Hugocentric Tendencies or Can One Approach Hugo Counting Words

Étienne Brunet

Hugocentric Tendencies or Can One Approach Hugo Counting Words[1]

Etienne Brunet

Le corpus de Hugo offre de *grands avantages* à qui veut éprouver la validité des méthodes lexicométriques. D'abord il s'agit d'une oeuvre familière, où la critique s'est exercée depuis longtemps et qui a rencontré dès l'origine la faveur du public. Le sentiment des lecteurs offre donc ici une garantie et un étalon de mesure qui permettent d'apprécier les enseignements des chiffres et de rejeter, le cas échéant, leurs extravagances. D'autre part l'étendue du corpus permet de bénéficier de la loi des grands nombres. Mais aussi l'étendue chronologique parcourue par un auteur précoce qui meurt octogénaire est assez ample pour qu'on puisse espérer saisir les phénomènes d'évolution. Enfin la variété des domaines de l'écriture où Hugo s'est illustré avec une égale maîtrise rend possible l'étude des genres littéraires. En s'attachant à un même écrivain, on isole avec plus de pureté les variables du temps et du genre, puisqu'on neutralise la diversité des tempéraments. Et en cela la monographie de Hugo acquiert le statut d'une expérience exemplaire.

Les données exploitées sont de même type que celles de nos précédentes monographies, consacrées à Giraudoux, Rousseau, Proust, Zola. On fera donc l'économie des longues explications. Ces données ont été comme d'habitude communiquées par l'Institut National de la Langue Fançaise, sous la forme de « fichiers-répertoires », sorte d'index bruts établis sur bande magnétique. La difficulté principale du traitement[2] est venue de ce que le dépouillement de ces données reposait sur des éditions qui ne

1. Article paru dans *Literary and Linguistic Computing*, vol.3, n°2, 1988, p.79-93. Seule la mise en page a été revue et corrigée.

2. Ce traitement a été réalisé par la voie télématique, en utilisant le serveur IBM du CNUSC de Montpellier, le système VM et le langage PL1.

correspondaient pas à nos préférences. Dans le cas de Hugo, comme dans celui de Proust ou de Zola, l'édition de la Pléiade semble en effet devoir servir de référence non seulement en vertu de son excellence, mais aussi pour son homogénéité et sa disponibilité. Il a donc fallu transposer la pagination d'une édition à l'autre, et cette opération a concerné 5000 pages. Nous ne nous attarderons pas aux lenteurs manuelles du repérage des pages, non plus qu'aux difficultés et aux approximations de la lemmatisation, ni aux particularités de la présentation synoptique de l'Index. Il suffit d'indiquer ici que le corpus s'étend sur plus de 60 ans (de 1822 à 1885), vingt textes complets[3] et plus de deux millions de mots[4] et qu'on y trouve huit recueils poétiques, quatre pièces de théâtre, dont deux en vers *(Hernani* et *Ruy Blas),* trois romans *(Notre-Dame de Paris, Les Misérables* et *Les Travailleurs de la mer),* un récit de voyage *(Le Rhin)* et quatre recueils de correspondance. On peut évidemment regretter l'absence des *Châtiments* et de quelques autres textes de premier plan, contester la préférence donnée à certains recueils *Les Feuilles d'automne* et *Les Rayons et les ombres* plutôt qu'à d'autres : *Orientales, Chants du crépuscule* ou *Voix intérieures.* Mais tel qu'il est – et tel que nous l'avons reçu – ce corpus donne des garanties suffisantes quant à l'impartialité (puisque nous n'avons pas choisi ses composants), quant à la taille (c'est en volume l'un des plus importants qu'on puisse extraire des données du *Trésor)* et quant à la variété (on y trouve tous les genres abordés par Hugo et toutes les périodes de sa carrière). La représentativité de l'échantillonnage est elle-même largement reconnue par le choeur des spécialistes de Hugo qui sont unanimes à voir dans *Les Misérables* le chef-d'oeuvre de ses romans, dans *Les Contemplations* et *La Légende des siècles* la fine fleur de sa production poétique et dans *Hernani* et *Ruy Blas* les meilleures de ses pièces. Ainsi fixées les frontières du corpus – et en admettant qu'elles puissent être déplacées ou agrandies – le dictionnaire de Hugo prend la forme de l'extrait reproduit dans le tableau 1. Chaque mot (il

3. Comme *Les Misérables* risquaient de déséquilibrer l'ensemble par leur masse imposante, on a divisé ce roman-fleuve en trois tronçons. Quant à *La Légende des siècles,* cette oeuvre se présentait d'elle-même en trois publications. Nous avons respecté ce fractionnement, qui remonte à l'auteur. C'est donc 22 textes (ou sous-textes) qui forment le corpus.

4. C'est moins que Zola, mais plus que Proust et Chateaubriand. Bien entendu l'oeuvre complète de Hugo est loin d'être contenue dans ce corpus

s'agit de vocables, non de formes)[5] y occupe deux ou trois lignes, qui contiennent la fréquence totale et les sous-fréquences, et aussi, lorsque le calcul est rendu possible par une fréquence suffisante, des écarts réduits calculés selon une norme interne (le corpus dans son ensemble) pour chacun des textes et chacun des genres, ou selon une norme externe, elle-même diversifiée (le corpus total XIX-XX[e] siècles, ou celui de l'époque de Hugo, de 1815 à 1885, ou seulement la poésie de la même époque, ou enfin la prose littéraire du temps de Hugo).

				f= fréquence		r= tendance			s= dispersion			<--comparaison externe (corpus TLF)--> c=corpus e=époque p=prose V=vers							
compar.interne -->				pa=poésie		ty= théâtre.vers			tp= théâtre.prose			ro= romans			co=corresp.				
LET	HER	AUT	DAM	BOR	TUD	RUY	RAY	RHI	COI	CON	LS1	MI1	MI2	MI3	FIN	RUE	CO2	MER	CO3	LS2	LS3
amour				f:	1183		r:	-0.399			s:	1.631		c:	-0.9	e:	-3.1	p:	-9.0	v:	-11.6
			467	pa	16.2		ty	5.9			tp	3.8			306	ro	-10.2		144	co	-6.3
193	34	37	68	23	22	28	51	27	46	141	34	35	91	93	35	71	26	24	300	69	29
19.5	5.8	6.1	-3.4	3.3	2.1	2.7	8.7	-8.6	-2.6	13.1	-1.4	-7.2	-1.6	-0.2	1.4	12.6	-8.1	-6.2	-7.1	2.3	1.8
amouraché				f:	1																
0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
amouracher				f:	2																
0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0
amourette				f:	29		r:	-0.050			s:	0.991		c:	12.6	e:	9.7	p:	9.0	v:	2.3
			4	pa	-0.9		ty	0.3			tp	6.8			18	ro	2.3		0	co	-3.9
0	0	0	0	4	2	1	0	0	0	0	0	6	5	4	1	2	0	3	0	0	1
amoureuse				f:	25		r:	-0.367			s:	0.885		c:	-2.0	e:	-2.1	p:	-1.7	v:	-3.4
			3	pa	-1.1		ty	1.7			tp	6.1			14	ro	1.6		1	co	-3.2
0	1	0	4	1	4	1	1	0	0	2	0	3	2	3	0	0	0	2	1	0	0
amoureusement				f:	7									c:	0.2	e:	0.1	p:	-0.3	v:	0.7
0	1	0	1	0	0	1	0	2	0	0	0	0	0	2	0	0	0	0	0	0	0
amoureux				f:	139		r:	-0.113			s:	0.864		c:	1.5	e:	-1.7	p:	-2.6	v:	-2.7
			33	pa	1.0		ty	2.9			tp	5.7			69	ro	2.1		15	co	-5.8
0	1	3	12	5	8	8	2	11	0	7	13	7	18	25	1	8	2	7	3	5	4
-2.6	-0.4	1.0	0.1	3.1	4.9	4.3	0.1	-0.7	-2.9	0.5	-1.4	-1.5	1.6	4.3	-1.3	4.8	-3.1	-0.7	-2.6	-0.5	1.0
amphibie				f:	5									c:	0.2	e:	1.0	p:	1.7	v:	-0.8
0	0	0	1	0	0	0	0	1	0	0	0	0	1	0	0	0	0	2	0	0	0
amphibologique				f:	2																
0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0
amphictyon				f:	6																
0	0	0	0	0	0	0	0	0	0	0	1	0	0	3	0	0	0	0	0	1	1

Tableau 1. Extrait du dictionnaire des fréquences de Hugo

Cette riche panoplie de tests permet de neutraliser à volonté l'effet du genre littéraire ou celui de la chronologie, et d'harmoniser étroitement les conditions auxquelles sont soumis les termes de la comparaison. Un coefficient de corrélation chronologique et une mesure de la dispersion complètent cet ensemble. Ainsi le premier mot de notre extrait, l'*amour* (1.183 occurrences) montre des excédents – attendus – dans les *Lettres à la fiancée* (19,5) et dans les *Chansons des rues et des bois* (12,6), un déclin sur l'ensemble de la série (r= -0,40), et une relative discrétion de Hugo, si on tient compte de l'époque (- 3,1) et du genre (prose littéraire -9 ; vers - 11,3).

5. *L'Index synoptique* reprend toutes les informations quantitatives du *Dictionnaire de Hugo*, en y ajoutant les références. Il y ajoute aussi le détail des formes rattachées aux vocables.

1. La structure lexicale

1.1. Le tableau 2, représenté ci-dessous, donne les éléments de base sur lesquels s'appuient tous les calculs de notre étude. Une colonne y précise l'étendue (en occurrences) de chaque texte, et une autre le nombre de vocables.

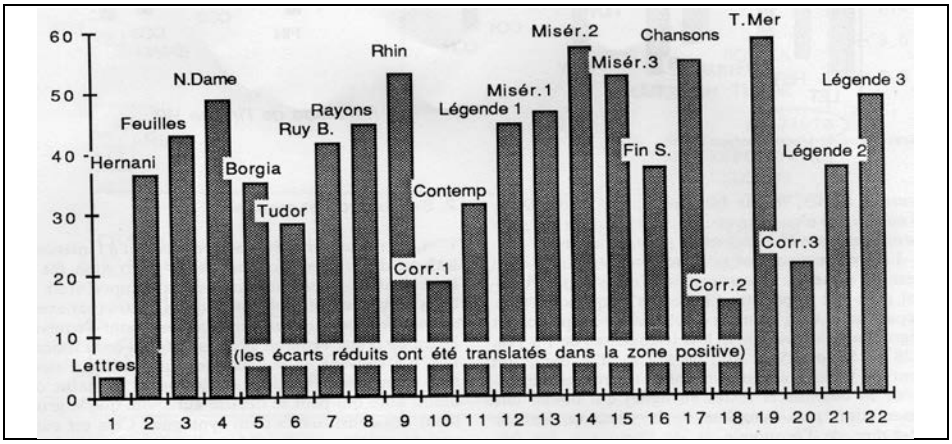
	vocables	occurrences		vocables	occurrences
Lettres	3731	93890	Légende1	5892	74820
Hernani	2476	22833	Misérables1	9091	180672
Feuilles	3001	24866	Misérables2	10005	187724
Notre-Dame	9168	176515	Misérables3	9183	165869
Lucrèce B.	2230	20719	Fin de Satan	4174	48237
Marie Tud.	2179	24947	Chansons..	4213	31580
Ruy Blas	3306	29868	Corresp2	6961	186022
Rayons	3384	28597	Travailleurs	8824	134135
Le Rhin	9949	199799	Corresp3	5483	108914
Corresp1	5527	117402	Légende2	6058	92763
Contempl	5475	87827	Légende3	4189	36287

Tableau 2. Les données de bases

C'est assez pour s'interroger sur la question de la *richesse lexicale,* qui revêt un intérêt particulier chez Hugo. Dans *Réponse à un acte d'accusation,* Hugo prétend avoir mis « un bonnet rouge au vieux dictionnaire » et donné la citoyenneté à tous les mots que le goût classique rejetait au ruisseau. Avant de voir si cette prétention est justifiée, on peut remarquer que cet affranchissement n'est pas absolu et que des limitations sont imposées par le genre littéraire.

Le graphique 1 montre bien l'étagement des genres[6] : en bas la variété lexicale est faible dans les textes épistolaires, car les mêmes préoccupations – et donc les mêmes mots – se répètent d'une lettre à l'autre, et particulièrement les protestations enflammées que le jeune Hugo adresse à sa fiancée dans le premier recueil. Puis vient le théâtre dont le discours s'adresse à l'oreille. Ici un peu de redondance est nécessaire et une certaine simplicité dans le choix des mots, car le spectateur n'a pas la possibilité de consulter un dictionnaire. La poésie occupe l'étage du dessus (les pièces en vers se situant à l'entresol).

6. Le calcul se fonde non seulement sur N et V (l'effectif des occurrences et des vocables de chaque texte), mais aussi sur le tableau de distribution des fréquences dans le corpus. La méthode utilisée ici est celle de la loi binomiale, selon la démarche établie par Charles Muller.



Graphique 1. La richesse lexicale par la loi binomiale

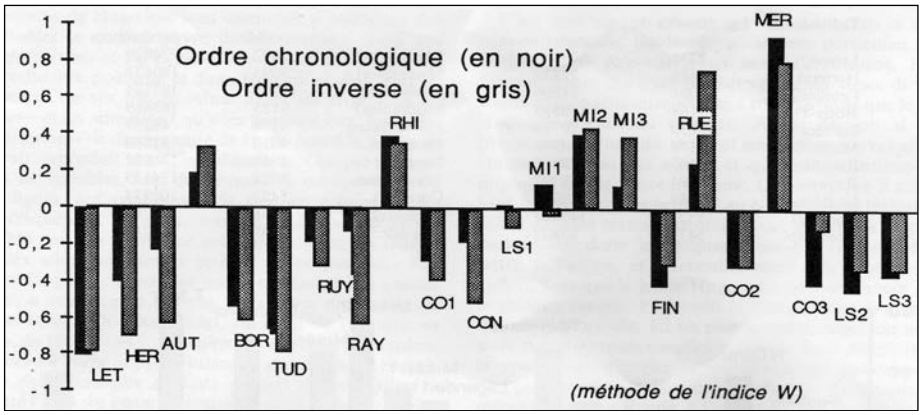
Ses exigences sont très précises et se situent entre le trop et le trop peu : d'une part elle a de la tenue, le goût de la parure, et ne se contente pas des mots passe-partout de la conversation, mais elle refuse les excès lexicaux, les mots trop techniques, trop rares, trop pédants et ceux dont le registre convient mal à la dignité poétique. Hugo a beau dire : *« Je nommai le cochon par son nom ; pourquoi pas ? »*, cette promotion du cochon est très éphémère et les vers de Hugo l'ignorent complètement, avant comme après les *Contemplations*[7]. A l'étage supérieur enfin s'installe le roman, de *Notre-Dame de Paris* aux *Travailleurs de la mer.* C'est là que le foisonnement lexical se donne libre cours, d'une part parce que le roman est libre de visiter tous les recoins de l'univers (c'est aussi le cas des récits de voyage, comme *Le Rhin)*, d'autre part parce que les convenances s'y effacent et permettent des excursions du côté de l'argot, des régionalismes, des mots étrangers et des vocabulaires techniques. Or Hugo est friand des curiosités que lui offrent les voyages, l'histoire ou les dictionnaires. Et il n'aime pas garder pour lui seul ses trouvailles. Le roman devient ainsi l'exutoire de son intempérance lexicale.

Mais cette intempérance ne peut être démontrée que par une comparaison extérieure. Avec 20.602 vocables pour 2.074.286 occurrences,

7. Journet et Robert n'ont pas trouvé la trace du *cochon* dans la poésie hugolienne, en dehors de cette citation. Pour être exact, le mot réapparaît une fois dans la *Légende des siècles.*

le corpus de Hugo peut être rapproché de celui de Zola, qui compte moins de mots différents (19.337) pour une étendue nettement plus grande (2.874.755). La comparaison avec Proust ou Chateaubriand tourne pareillement à l'avantage de Hugo, du moins si l'on s'en tient à la prose littéraire et au roman[8], c'est-à-dire au genre littéraire commun à tous.

1.2. Les faits de structure lexicale peuvent être observés selon un point de vue un peu différent, qui s'attache à relever l'accroissement du vocabulaire, la situation étant celle d'un lecteur méthodique qui dépouillerait l'un après l'autre dans l'ordre chronologique les textes de Hugo, en notant les mots nouveaux, qui n'ont pas d'exemple dans la production antérieure. Les effectifs obtenus, qui vont nécessairement en s'amenuisant, sont appréciés par référence à un modèle (ce peut être encore la loi binomiale, ou une formule d'approximation comme notre indice *w*) et les écarts donnent lieu à une représentation qui offre des traits assez semblables à ceux du graphique 1 : même prédominance des romans, même suprématie des *Travailleurs de la mer*, même écrasement du théâtre et de la correspondance. On note seulement que la poésie est aussi rebelle au changement, et que ses tendances très conservatrices lui déconseillent de renouveler son stock lexical.



Graphique 2. L'accroissement lexical

8. Le roman de Giraudoux reste cependant hors d'atteinte. Mais, né plus tard, Giraudoux jouit de circonstances plus favorables, le vocabulaire français, gonflé par l'inflation lexicale, offrant un choix plus large au XX^{e} siècle.

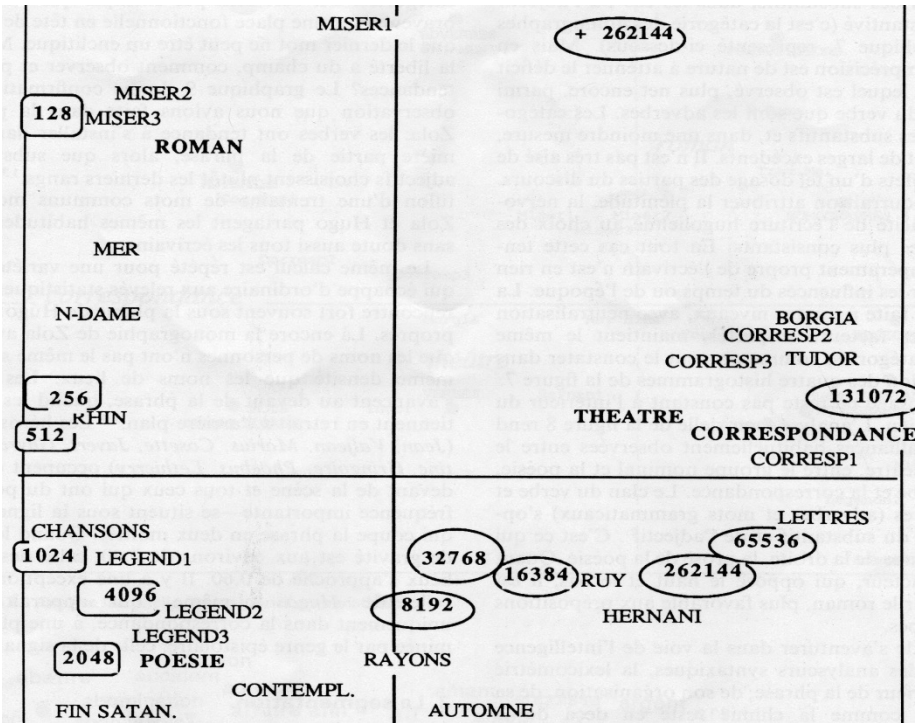
La poésie gère son bien à l'économie, en s'interdisant l'aventure. Les observations sont de même nature si l'on fait le voyage retour dans la chronologie et qu'on relève les nouveautés de l'avant-dernier texte après avoir lu le dernier, en remontant le temps. C'est ce que montrent les zones grises du graphique 2. On remarquera toutefois que, dans cette perspective inversée, le troisième sous-ensemble des *Misérables* plus de nouveautés que le début du roman.

1.3. On a poussé plus loin encore l'expérience, en relevant non pas les mots nouveaux, mais les mots exclusifs, qui n'ont qu'une occurrence chez Hugo et qui n'en ont pas davantage dans le corpus XIX-XX^{e} du *Trésor de la langue française.* C'est ce qu'on appelle les *hapax.* Il est parfois délicat de décider si l'on doit accorder ou refuser l'entrée du dictionnaire à ces marginaux dont beaucoup sont des immigrés arrêtés à la frontière du lexique. Sur un total de 454 hapax homologués[9], 106, soit près du quart, appartiennent aux *Travailleurs de la mer,* 98 aux *Misérables,* 83 à *Notre-Dame* et 76 au *Rhin.* Il n'en reste que 42 pour la poésie, 38 pour la correspondance et 2 seulement pour le théâtre.

1.4. Les perspectives que nous venons d'explorer, qu'il s'agisse de richesse, d'accroissement ou d'hapax, donnaient peut-être la part trop belle aux fréquences rares. En répartissant tous les mots en classes de fréquences et en regroupant celles-ci en 13 lots (fréquences de 1 à 128, de 128 à 256, de 256 à 512, de 512 à 1.024, etc.)[10], on obtient un tableau à deux dimensions (groupes de fréquences en colonne et textes en ligne) qui donne une représentation plus complète de la structure lexicale (c'est-à-dire de l'économie et de l'équilibre des fréquences) et qui peut faire l'objet d'une analyse factorielle. Le résultat d'une telle analyse est représenté dans le graphique 3.

9. Voici les premiers dans l'ordre alphabétique : *abaca, abax, abodrite, abreuveur, acarien, accompagneresse, accortement, aciculiforme, affieffeur, afflation, aiguilleter, alète, aise, alunière, andryade, antenale, anthylle,* etc.

10. Ces jalons (128, 256, 512, ..., 262144) appartiennent au corpus des XIX^{e} et XX^{e} siècles.



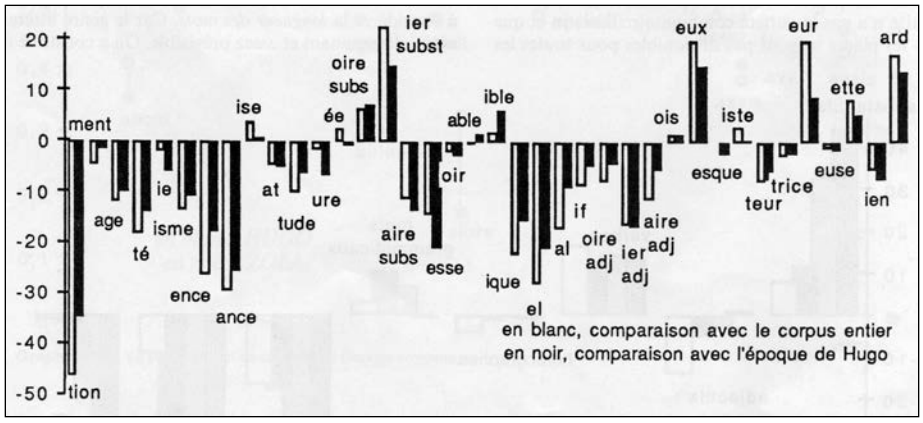
Graphique 3. Analyse factorielle des groupes de fréquence

Les groupes de fréquences y dessinent un arc de cercle caractéristique des données sérielles : des mots rares (classe 128, 256 et 512) établis dans le quadrant supérieur gauche, là où se concentrent les romans, on descend à la moitié inférieure, d'abord à gauche (classes 1.024, 4.096 et 2.048), puis à droite (classes 8.192, 32.768, 16.384). Ici, parmi les fréquences moyennes, s'étend le royaume du vers où tous les recueils poétiques prennent place. Enfin la boucle tend à se refermer en s'écartant sur la droite, puis en rebroussant chemin vers le haut. Les classes de haute fréquence occupent cet espace (262.144, 65.536, 131.072), que fréquentent le théâtre et la correspondance. Les pièces en vers *Hernani* et *Ruy Blas* hésitent dans le quadrant inférieur droit et guignent du côté de la poésie, tandis que les pièces en prose voisinent avec la correspondance dans le quadrant supérieur droit. Si les classes extrêmes se portent vers le haut, là où s'est établi le roman, c'est que le genre romanesque cultive à la fois les mots rares et les fréquences très

hautes, tandis que la poésie privilégie les mots de fréquence moyenne. Cette structure, qui avait été déjà décelée dans notre étude du vocabulaire français, de 1789 à nos jours, se trouve donc confirmée dans le cas de Hugo qui, malgré *la Préface de Cromwell*, n'a nullement aboli la barrière des genres.

2. Suffixation et syntaxe

2.1. La syntaxe s'introduit subrepticement à l'intérieur du lexique dans le phénomène de la suffixation (et plus largement de la préfixation et de la composition). Certains mots abstraits qui cumulent les affixes (par exemple *désintoxication* ou *anticonstitutionnel)* sont l'équivalent lexicalisé d'une véritable proposition. Or cette lexicalisation n'a souvent qu'une existence provisoire, l'association d'une racine et d'un suffixe ayant le statut d'une union libre qui peut se défaire aussi vite que le groupement des constituants d'un syntagme. Cela est surtout vrai de certaines variétés abstraites de suffixes qui ont la rapidité de reproduction des cellules cancéreuses. Or Hugo s'en éloigne comme s'il y avait un danger. Il n'y a pas chez lui de déclaration de guerre à l'abstraction, comme on peut en trouver sous la plume de Zola ou de Giraudoux. Mais instinctivement Hugo donne à l'idée la forme d'une image ou d'un symbole, comme Zola lui donne la forme d'un objet. Et le choix précautionneux qu'il fait dans les variétés de suffixes est à peu près celui de Zola. Ceux qu'il écarte – c'est la majorité – se situent dans la partie inférieure de l'histogramme (graphique 4).



Graphique 4. La suffixation chez Hugo. Comparaison externe

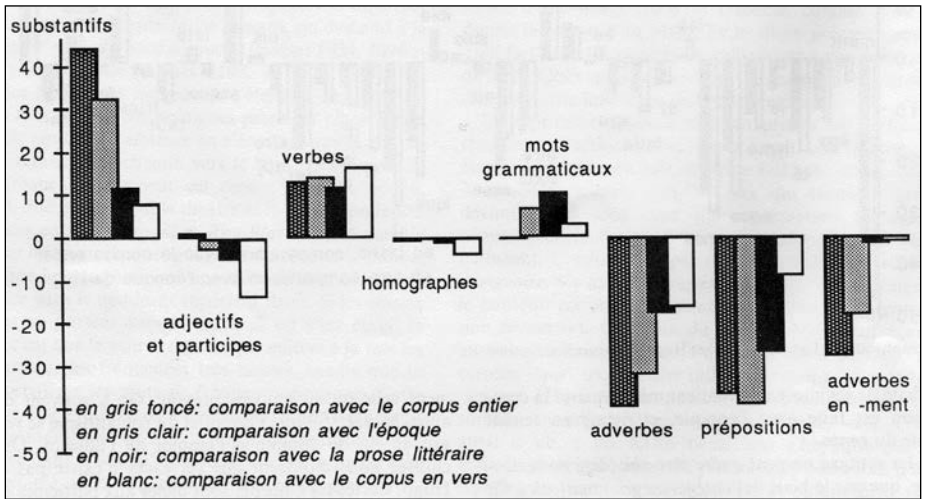
Ils expriment généralement un procès ou parfois le résultat de l'action *(-tion, -ment, -age),* une qualité, un état, une science ou une doctrine *(-té, -ie, -isme, -tude, -ure, -ente, -ante, -at)*[11]. Ceux qui trouvent grâce devant Hugo sont ceux qui représentent un agent, comme *-ier* (subst.) ou un instrument, comme *-oire* (subst.), selon le type : *ouvrier, charretier, ciboire, mangeoire.* S'y ajoute la variété en *-ée* qui sert à désigner le contenu concret d'un contenant précis (une *ventrée,* une *brouettée).* Ce refus du langage spéculatif écarte aussi les suffixes d'adjectifs qui sont liés à un concept, comme *-ique, -el, -al, -oire (adj.), -ier (adj.), -aire (adj.), -ien, -if.* Hugo accepte par contre les suffixes (mi-substantifs, mi-adjectifs) en *-eur, -eux, -ette, -ard.* C'est aussi le choix du XIX^{e} siècle, qui s'explique par le goût du descriptif, voire du pittoresque. Mais loin d'accompagner le mouvement, Hugo surenchérit. La spécificité de ses choix se maintient même quand la comparaison est faite avec l'époque, et même en tenant compte du genre.[12]

2.2. La syntaxe ne peut guère être abordée, en lexicométrie, que par le biais des *catégories grammaticales.* Or la répartition des parties du discours n'est pas exactement celle qu'on pouvait prévoir. On a souvent dit – après Musset[13] – que l'essentiel du romantisme se réduisait en fin de compte à l'emploi de l'adjectif. Or les chiffres nous montrent que cet excès n'existe pas chez Hugo. Certes les adjectifs sont mêlés aux participes dans la lemmatisation héritée du *T.L.F.*, et certains autres se dégagent mal des substantifs, dans les cas où l'adjectif peut être substantivé (c'est la catégorie des homographes dans le graphique 5, représenté ci-dessous). Mais en réalité cette imprécision est de nature à atténuer le déficit des adjectifs, lequel est observé, plus net encore, parmi ces adjectifs du verbe que sont les adverbes. Les catégories pleines, les substantifs et, dans une moindre mesure, les verbes, ont de larges excédents.

11. Le suffixe *-ise* est épargné, non point tellement à cause de ses sonorités exquises, fort goûtées de Verlaine, mais parce que cette désinence est l'apanage de quelques mots fréquents, comme *église,* qu'on ne devrait pas mêler aux vrais suffixes comme *franchise.* Le suffixe *-ise* n'est d'ailleurs pas le seul à souffrir de l'ambiguïté, même après quelques filtrages.

12. En réalité ces traits sont dûs surtout aux exigences du style poétique. Le roman hugolien n'est pas vraiment avare de suffixes, mêmes abstraits.

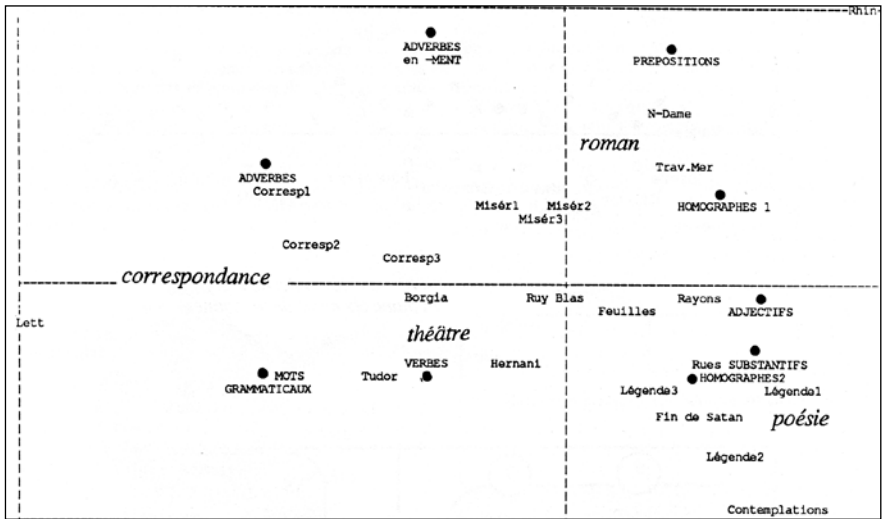
13. Dans *Dupuis et Cotonet.*



Graphique 5. Les catégories. Comparaison externe

Il n'est pas très aisé de mesurer les effets d'un tel dosage des parties du discours. Sans doute pourrait-on attribuer la plénitude, la nervosité et l'efficacité de l'écriture hugolienne au choix des ingrédients les plus consistants. En tout cas cette tendance du tempérament propre de l'écrivain n'est en rien perturbée par les influences du temps ou de l'époque. La comparaison faite à quatre niveaux, avec neutralisation successive des facteurs suspectés, maintient le même dosage des catégories, comme on peut le constater dans la superposition des quatre histogrammes du graphique 5.

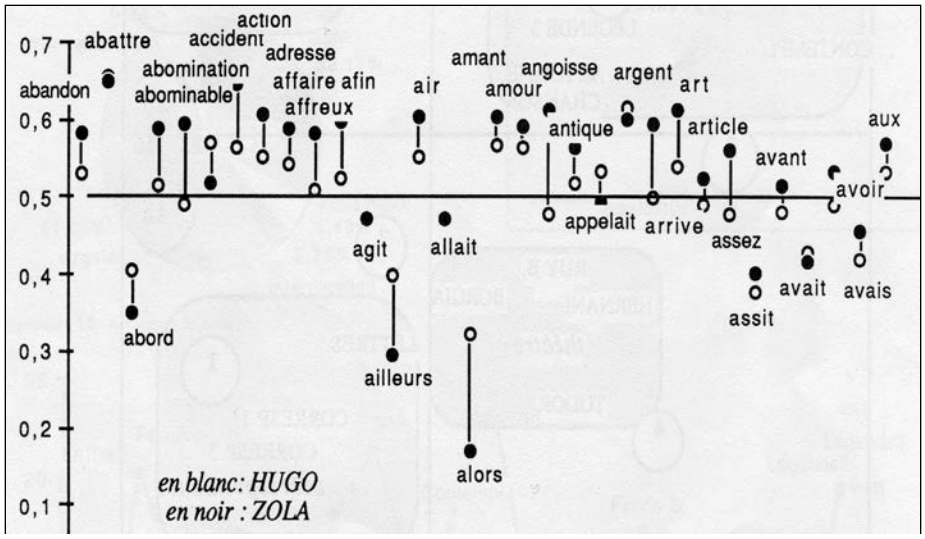
Mais ce dosage ne reste pas constant à l'intérieur du corpus hugolien. L'analyse factorielle du graphique 6 rend compte des alliances habituellement observées – entre le verbe et le théâtre, entre le groupe nominal et la poésie, entre l'adverbe et la correspondance. Le clan du verbe et de ses acolytes (adverbes et mots grammaticaux) s'oppose au clan du substantif et de l'adjectif. C'est ce qui sépare la gauche de la droite, la prose de la poésie. Quant au second facteur, qui oppose le haut et le bas, il est déterminé par le roman, plus favorable aux prépositions et aux adverbes.



Graphique 6. Analyse factorielle des catégories grammaticales

2.3. Faute de s'aventurer dans la voie de l'intelligence artificielle et des analyseurs syntaxiques, la lexicométrie reste à l'extérieur de la phrase, de son organisation, de sa signification, comme la chimie reste en deçà de la biologie. Mais elle peut ne pas se borner à distinguer les éléments, à compter les espèces et à établir la nature de chacune. Elle peut aussi s'attacher à relever l'ordre des éléments et la place des mots. Chacun sait que la phrase française n'a pas la variété combinatoire du latin et que toutes les places ne sont pas disponibles pour toutes les catégories. On voit bien par exemple que certains embrayeurs ont une place fonctionnelle en tête de phrase et que le dernier mot ne peut être un enclitique. Mais là où la liberté a du champ, comment observer et prévoir les tendances ? Le graphique 7 donne confirmation à une observation que nous avions faite dans la phrase de Zola : les verbes ont tendance à s'installer dans la première partie de la phrase, alors que substantifs et adjectifs choisissent plutôt les derniers rangs.[14] L'échantillon d'une trentaine de mots communs montre que Zola et Hugo partagent les mêmes habitudes, comme sans doute aussi tous les écrivains.

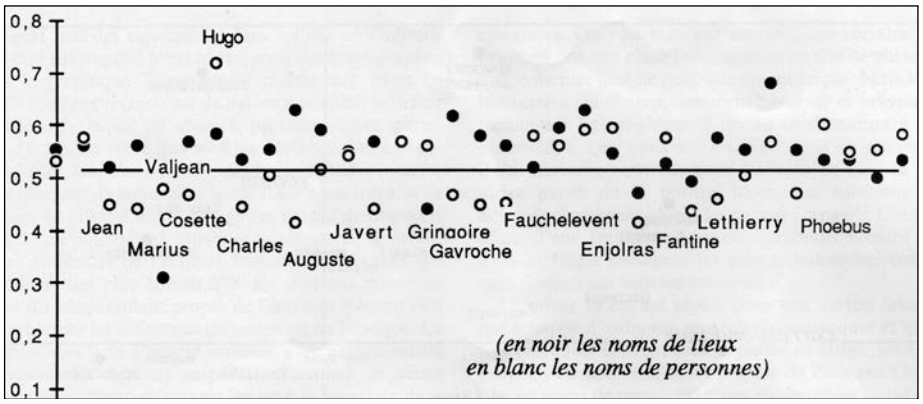
14. Le calcul consiste à établir la place moyenne occupée par un mot, compte tenu de la longueur de toutes les phrases où apparaît ce mot.



Graphique 7. La place du mot dans la phrase (lexique commun)

Le même calcul est répété pour une variété lexicale, qui échappe d'ordinaire aux relevés statistiques et qu'on rencontre fort souvent sous la plume de Hugo : les noms propres (graphique 8). Là encore la monographie de Zola avait établi que les noms de personnes n'ont pas le même statut et la même densité que les noms de lieux. Les premiers s'avancent au devant de la phrase, quand les autres se tiennent en retrait à l'arrière-plan.[15] Les héros de Hugo *(Jean, Valjean, Marius, Cosette, Javert, Gavroche, Fantine, Gringoire, Phoebus, Lethierry)* occupent en effet le devant de la scène et tous ceux qui ont du poids – une fréquence importante – se situent sous la ligne médiane qui coupe la phrase en deux moitiés. Quand leur centre de gravité est aux environs de 0,40, celui des noms de lieux s'approche de 0,60. Il y a une exception notable : celle de *Hugo* lui-même, qui apparaît presque uniquement dans la correspondance, à une place déterminée par le genre épistolaire : celle de la signature.

15. Le fait a été aussi constaté dans trois romans de Giraudoux.



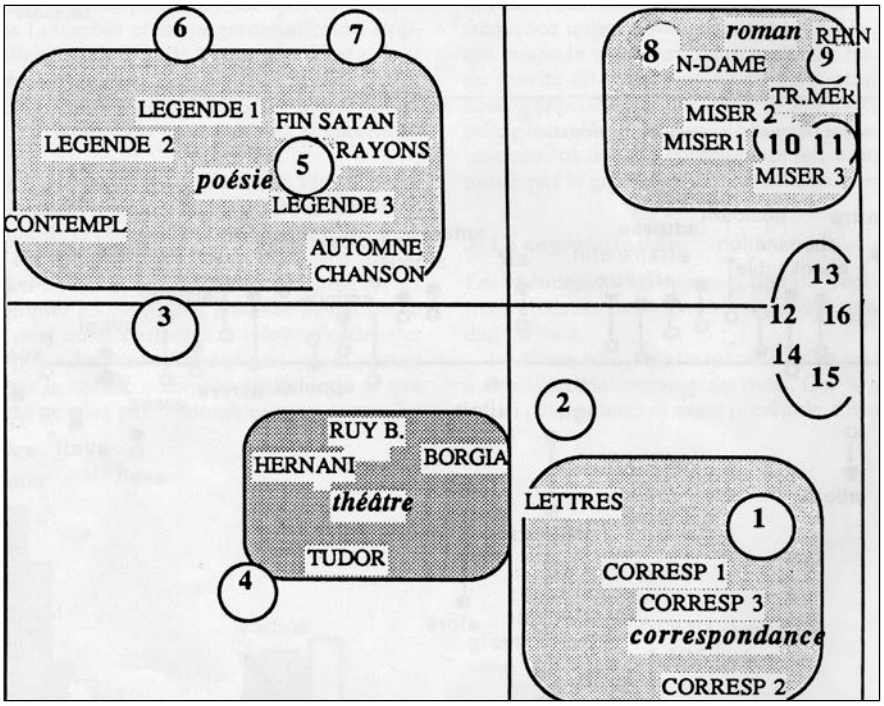
Graphique 8. Place des noms de lieux et de personnes

3. La segmentation

Les phénomènes de segmentation peuvent être abordés à trois niveaux : dans le mot lui-même, dans la phrase et dans le vers.

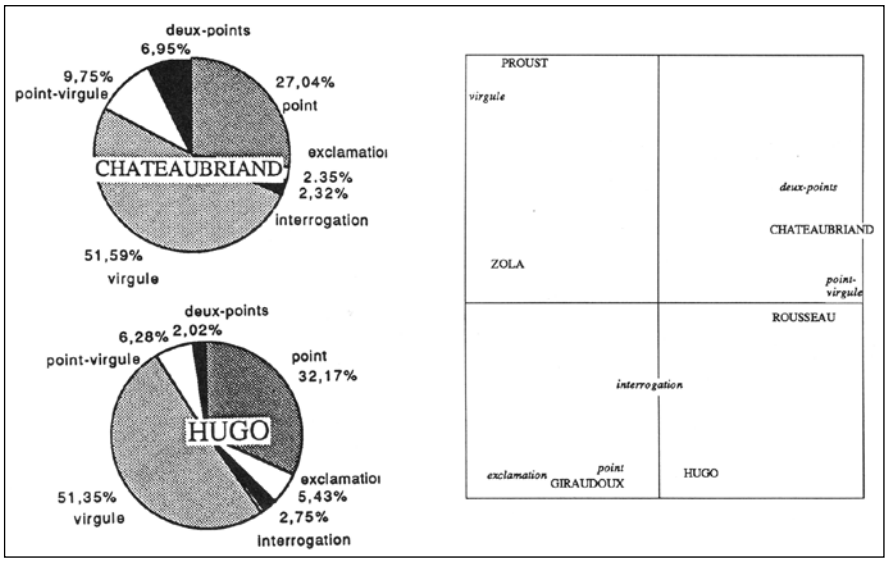
3.1. Nous ne nous attarderons guère au premier stade, à considérer la *longueur des mots.* Car le genre littéraire est ici déterminant et assez prévisible. On a constitué des classes de mots de 1, 2 , …, 16 lettres, en regroupant dans le dernier effectif les mots qui vont au delà de la limite. Le croisement de ces classes avec les textes produit un tableau à deux dimensions qui s'offre à l'analyse factorielle. Le résultat, reproduit dans le graphique 9, illustre les clivages du genre. Le théâtre utilise de préférence les mots courts et le roman les mots longs. Ces deux genres occupent deux quadrants opposés du graphique. Un troisième quadrant, intermédiaire, est réservé à la poésie, laquelle cultive les espèces de calibre moyen (mots de 5, 6 et 7 mots). Enfin la correspondance s'établit dans le dernier coin, où se rejoignent les deux extrêmes : mots très longs et mots très courts. En réalité, ces classes de longueur ne sont pas sans rappeler les classes de fréquences que les genres s'étaient partagées de la même façon dans le graphique 3. Il y a en effet une liaison, soulignée par Pierre Guiraud[16], entre la fréquence des mots et leur coût articulatoire, qui peut s'exprimer en nombre de phonèmes (ou de lettres).

16. Pierre Guiraud, *Problèmes et méthodes de la statistique linguistique*, PUF, 1960, p. 79.



Graphique 9. Analyse factorielle de la longueur du mot

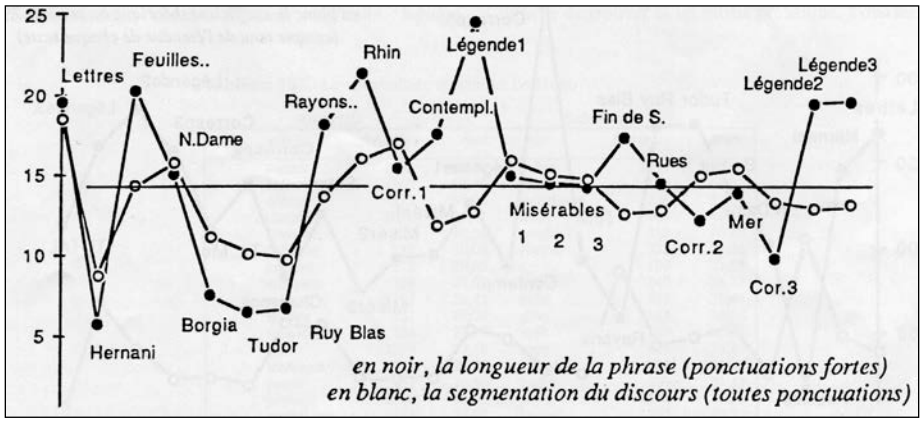
3.2. De la longueur du mot, on passe facilement à la *longueur de la phrase*. L'unité de mesure est cette fois le mot lui-même, la phrase étant définie comme l'ensemble des mots compris entre deux ponctuations fortes. Cet espace est en moyenne de 14,41 mots, soit tout près de la valeur obtenue pour Zola (14), et loin de Chateaubriand (22), Rousseau (27) et Proust (31). La phrase de Hugo est donc courte. La proportion du point est forte comme celle du point d'exclamation. Le point-virgule a déjà perdu chez lui les privilèges dont il jouissait chez Rousseau et Chateaubriand et plus généralement au XVIII^{e} siècle. Et les deux-points suivent le même sort. Les deux histogrammes en secteurs représentés ci-dessous (graphique 10) mettent en parallèle le système des ponctuations chez Hugo et Chateaubriand. Les différences sont considérables. Le même graphique étend la recherche aux six écrivains dont les données nous sont connues.



Graphique 10. Les signes de ponctuation

La synthèse est claire : d'un côté Rousseau et Chateaubriand gardent l'héritage du XVIII[e] siècle et maintiennent en faveur les signes atténués : deux-points et point-virgule. De l'autre côté les ponctuations fortes, non seulement le point, mais aussi les signes affectifs (! et ?) sont le fait de Giraudoux et de Hugo (bas du graphique). Proust, dont la phrase possède un rythme si particulier, s'installe tout seul en haut du graphique, en n'admettant que la virgule à ses côtés. Enfin Zola, sur la ligne de démarcation, reste dans l'expectative.

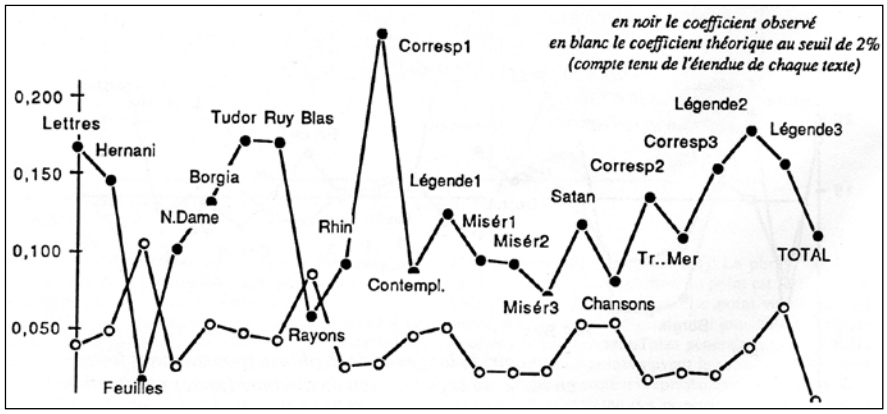
Mais le rythme de la phrase de Hugo n'est pas constant d'un texte à l'autre, ni surtout d'un genre à l'autre. La phrase se raccourcit au théâtre et se racornit dans la correspondance. Elle prend ses aises dans le roman, et particulièrement dans la prose du *Rhin.* Mais c'est surtout en poésie qu'elle se développe et prend de l'ampleur, atteignant les sommets dans l'épopée de la *Légende des siècles* (graphique 11).



Graphique 11. La longueur de la phrase

3.3. Le phénomène du rythme ne se réduit pas à une moyenne. Pour qualifier le tempo d'un morceau de musique, il serait insuffisant de calculer la durée moyenne des notes. L'étude du *rythme* suppose une approche dynamique et non un constat statique. Il faut mesurer la succession des phrases, qui peut être rapide ou lente, régulière ou heurtée. On a donc examiné les 113.456 phrases qui se terminent par un point, en comparant la longueur de chacune à celle de la précédente. Si ces écarts entre phrases consécutives sont importants, le rythme est court et morcelé ; l'auteur spécule sur les effets de surprise, de variété, de choc. Si les écarts sont faibles, ce devrait être le signe d'une rhétorique ample, volontiers périodique, qui, une fois lancée, va jusqu'au bout de son mouvement. Naturellement on dispose d'un modèle qui permet de classer les observations dans la première catégorie, ou dans la seconde, ou dans aucune. En ce qui concerne Hugo, les résultats rejoignent ceux qui ont été obtenus chez Rousseau, Proust et Zola : la phrase longue appelle la phrase longue, et la courte, la courte, l'effet étant celui d'une houle qui tantôt s'agite, et tantôt s'assagit. Le seuil est dépassé de très loin dans 20 textes de Hugo (graphique 12). Deux textes seulement, qui sont courts et poétiques, restent en deçà de ce seuil : *Les Feuilles d'automne* et *Les rayons et les ombres.* Et deux autres, *Les Contemplations* et *Les Chansons des rues et des bois* le dépassent d'assez peu. Il y a là sans doute

une indication : le désir de varier le ton et de casser le rythme s'oppose, dans ces recueils, à la périodicité naturelle de la rhétorique.[17]



Graphique 12. Le rythme de la phrase. L'autocorrélation

Mais le genre poétique dispose d'un autre rythme que celui de la phrase : celui du vers. Il est certes difficile de formaliser, en français, la différence entre les temps forts et les pauses, entre un pied qui se soulève et un pied qui retombe, et même le problème des césures n'est pas simple à expliquer à l'ordinateur, là où manque un signe de ponctuation. Mais il existe un repère rythmique facile à isoler : *la rime. Il* suffit de préciser le moule de la strophe et les couples de rimes sont accordés par la machine. Le traitement a porté sur deux recueils : *Les Feuilles d'automne* et *Les Chansons des rues et des bois.* Dans les deux cas le couple le plus fréquent est la rime *sombre-ombre,* qui apparaît respectivement 10 et 9 fois.[18] Cela n'est pas pour étonner quand on sait que *sombre* est l'adjectif qui vient en tête dans le vocabulaire caractéristique de Hugo, et qu'il en est de même de l'*ombre* parmi les substantifs (voir la liste plus loin). Hugo a naturellement tendance à accorder deux privilèges aux mots qu'il chérit : il les invite plus souvent et il leur

17. Les recueils poétiques sont constitués par Hugo de poèmes indépendants qui ont chacun leur rythme propre et dont l'entrelacement évite la monotonie. La mise en page y est pleine d'imprévu, parfois de ruptures.

18. Dans le sens inverse, le couple *ombre-sombre* est relevé respectivement 5 et *2* fois.

accorde les meilleures places, notamment le fauteuil de la rime.[19] Mais la rime a deux fauteuils tout à la fois associés et dépareillés. L'un, le second, a la prééminence protocolaire et chronologique sur l'autre. Les versificateurs ont souvent reconnu qu'ils trouvaient le second vers avant le premier, ou plus exactement qu'après avoir créé un vers ils se souciaient de lui associer un compagnon, ou mieux un serviteur chargé d'introduire le maître dans la place. Et ce serviteur précède, pour ouvrir les portes et ménager les articulations syntaxiques de la phrase. En somme ils trouvaient le second vers, par inspiration, et fabriquaient le premier, par transpiration. Comment peut-on vérifier le fait ? Un premier indice réside dans l'asymétrie des couples de rimes : les associations ne sont pas réversibles. Le couple *sombre-ombre* est beaucoup plus fréquent que le couple *ombre-sombre* et il en est ainsi des rimes *jour-amour, superbe-herbe, vainqueur-coeur, flamme-âme, encor-or, voiles-étoiles, vermeil-soleil, vivant-vent, mystérieux-cieux, choses-roses, sommes-hommes.* Le terme fort de chacun de ces couples est le second, presque toujours un substantif. Le premier est souvent un adjectif, parfois un verbe ou un adverbe, et sa charge sémantique est moins riche. On soupçonne certaines de ces premières rimes de n'être que des chevilles quasi automatiques et de n'avoir que l'importance secondaire qu'ont les confidents des héros dans le théâtre classique. Le second indice est à chercher dans l'inégalité des choix offerts aux deux places. Les secondes rimes ont tendance à être monopolisées par les mêmes habitués, qui semblent avoir souscrit à un abonnement, alors qu'en première place la variété est bien plus grande, comme si Hugo consultait à ce moment un dictionnaire de rimes. Et il lui arrive alors de choisir les moins attendues. Les rimes rares se trouvent plus souvent en première place qu'en seconde : 782 rimes hapax contre 716 dans les *Feuilles d'automne,* 1.728 contre 1.562 dans les *Chansons.* Et inversement les rimes communes (fréquence au moins égale à 5) se rencontrent moins en première place qu'en seconde (144 contre 221 dans le premier recueil, et respectivement 205 et 415 dans le second).

19. Voici pour *Les feuilles d'automne* le classement des rimes: *âme* et *ombre* (*23* occurrences en fin de vers), *pas* et *sombre* (17 occ.), *ciel* et *flamme* (15), *foule* et *monde* (14), *onde* (12), *fleurs, jour, nuit, terre, Dieu* (11). Dans les *Chansons* la hiérarchie est la suivante: *bois, nuit, ombre, amour, cieux, fleurs, infini, superbe, elle, rose, soir, sombre, herbe, vent, bleu, Dieu, jour, noir, oiseaux, printemps, roses, sourires.*

4. Le contenu lexical

Chacun connait le vers célèbre de Hugo : *Car le mot, c'est le Verbe, et le Verbe c'est Dieu.* Ce n'est pas seulement un jeu de mot, c'est aussi l'expression d'une théorie du mot, conçu comme une unité indissociable qui soude le fond et la forme, le référent et le signe linguistique. Hugo souscrit donc par avance aux enquêtes lexicales qui portent sur son oeuvre, puisque le mot *est un être vivant,* et que le sens – et la réalité même – est dans les mots et non derrière eux. Encore faut-il que l'enquête ouvre par quelque moyen une perspective sur le contenu.

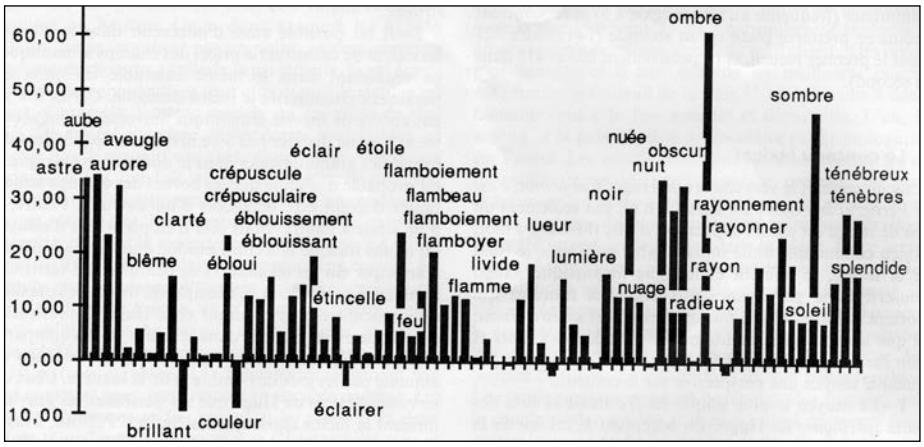
4.1. Le moyen le plus simple est d'obtenir la liste des *mots spécifiques* de Hugo. En acceptant le critère de la fréquence et en prenant pour outil de sélection l'écart réduit, on met en relief les excédents, mais aussi les déficits, significatifs. Il est conseillé de séparer les catégories grammaticales, afin d'augmenter la lisibilité et l'interprétation des listes de spécificités. Nous en donnerons deux extraits dans le tableau 3. L'un concerne les adjectifs, l'autre les substantifs. Dans le cas de ces derniers on voit affleurer à la surface les mots qui sont associés à un thème, ou à une intrigue ou seulement à un passage longuement développé. Tel épisode des *Misérables* explique la présence des *barricades,* des *égouts,* de *l'évêque.* La protection particulière de *Notre-Dame* vaut leur promotion aux *truands* et à l'*archidiacre.* Et de la même façon le *mess,* la *panse* et la *pieuvre* sont échoués dans la liste par le flux des *Travailleurs de la mer.* Mais sous cette écume un peu opaque et trop triviale, on devine en transparence les constantes, les obsessions de Hugo qui ne tiennent à aucune situation particulière et qui s'attachent à des mots hantés : *ombre, gouffre, spectre, aube, gibet, nuée, astre, hydre, brume, nuit, ténèbres, ange, vent, ciel, abîme,* etc. La liste des adjectifs est tout aussi éloquente à ce point de vue. Il suffit de citer les premiers : *sombre, hideux, obscur, farouche, difforme, lugubre, effrayant, noir, sinistre, formidable, pensif, profond, monstrueux.*[20]

20. On n'attachera pas d'importance particulière aux adjectifs *charmant, cher* et *cordial,* qui ne sont que des amabilités coutumières à la correspondance.

Adjectifs					**Substantifs**
Mot	fréquence	écart	mot	fréquence	écart
sombre	1198	46,50	*ombre*	2051	61,43
hideux	360	37,49	*barricade*	349	64,37
obscur	573	33,05	*archidiacre*	208	51,04
charmant	1031	31,41	*mess*	169	48,30
farouche	319	30,56	*gouffre*	415	46,80
difforme	131	30,08	*spectre*	328	37,36
lugubre	272	29,65	*égout*	189	37,25
effrayant	354	28,97	*panse*	140	36,27
noir	1701	28,81	*aube*	337	35,44
sinistre	331	28,24	*gibet*	145	34,54
formidable	284	27,73	*nuée*	289	33,91
cher	1379	27,16	*astre*	418	33,83
ténébreux	194	27,01	*truand*	83	31,84
pensif	251	26,96	*hydre*	109	31,13
profond	1108	26,75	*lettre*	2558	29,76
cordial	163	26,07	*brume*	292	29,37
monstrueux	250	24,01	*antre*	186	29,29

Tableau 3. Le vocabulaire spécifique de Hugo

4.2. I1 est possible aussi d'intervenir dans le contenu lexical, et de constituer a priori des champs sémantiques, en réunissant dans le même ensemble les mots qui paraissent circonscrire le même domaine. Certes il n'y a pas encore de théorie sémantique universellement acceptée et l'on ne dispose pas à ce niveau de l'équivalent des catégories grammaticales dans le domaine syntaxique. Il est probable d'ailleurs que les bornes des champs sémantiques doivent être déplacées d'un auteur à l'autre, et d'un siècle à l'autre. Mais cela n'empêche pas d'essayer, du moins lorsque la lecture assidue d'un écrivain permet d'anticiper sur les résultats et de délimiter des territoires familiers à cet écrivain. L'exemple du thème de la *lumière* est particulièrement éclairant chez Hugo. L'illustration en est donnée dans le graphique 13.

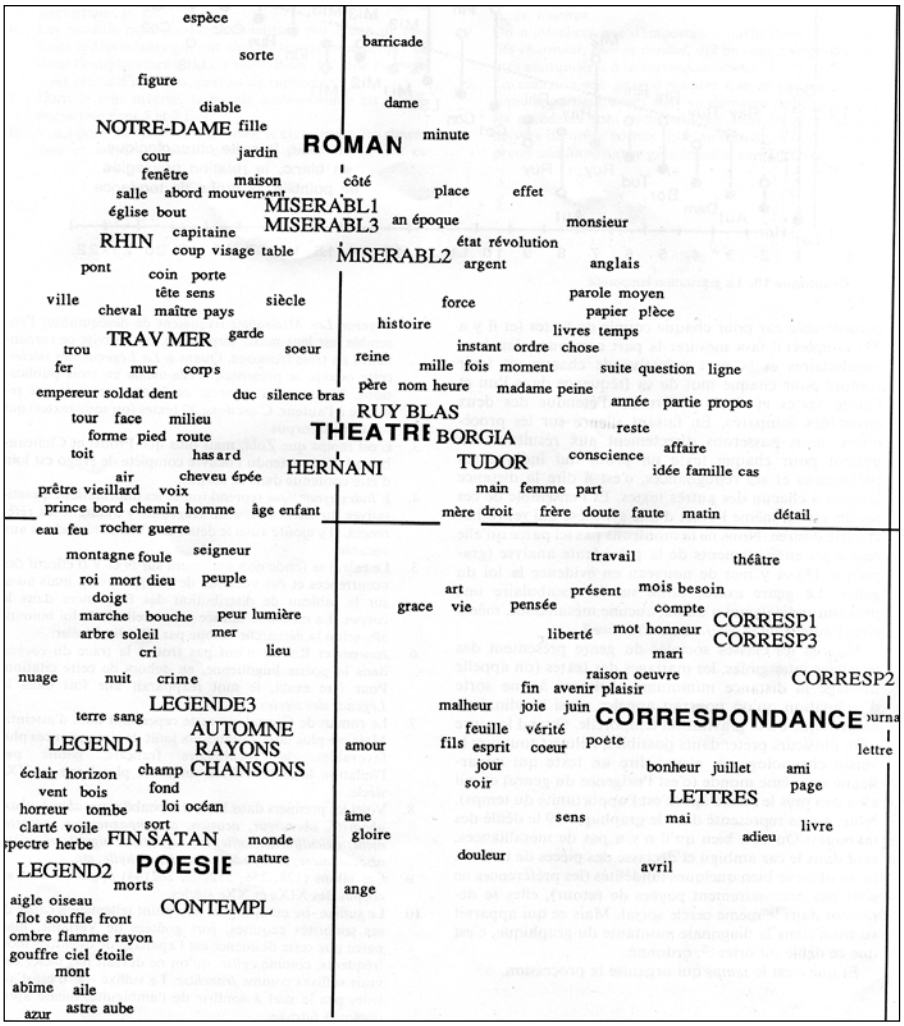


Graphique 13. La lumière. Comparaison externe

La comparaison avec l'époque montre que le regard de Hugo est aimanté par les jeux de l'ombre et de la lumière. Ceux qui ne connaîtraient de Hugo que ses dessins et ses gravures diraient la même chose que les chiffres. Partout, sous sa plume comme sous son pinceau, Hugo fait jaillir le feu et les rayons, mais aussi les ténèbres et le noir profond. Son univers est un ciel sombre zébré d'éclairs. Le graphique 13 le dit clairement, puisque sur 72 mots représentés 3 seulement s'inscrivent dans la zone négative. Et ceux qui s'élèvent le plus haut dans les excédents appartiennent à la face noire du champ : *ombre, sombre, obscur, nuée, nuit, noir, aveugle, ténébreux, ténèbres, crépuscule*.[21]

4.3. Si l'on s'enferme dans le corpus hugolien, le contenu se différencie selon les textes, selon le genre dont ils relèvent et selon la date qui est la leur. *L'analyse factorielle* est un moyen commode pour mettre en relief ces lignes de force qui structurent l'oeuvre d'un écrivain. Parmi beaucoup d'autres, nous avons choisi celle qui concerne quelque 200 substantifs (les plus fréquents du corpus). Le graphique 14 en restitue le détail qu'on laisse au lecteur le soin d'examiner.

21. La comparaison interne montre que ce champ sémantique appartient avant tout au domaine poétique et que les excédents sont systématiques dans les huit recueils de vers de notre corpus. Elle montre aussi que le thème prend une importance grandissante au fil des ans.

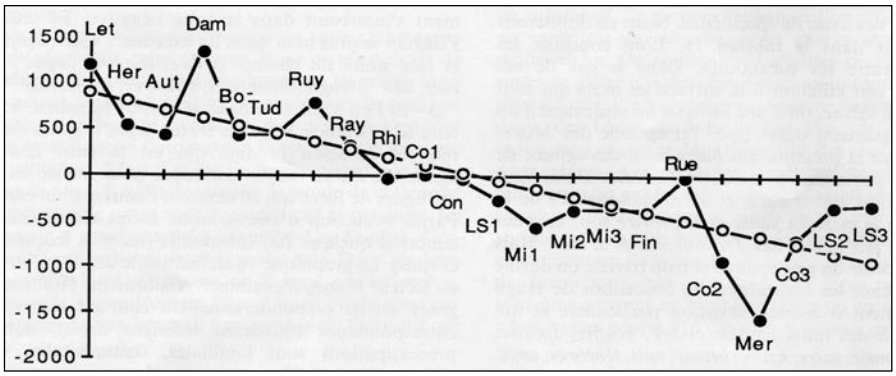


Graphique 14. Analyse factorielle des substantifs fréquents

Visiblement l'influence du genre est ici prépondérante. Un coin est réservé à la correspondance (quadrant inférieur droit), dont les

préoccupations sont familiales sentimentales, éditoriales, utilitaires. Un autre coin jalousement protégé est dévolu à la poésie (quadrant inférieur gauche). Il serait difficile de trouver là un intrus prosaïque.Tous les mots appartiennent au clan fermé du noble langage, et le *cochon y* serait accueilli par des huées. La moitié supérieure est le domaine du roman, chacun des titres se taillant sa sphère d'influence. Bien entendu les trois sous-ensembles des *Misérables* font cause commune. Quant au théâtre, il a quelque peine à affirmer sa spécificité, d'une part parce qu'il est intérieurement divisé (et c'est pourquoi les pièces en vers se rapprochent de la poésie), mais aussi parce le dialogue au théâtre participe tout à la fois au langage de tous les jours, qu'on trouve dans les lettres, mais aussi aux intentions littéraires, qui sont communes au roman. Déchiré par ces tensions internes, le théâtre hésite donc au milieu du graphique, non loin de l'origine, à mi-distance du roman et de la correspondance.

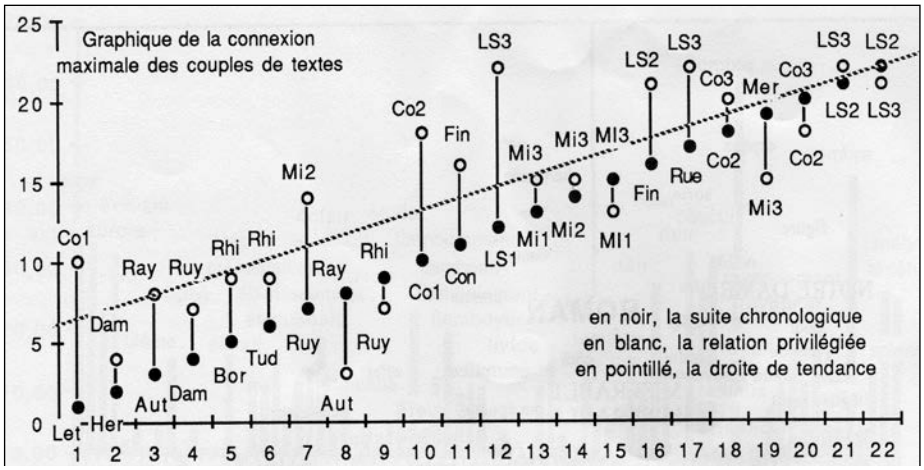
Le temps n'a-t-il donc aucune influence décelable ? Si, mais ses effets n'apparaissent que lorsque les prérogatives du genre sont satisfaites. En général l'évolution chronologique coïncide avec le troisième facteur des analyses factorielles. En reportant sur un graphique les coordonnées de ce troisième facteur et en réunissant cinq jeux de données indépendants (quelques centaines de substantifs, autant de verbes, une centaine d'adjectifs, de participes, d'adverbes), on obtient la courbe suivante, dont la pente ne fait aucun doute (graphique 15).



Graphique 15. Le facteur TEMPS. (Récapitulation de 5 analyses factorielles)

4.4. La perspective interne au corpus gouverne enfin notre dernière tentative pour appréhender le contenu lexical du corpus et y saisir les marques du temps. Cette tentative – à laquelle Muller a donné le nom de *connexion lexicale* – est aussi la plus ambitieuse et la plus impersonnelle. Car elle ne suppose aucun choix préalable. C'est tout le vocabulaire qui est considéré, dans chacun des textes. L'objectif est de mesurer la distance qui sépare chaque texte de tous les autres et de constituer une sorte de carte géographique où les oeuvres représenteront les régions et les mots les habitants. Le calcul est ici considérable car pour chaque couple de textes (et il y a 2.231 couples) il faut mesurer la part commune des deux vocabulaires et la part privative de chacun, et tenir compte pour chaque mot de sa fréquence dans l'un et l'autre textes et de la différence d'étendue des deux ensembles comparés. En faisant silence sur les procédures, nous passerons directement aux résultats. On obtient pour chaque texte un profil qui indique ses préférences et ses répugnances, c'est-à-dire la distance lexicale à chacun des autres textes. Et l'ensemble de ces profils est lui-même l'objet d'une synthèse qui reproduit la carte désirée. Nous ne la montrons pas ici parce qu'elle répète les enseignements de la précédente analyse (graphique 14) et y met de nouveau en évidence la loi du genre. Le genre exerce donc sur le vocabulaire une pression radicale qui n'admet aucune mésalliance, même chez l'auteur de la *Préface de Cromwell.*

Mais si les classes sociales du genre présentent des barrières intangibles, les mariages des textes (on appelle mariage la distance minimum) obéissent à une sorte d'inclination qu'on pourrait appeler aussi l'inclinaison magnétique ou la gravitation temporelle. Quand le genre offre plusieurs prétendants possibles, l'élu est toujours le voisin chronologique, c'est-à-dire un texte qui appartienne au même monde (c'est l'exigence du genre) et qui ait à peu près le même âge (c'est l'opportunité du temps). Nous avons représenté dans le graphique 16 le défilé des mariages.



Graphique 16. La gravitation temporelle

On voit bien qu'il n'y a pas de mésalliances, sauf dans le cas ambigu et déclassé des pièces de théâtre. Si on observe bien quelques infidélités (les préférences ne sont pas nécessairement payées de retour), elles se déploient dans le même cercle social. Mais ce qui apparaît surtout, dans la diagonale montante du graphique, c'est que ce défilé est orienté, ordonné.

Et c'est le *temps* qui organise la procession.