou moins éloignés dans l'espace du texte. La fenêtre de recherche de la cooccurrence peut être le paragraphe ou le texte tout entier; dans le premier cas, sont maintenus les liens syntaxiques qui peuvent unir les unités; dans le second cas, est davantage mis en évidence l'univers imaginaire ou

intellectuel d'un écrivain, amorçant une possible contextualisation intertextuelle. La mise en contraste de plusieurs écrivains, réunis autour de l'emploi d'un même mot, peut affiner l'analyse littéraire, voire mener à l'histoire des idées.

La textométrie apparaît comme complémentaire de l'analyse cursive d'un texte littéraire; elle en renouvelle cependant la lecture en multipliant les parcours possibles, présentant le texte sous forme de listes ou de réseaux. Elle ouvre enfin à une lecture sérielle[18] qui peut déduire des paradigmes textuels articulés sur une différenciation auctoriale ou générique, lorsque plusieurs grands corpus sont mis en contraste. Cette lecture est à même d'établir une série ou un réseau de faits homogènes selon le point de vue choisi pour l'analyse, susceptible de caractériser de grands ensembles textuels.

Bibliographie

Brunet, É. (2007): Fréquences et séquences. Mise en œuvre dans Hyperbase. In: Lexicometrica: Topographie et topologie textuelles. Disponible: http://lexicometrica.univ-paris3.fr/numspeciaux/special9/brunet.pdf.

— (2009): Comptes d'auteurs. Études statistiques, de Rabelais à Gracq. Paris (Champion). Avec un Dvd contenant les Bases littéraires.

— (2011): Hyperbase. Logiciel hypertexte pour le traitement documentaire et statistique des corpus textuels. Manuel de référence. Disponible: ftp://ancilla.unice.fr/manuel.pdf

Launay, L. & M. (1979): Le Vocabulaire littéraire de J.-J. Rousseau. Genève (Slatkine)/Paris (Champion). Coll. des Études rousseauistes et index des œuvres de J.-J. Rousseau. Série A. Champs sémantiques, v. 2.

Launay, M. (1977): Le Vocabulaire politique de J.-J. Rousseau. Genève (Slatkine)/Paris (Champion). Coll. des index et concordances de J.-J. Rousseau. Série A. Champs sémantiques, v. 1.

Muller, Ch. (1964): Essai de statistique lexicale. Paris (Klincksieck).

Rastier, F. (1987): Sémantique interprétative, Paris (PUF).

— (1996): La sémantique des thèmes - ou le voyage sentimental. In: Texto! Disponible: <http://www.revue-texto.net/Inedits/Rastier/Rastier_Themes.html>

— (2001): Arts et sciences du texte. Paris (PUF).

— (juin-sept. 2003): De la signification au sens. Pour une sémiotique sans ontologie. In: Texto! Disponible: <http://www.revue-texto.net/Inedits/Rastier/Rastier_Semiotique-ontologie.html>

Dictionnaire de l'Académie française. (1762): Disponible: <http://artfl.atilf.fr/dictionnaires/ACADEMIE/QUATRIEME/quatrieme.fr.html>

Encyclopédie, ou Dictionnaire Raisonné des Sciences, des Arts et des Métiers. (1751-1772): Disponible: <http://portail.atilf.fr/encyclopedie/>

Trésor de la langue française. Disponible: <http://atilf.atilf.fr/tlf.htm>

[18] Par référence à la stylistique sérielle établie par Molinié, G. (1986): Éléments de stylistique française. Paris (PUF).