LA STRUCTURE LEXICALE dans l'oeuvre de Hugo

Étienne Brunet

LA STRUCTURE LEXICALE
dans l'oeuvre de Hugo

Etienne Brunet

La présente étude est extraite de notre ouvrage paru aux éditions Slatkine: Le Vocabulaire de Hugo. *On trouvera dans le tome 1(pages 19-60), le détail des tableaux et des calculs sur lesquels se fonde cet article.*

Résumé

Après avoir exploré l'oeuvre de Proust et celle de Zola, on aborde ici le corpus de Hugo. Les données fournies par *l'Institut national de la langue française* constituent un corpus de deux millions de mots, où ont été traités une vingtaine de textes complets, dont les *Misérables,* les *Contemplations,* la *Légende des siècles.* Les questions que la statistique lexicale pose habituellement offrent ici un intérêt particulier du fait que Hugo prétend avoir libéré le goût et ouvert les frontières du lexique. On vérifie ici si ces prétentions sont fondées, en mesurant la richesse lexicale de Hugo, le flux dynamique de l'accroissement et du renouvellement, la part du vocabulaire exclusif et, plus généralement, l'économie des fréquences. Tous ces phénomènes, qui relèvent de la structure lexicale, sont étudiés en tenant compte du genre littéraire abordé, de l'évolution constatée chez l'écrivain et des tendances observées dans la littérature de l'époque.

- I -
La richesse lexicale

La première question que la statistique linguistique se pose quand elle aborde un écrivain peut paraître lassante ou oiseuse: se trouve-t-on devant un vocabulaire riche ou pauvre? Mais s'agissant de Hugo, l'affaire est d'importance puisque Hugo prétend ouvrir les portes du dictionnaire à tous les mots exclus par un goût trop étroit et accorder la citoyenneté lexicale aux représentants du peuple.

J'ai mis un bonnet rouge au vieux dictionnaire...
Oui, je suis ce Danton! je suis ce Robespierre!
J'ai, contre le mot noble à la longue rapière,
Insurgé le vocable ignoble, son valet,
Et j'ai, sur Dangeau mort, égorgé Richelet.
(Réponse à un acte d'accusation, Contemplations)

Un tel libéralisme devrait multiplier les variétés. Notre relevé en recense 20602. Est-ce peu? Est-ce beaucoup? La réponse est dans la comparaison. Précisons d'abord que ce chiffre correspond à celui des vocables lemmatisés selon les normes du TLF et qu'il rend compte d'un corpus de plus de deux millions d'occurrences[1]. Zola compte moins de mots différents (19337 vocables), alors que pourtant l'étendue du corpus zolien est nettement plus grande (2 874 755 occurrences)[2]. Le vocabulaire de Hugo l'emporte donc en variété sur celui de Zola. La comparaison est moins facile avec Chateaubriand, Proust et Giraudoux, dont les données nous sont connues. Car si l'effectif des vocables est plus grand chez Hugo, celui des occurrences l'est aussi, et la conclusion reste en suspens à cause de la corrélation évidente qui lie *N* et V (les vocables et les occurrences). Il faut donc utiliser une formule de pondération qui neutralise les différences de taille. L'indice *w*, que nous avons proposé à cet effet [3], donne les résultats suivants:

	occurrences	*vocables*	*indice w*
Hugo	2074286	20602	10.13
Chateaubriand	1398984	19606	9.71
Zola	2874755	19337	10.97
Proust	1267069	18322	9.84
Giraudoux	671364	15771	9.43

Comme cet indice, dont la valeur est habituellement proche de 10, évolue en raison inverse de la richesse lexicale, on serait amené à admettre que la variété lexicale de Hugo n'atteint pas celle de Chateaubriand, ni celle de Proust ou de Giraudoux. Mais deux ou trois faits sont à prendre en considération: d'une part, de 1789 à nos jours, un phénomène d'inflation lexicale s'est donné libre cours et il n'est peut-être pas légitime de mesurer à la même toise des écrivains que plus d'un siècle sépare. D'autre part la notion de *richesse lexicale* est assez ambiguë, selon qu'elle repose ou non sur les

[1] Ce vaste dépouillement, vient, comme tous ceux que nous avons traités, des dépouilles du *Trésor de la langue française.* Il contient vingt textes complets, dont trois romans (y compris l'intégralité des *Misérables),* quatre pièces de théâtre, huit textes poétiques (dont les *Contemplations* et la *Légende des siècles),* et quatre recueils de correspondance. S'y ajoute un récit de voyage, le *Rhin,* dont le genre est hybride. Comme le poids des *Misérables* risquait de détruire l'équilibre, on a réparti leur masse en trois sous-ensembles, ce qui porte à 22 le nombre de textes considérés.

[2] Bien entendu les mêmes principes ont été respectés pour la lemmatisation des deux corpus. Le nombre d'occurrences est à prendre dans un sens restrictif, en excluant les noms propres et les mots étrangers.

[3] En voici la formule : $w = N^{V^{-a}}$ pour a = 0.185,

facilités de la suffixation, de la préfixation et de l'invention lexicale. On a pu montrer que Zola, méfiant à l'égard de l'abstraction, a peu souvent recours aux affixes et aux ressources de la langue qui permettent de faire du neuf avec du vieux. On observe la même réserve chez Hugo à l'égard des suffixes abstraits et cela représente un handicap numérique par rapport aux écrivains qui n'ont pas ces scrupules. Enfin et surtout le genre littéraire perturbe les résultats. Il se trouve que le corpus de Hugo, à la différence des autres corpus précités, incorpore des textes en vers et des lettres. Or ni la poésie, ni la correspondance ne se prêtent facilement aux curiosités lexicales et à l'enrichissement de la langue. Mais si on opère la confrontation à l'intérieur d'un genre commun, par exemple la prose romanesque, Hugo reprend l'avantage, sinon sur Giraudoux, du moins sur Chateaubriand et Proust [1] :

	occurrences	*vocables*	*indice w*
Hugo(roman)	844915	17212	**9.45**

Cette valeur de 9.45 n'est d'ailleurs atteinte à aucun moment par la prose du XIX^{e} siècle, les huit tranches qu'on y distingue se distribuant comme suit:

1800	*1825*	*1835*	*1845*	*1855*	*1865*	*1875*	*1885*
11.35	10.84	10.67	10.11	10.23	9.70	9.72	9.84

Mais si l'on isole le genre poétique l'originalité du vers hugolien se dissout dans la moyenne du siècle[2], puisqu'on observe la valeur 10.29 dans la poésie de Hugo contre respectivement 10.29, 10.45, 10.42, 9.45, 9.74, 9.08, 9.48, 8.97 dans les huit tranches chronologiques du corpus en vers au XIX^{e} siècle[3]. La conclusion doit donc être nuancée: certes le lexique virtuel de Hugo, largement alimenté d'ailleurs par la lecture systématique des dictionnaires, a une dimension considérable et il se déploie largement dans son discours lorsque le genre et la situation le permettent. Mais précisément, quoi qu'il prétende dans *Réponse à un acte d'accusation,* la liberté n'est pas entière dans le genre poétique, dont Hugo atteint assez vite les limites, et, s'il se montre généralement accueillant aux nouveautés, aux curiosités, et parfois même aux trivialités du vocabulaire, il entasse ses trouvailles au grenier - c'est-à-dire dans la prose, et dans un genre, le roman, où les règles existent à peine - mais non à l'étage noble de la poésie. A ce point de vue, la séparation des genres n'est nullement abolie, malgré la *Préface de Cromwell* .

[1] La comparaison directe, sans calcul, peut être faite entre les *Travailleurs de la mer* (N=134135, V=8824), et le *Temps retrouvé* (respectivement 144752 et 8240). Le roman de Hugo, qui est plus court, a pourtant plus de vocables .

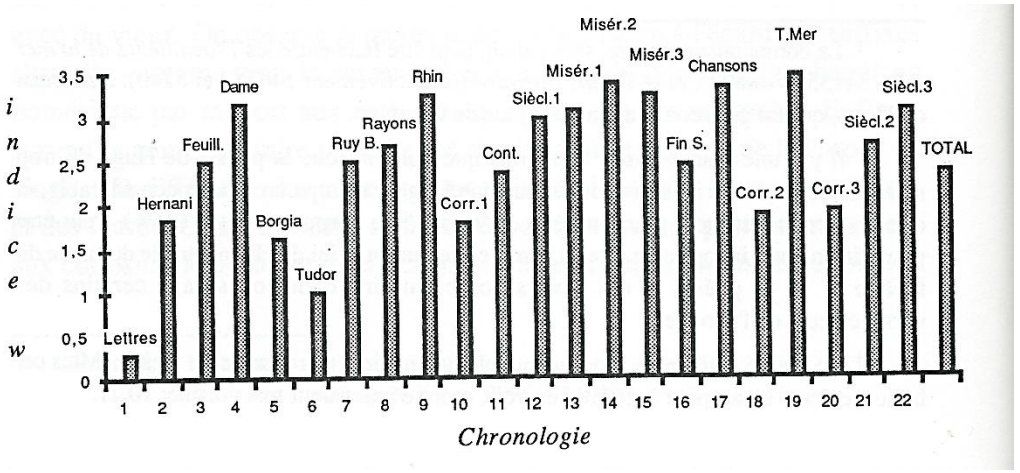
[2] il y a une bonne raison mathématique qui empêche la poésie de Hugo de trop s'écarter de la poésie de son époque, puisque Hugo y occupe un espace considérable, et que ses vers constituent pour un quart la "norme" du corpus poétique (26%). Peut-être pourrait-on aussi invoquer le rôle de chef de file qui fut celui de Hugo dans le domaine du théâtre et de la poésie et qui a sans doute entraîné dans son sillage certains des versificateurs de l'époque.

[3] En réalité l'indice de la poésie hugolienne ne tient pas compte du théâtre. Mais cet indice, calculé à part pour le théâtre en vers, montre une valeur très voisine: 10.21.

Tableau 1 . Les effectifs dans le corpus Hugo

	Rang chrono	Occurr. Formes (brutes)		Occurr. Vocables (lemmatis.)		Accroiss. normal	Accroiss. inverse	Hapax	Fréq.1	Sous-freq.1
Lettres à la fiancée	1	110356	6428	93890	3731	3731	58	0	46	1260
Hernani	2	33821	3685	22833	2476	998	20	0	18	1133
Les Feuilles d'aut.	3	29848	4473	24866	3001	998	27	1	23	1359
Notre-Dame de P.	4	218205	16624	176515	9168	4813	741	83	543	3257
Lucrèce Borgia	5	28096	3304	20719	2230	132	26	0	14	1073
Marie Tudor	6	34274	3241	24947	2179	105	17	1	8	966
Ruy Blas	7	41287	4712	29868	3306	319	68	1	46	1611
Les Rayons et..	8	34868	5146	28597	3384	306	35	0	24	1569
Le Rhin	9	247856	20170	199799	9949	2601	967	76	579	3556
Correspondance1	10	145197	9887	117402	5527	603	275	8	153	2196
Contemplations	11	110121	9692	87827	5475	474	170	2	108	2153
La Légende 1	12	93885	10126	74820	5892	467	275	12	132	2376
Les Misérables 1	13	221861	16155	180672	9091	1116	771	17	391	3399
Les Misérables 2	14	233700	18125	187724	10005	1210	1398	69	613	3792
Les Misérables 3	15	205452	15849	165869	9183	747	1486	21	423	3525
La Fin de Satan	16	60106	6975	48237	4174	106	245	5	52	1737
Les Chansons ..	17	39499	6464	31580	4213	147	435	4	94	2162
Correspondance2	18	240725	13316	186022	6961	454	1133	20	249	2551
Les Travailleurs ..	19	164553	15009	134135	8824	859	3178	106	615	3416
Correspondance3	20	149353	9847	108914	5483	195	2266	10	159	2217
La Légende 2	21	115013	11009	92763	6058	153	2822	11	127	2365
La Légende 3	22	44433	6564	36287	4189	68	4189	7	65	1989
TOTAL		2602509	216801	2074286	20602	20602	20602	454	4482	4482

Figure 2. La richesse lexicale (par l'indice *w*)

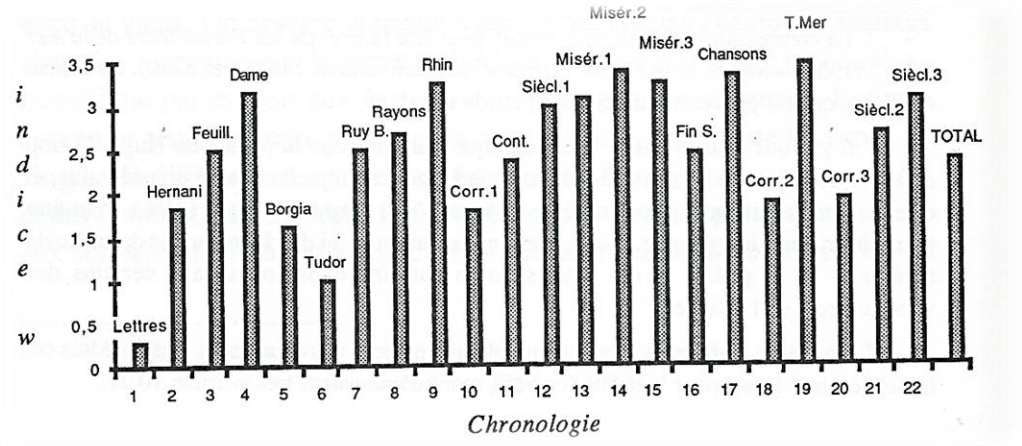


Observons en effet les données du tableau 1 et leur représentation graphique dans la figure 2. À l'étage le plus bas Hugo entrepose les objets utilitaires, les mots de tous les jours qui se répètent sans cesse et qui ont cours dans la correspondance et le théâtre en prose. L'étage intermédiaire est celui du théâtre en vers et de la poésie et l'indice y montre une valeur moyenne. Enfin les romans – le *Rhin* se joint à eux -

occupent le plan supérieur, où la richesse lexicale est la plus forte. Si on néglige la distinction des genres, on peut mesurer l'évolution chronologique de Hugo[1] en considérant la pente de la courbe, qui est ascendante. Contrairement à ce qu'on a observé chez d'autres écrivains, le lexique de Hugo n'a donc pas tendance à se replier ou à se recroqueviller avec l'âge.[2] Le coefficient de corrélation chronologique (r= 0.47) atteint le seuil significatif de 3%.[3] Et le progrès est sensible à l'intérieur des genres, et notamment dans la série des textes en vers: 10.64, 9.99, 9.99, 9.79, 10.13, 9.51, 10.04, 9.14, 9.81, 9.43. La variété lexicale culmine à la fin de la série,dans un recueil plus libre que les autres: *Les Chansons des rues et des bois* et dans le foisonnement épique de la *Légende des siècles.*[4]

Si l'on se méfie de l'indice *w*, qui n'est guère qu'une approximation empirique, on peut recourir à la logique irréprochable de la loi binomiale. Sans détailler un raisonnement d'abord proposé par Muller[5] et maintes fois exposé, bornons-nous à examiner les résultats reproduits dans le graphique 3.

Tableau 3 . La richesse lexicale (par la méthode binomiale)



[1] On ne se cache pas que la date de publication et la date avouée de composition n'ont qu'une relation lâche avec la date réelle de composition. Tout ce qu'on peut dire d'un recueil de Hugo, c'est qu'aucune pièce n'en a été rédigée après la date de parution du recueil. Cette règle triviale ne suffit pas à assurer la chronologie, et dans au moins un cas - celui de *La fin de Satan* - nous avons préféré l'époque de la rédaction à celle de la publication, trop longtemps différée.

[2] Chateaubriand résiste pareillement à la fatigue lexicale et à l'usure des ans.

[3] Rappelons que cette notion de seuil signifie qu'on a 3 chances sur 100 de se tromper lorsqu'on affirme la tendance de Hugo à enrichir son vocabulaire

[4] On peut toutefois hésiter à attribuer au temps ce qui n'est peut-être que l'effet du choix des thèmes. Le lyrisme plus intime qui a la faveur de Hugo dans les *Feuilles d'automne* , les *Rayons et les ombres* et les *Contemplations* n'appelle pas les mêmes ressources lexicales que la poésie ornementale, satirique ou épique. Si notre corpus avait enregistré les *Orientales* et les *Châtiments,* l'évolution ne serait probablement plus si visible.

[5] Voir le chapitre 16: "Modèle binomial d'une distribution théorique" dans *Principes et méthodes de statistique lexicale,* p.101-109.

En prenant appui sur le tableau des distributions des fréquences dans le corpus Hugo [1], on mesure la part du vocabulaire théoriquement absent (et par suite celle du vocabulaire théoriquement présent) dans chacun des textes. Cet effectif attendu est comparé à celui qu'on observe en réalité, et la distance est appréciée par un écart réduit. Une fois de plus on constate que tous ces écarts sont négatifs, ce qui traduit le phénomène partout observé de la spécialisation lexicale: les mêmes mots se retrouvent dans les mêmes textes, en vertu des contraintes, thématiques ou stylistiques, de la situation de discours. Cela provoque des grumeaux dans la pâte lexicale, l'homogénéité en est imparfaite et s'éloigne assez du saupoudrage égal et étale que supposent le hasard et la loi binomiale. Quant aux conclusions suggérées par ces écarts, elles recouvrent exactement celles du tableau précédent. Le graphique 3 montre pareillement l'étagement des genres: 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 vient ensuite. 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.* Au haut du graphique 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.

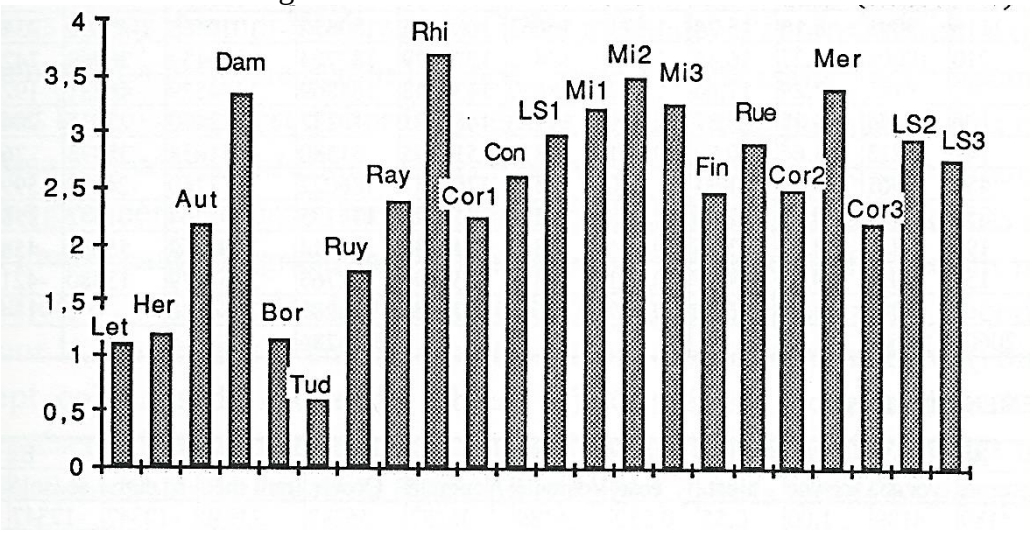
Le parallélisme se reproduit dans la figure 4 (ci-dessous) qui reprend la méthode plus simple de l'indice w, mais l'applique aux formes, avant toute lemmatisation. Les données sont celles qu'on trouve dans les premières colonnes du tableau 1. Observons qu'en théorie il n'y a pas de liaison étroite entre la diversité des formes et la variété des vocables. Et de la même façon les occurrences, qui comprennent cette fois les signes de ponctuations, les noms propres et les mots étrangers, peuvent subir les distorsions propres à ces éléments du discours. Il est

1 Ce tableau est détaillé dans notre ouvrage *Le Vocabulaire de Hugo,* p. 53. En voici les premiers éléments:

fréquence	1	2	3	4	5	6	7	8	9	10
effectif	4482	2226	1553	1099	872	727	598	544	404	342
fréquence	11	12	13	14	15	16	17	18	19	20
effectif	365	316	307	236	236	201	198	179	170	159

remarquable que les perturbations annoncées ne se manifestent pas et que les enseignements tirés de ces données moins pures et plus brutes (mais aussi plus objectives puisqu'elles échappent à toute intervention humaine) coïncident avec ceux que suggèrent les données lemmatisées. Cette convergence rassurera les optimistes.

Figure 4. La variété des formes brutes



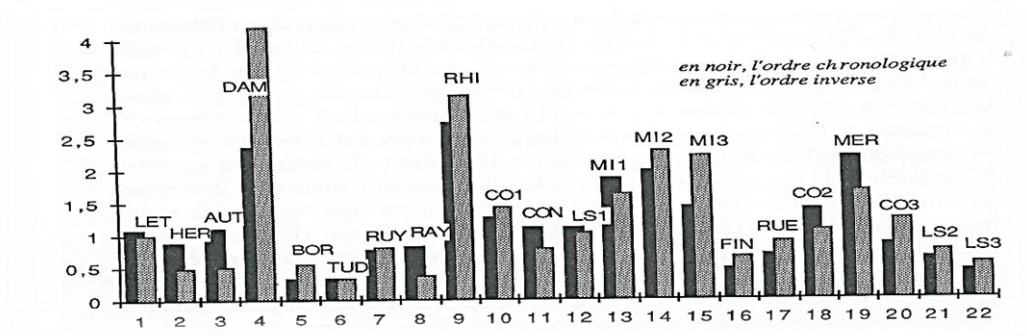
Aux pessimistes incrédules qui ont besoin de toucher l'objet du doigt nous proposons un contrôle très simple et presque sans calcul. Le rapport qui lie V à *N* et qui mesure la richesse lexicale n'est pas un simple quotient et il est difficile de le soustraire à l'influence de l'étendue des textes. Mais dans certains cas favorables, le simple bon sens - aidé par la démonstration de Ch. Muller - suffit à trancher en faveur du texte *a* et à estimer son vocabulaire plus riche, soit parce qu'il a plus de vocables que le texte *b* tout en ayant moins d'occurrences, soit parce qu'il a une fréquence moyenne plus faible, tout en ayant une étendue plus grande. Or sur les 231 confrontations possibles ($n^2/2 = 231$ pour n=22), 70 aboutissent dans notre corpus à une conclusion certaine, dont aucune n'est controuvée par les deux méthodes précédemment employées.

- II -

L'accroissement du vocabulaire

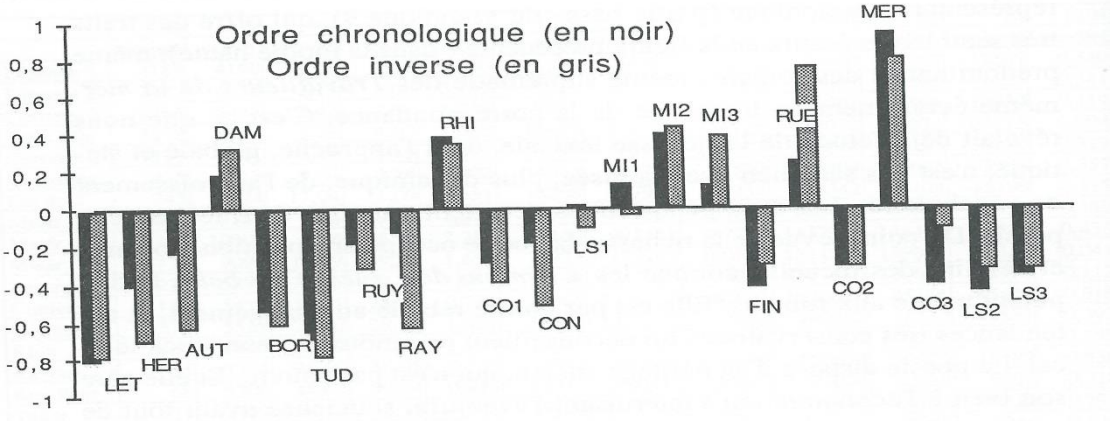
Les faits de structure lexicale peuvent être observés selon un point de vue un peu différent, qui s'attache à relever l'accroissement du vocabulaire, la situation étant celle d'un lecteur méthodique qui dépouillerait l'un après l'autre dans l'ordre chronologique les textes de Hugo, en notant à chaque étape les mots nouveaux, qui n'ont pas d'exemple dans la production antérieure. Les effectifs obtenus, qui vont nécessairement en s'amenuisant du premier au dernier texte de la chronologie, sont reproduits dans le tableau 1. Pour les apprécier, il faudrait disposer d'un modèle. Ce modèle peut encore être fourni par la loi binomiale, mais on a démontré qu'une distorsion était à craindre relativement au premier texte de la série. Nous avons donc préféré recourir à des méthodes d'approximation expérimentale. La première établit le rapport *V/Ac* (vocabulaire/accroissement) dont la croissance est d'abord ajustée par une fonction puissance, puis comparée à la courbe théorique. On peut suivre le progrès de ce calcul dans le graphique *5.*

Graphique 5. L'accroissement du vocabulaire



Une seconde méthode est fondée sur notre indice *w*, les données livrées au calcul étant cumulées, pour *N* comme pour *V*. On obtient là encore des effectifs théoriques et des écarts, lesquels donnent lieu à une représentation graphique (figure 6). Celle-ci offre des traits très semblables à ceux de la figure précédente: même prédominance des romans, même suprématie des *Travailleurs de la mer,* même écrasement du théâtre et de la correspondance. C'est ce que nous révélait déjà l'étude de la richesse lexicale, dont l'approche, globale et statique, n'est pas sans lien avec la visée, plus dynamique, de l'accroissement du vocabulaire. Les deux perspectives s'écartent toutefois dans le cas de la poésie. Du point de vue de la richesse, la poésie occupe une position moyenne et certains des recueils, comme les *Chansons des rues et des bois,* le disputent même aux romans. Elle est par contre rebelle au changement, et ses

Figure 6. Accroissement. Méthode de l'indice *w*



tendances très conservatrices lui déconseillent de renouveler son stock lexical. La poésie dispose d'un héritage ancien, qui n'est pas précisément pauvre. Mais elle gère son bien, comme font les pauvres, à l'économie, en s'interdisant l'aventure, en se souciant avant tout de maintenir le dépôt sacré.

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 des graphiques 5 et 6, à partir des données relevées dans le tableau 1. On remarquera toutefois que, dans cette perspective inversée, le dernier recueil de la *Légende des siècle* offre plus de surprises que le premier et le troisième sous-ensemble des *Misérables* plus de nouveautés que le début du roman, ce qui prouve l'homogénéité de ces deux textes.

- III -

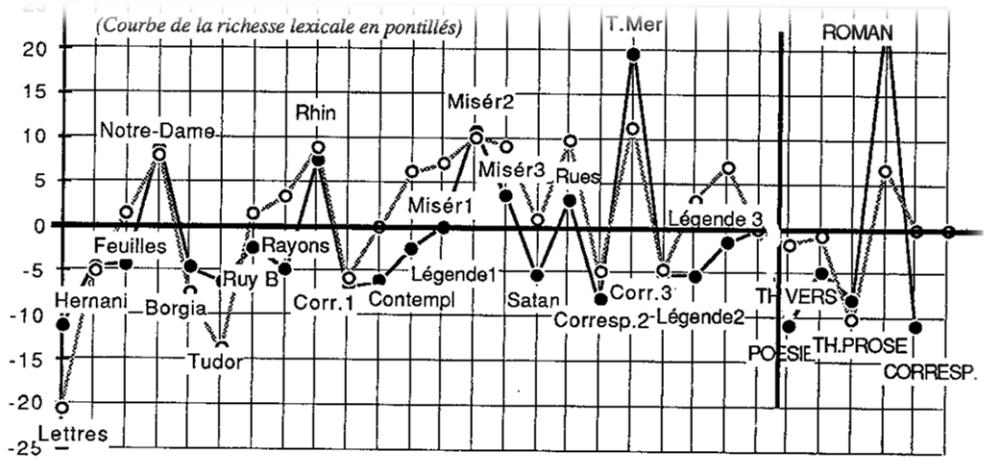
Les hapax et les mots rares

1 - On a poussé plus loin encore l'expérience, en relevant non pas seulement les mots nouveaux, mais les mots exclusifs de chaque texte, ceux qui n'ont qu'une occurrence chez Hugo et qui n'en ont pas davantage dans le corpus XIX-XXe du Trésor de la langue française[1]. 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. Nous sommes ici aux marges du dictionnaire, là où convergent les mots étrangers, les noms propres, les régionalismes, l'argot, les mots techniques ...et aussi les mots estropiés qui souffrent d'une lettre tordue ou perdue. Comme Hugo est assez friand de couleur locale, d'argot, de termes techniques, de patois, ces vocables ont un effectif non négligeable, et il a fallu les examiner l'un après l'autre. Nous en avons retenu près de 500. Ici encore l'audace lexicale de Hugo se déploie là où les risques sont les moindres, c'est-à-dire dans les genres qui imposent le moins de règles: roman et récit de voyage. La plupart des 454 hapax répertoriés appartiennent au *Rhin* (couleur locale), *à Notre-Dame de Paris* (vieux mots), aux *Misérables* (argot, termes militaires, mots techniques), et surtout aux *Travailleurs de la mer (*régionalismes, termes marins). Ni la correspondance, ni le théâtre, ni la poésie n'ont de part appréciable dans l'effectif des hapax. *Hernani* n'en contient aucun, non plus que les *Rayons et les ombres* . On n'en rencontre que deux au théâtre et 42 parmi les vers, dont 30 dans la seule *Légende des siècles.*[2] L'accueil fait aux hapax est donc déterminé par le genre littéraire. Il diffère aussi selon la nature grammaticale: plus de 400 de ces hapax appartiennent à la classe nominale; 289 sont des substantifs, 96 des adjectifs et 34 hésitent entre les deux statuts. Quant aux verbes on n'en relève que 34 (et 8 adverbes). Ce ne sont pas là les proportions qui régissent le partage habituel des parties du discours.

[1] Cela ne signifie pas que ces hapax ne se rencontrent pas à l'extérieur du corpus du TLF. La notion d'hapax est relative à un corpus. Celui qui nous sert de référence est vaste, mais il n'englobe qu'une faible partie des réalisations de la langue. Il est d'ailleurs assez difficile d'accorder à un mot une datation absolue ou une exclusivité radicale. En cette matière les attestations qu'on donne n'ont qu'une portée provisoire.

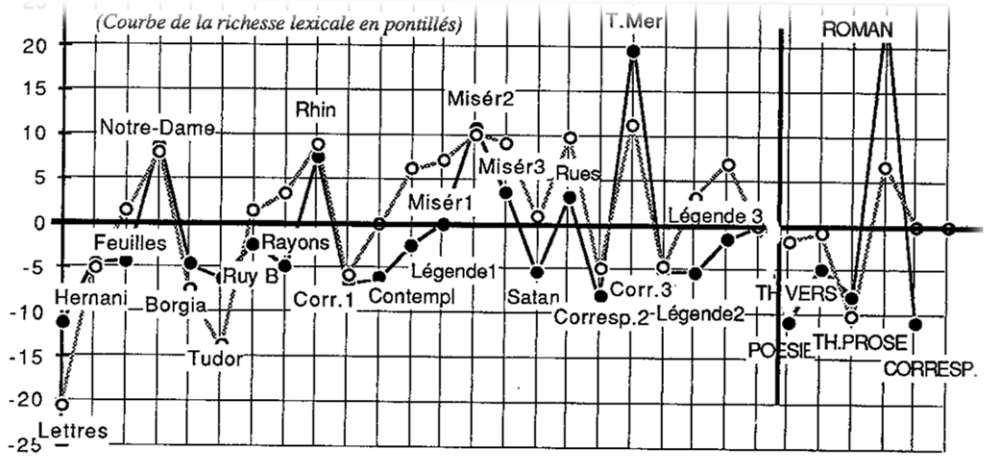
[2] Ces hapax ont un statut qui les assimile plutôt à des étrangers qu'aux membres de la tribu. En les mêlant au discours, Hugo produit plus d'étrangeté que de familiarité. Et sa plume accepte plus volontiers le dépaysement que le déclassement. Le "mot propre" n'est donc pas un "rustre" malpropre, mais plus volontiers l'ambassadeur d'une contrée lointaine. Ainsi se justifient - nous égrenons le début de la liste - ces termes bizarres dont Hugo parsème la *Légende des siècles : atèle, andryade, aragonal, argiraspide...*, et qu'on trouverait difficilement dans les dictionnaires, même encyclopédiques

Tableau 7. Les hapax de Hugo dans le corpus du TLF



2 Le mot hapax peut être entendu d'une autre façon et désigner les mots qu'on ne rencontre qu'une fois chez Hugo, sans examiner leur fréquence dans le corpus du TLF. Il s'agit là d'un sur-ensemble des hapax précédents. La contrainte extérieure ayant disparu, le filtre admet dix fois plus d'éléments, soit 4482 ou 22% du vocabulaire. Rien ne change pourtant dans la courbe, les romans et le *Rhin* s'élevant seuls dans la zone positive où les *Travailleurs de la mer* se maintiennent au sommet. Le graphique **8** reproduit les grands traits du précédent et l'on distingue à peine de faibles différences: la remontée dans la zone médiane de la *Légende des siècles* et l'accès des *Chansons à* la zone des excédents. Et la répartition des parties du discours reste toujours très favorable à la classe nominale : 2092 substantifs, 1410 adjectifs et 264 substantifs-adjectifs représentent 84% de l'effectif, ce qui est proche du partage léonin (de 91%) que la même classe nominale imposait parmi les hapax au sens étroit.

Figure 8. Les mots employés une fois chez Hugo



3 - En élargissant encore les mailles du filtre, on peut recueillir plus d'éléments qui peuvent prétendre au statut d'hapax, à condition de considérer chaque texte isolément. Mais on parlera plus volontiers de la sous-fréquence 1. Les effectifs sont considérables: plus de 3000 éléments dans les *Travailleurs de la mer* et autant dans *Notre-Dame* et les *Misérables.* La classe de fréquence 1 est donc mieux représentée, sinon en valeur absolue, du moins en pourcentage, dans les parties que dans l'ensemble: on comparera la proportion des *Travailleurs de la mer (* 3416 / 8824 = 0,387) ou, à l'autre extrême, des *Lettres à la fiancée (* 1260 / 3731 = 0,337), à celle du corpus Hugo(4482 / 20602 = 0,217) et à celle du corpus TLF (21198 / 71640 = 0,295)[1]. L'application de la loi binomiale permet de s'affranchir des distorsions dues aux différences d'étendue et de calculer un effectif théorique auquel le rée1 sera confronté. La formule est ici assez simple:

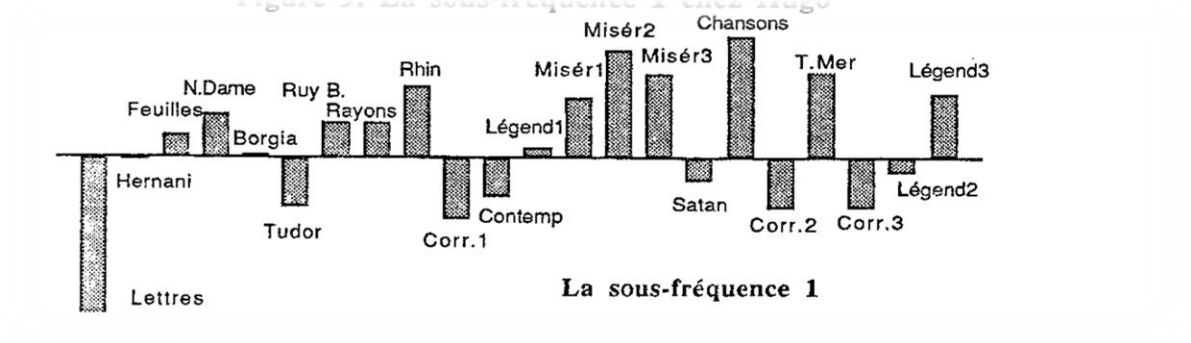
$$V1 = \Sigma\ V_i\ i\ p\ q^{i-1}$$

pour V_i = l'effectif des différentes classes de fréquence du corpus

p = la proportion du texte dans le corpus, *q* = le complément de *p à* l'unité

Les résultats apparaissent dans le tableau 1 et l'histogramme associé (figure **9).** La ressemblance avec les courbes **7** et **8** est frappante mais les recueils poétiques retrouvent de la vigueur: Les *Feuilles d'automne,* les *Rayons et les ombres* et plus encore la *Légende des siècles* sortent de la médiocrité et se rapprochent du niveau des romans. Et l'on voit même les *Chansons des rues et des bois* atteindre les sommets, comme cela s'était produit pour la richesse lexicale. C'est qu'en effet l'importance de la sous-fréquence 1 est prépondérante dans les paramètres qui influent sur V (elle en représente plus du tiers) et donc sur la richesse lexicale. On voit ainsi que, des hapax à la fréquence 1, et de la fréquence 1 à la sous-fréquence 1, on franchit par paliers l'espace qui sépare accroissement et richesse.

Figure 9. La sous-fréquence 1 chez Hugo



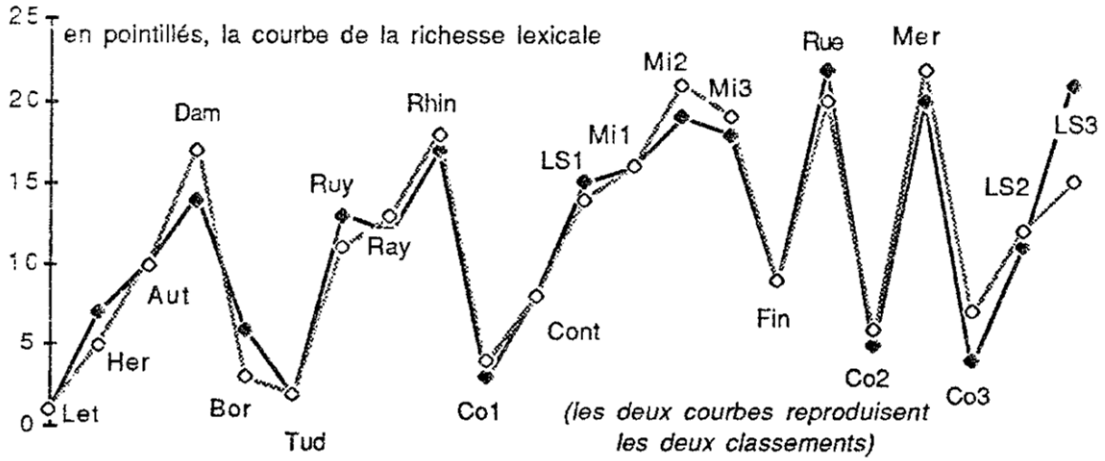
[1] La proportion des mots de fréquence 1 tend à décroître à mesure que s'étend la taille du corpus, comme si le réservoir des mots avait des limites, quoiqu'on ne l'épuise jamais. Or, contrairement à ce qu'on attendrait, la proportion relevée chez Hugo est moindre que celle du TLF, dont la taille est 35 fois supérieure. Le fait se voit aussi chez Proust (5034 / 18322 = 0,274) et, plus net encore, chez Zola (3750 / 19337 = 0,193). L'explication semble devoir être cherchée dans la nature, homogène ou hétérogène, des corpus considérés. Les monographies d'auteurs constituent une unité organique où les textes ne sont pas pleinement cumulatifs: même si les sujets ou les genres diffèrent, les productions d'un même écrivain ont une zone lexicale commune, et la part privative (c'est-à-dire la fréquence 1) tend à être plus étroite. Le corpus du TLF au contraire est fait de l'assemblage d'écrivains divers, de la superposition d'époques et de genres différents et la fréquence 1 témoigne de cette disparate.

4 - Un pas encore et l'on se rapproche de si près de la courbe de la richesse qu'on ne voit plus de différence. C'est ce qui se produit lorsqu'on envisage les *hapax de croisement.* On entend par là les mots qui n'ont qu'une occurrence dans l'ensemble formé par deux textes. Prenons un exemple: celui du croisement des *Contemplations* et des *Travailleurs de la mer.* I1 y a dans ce couple 3702 mots de fréquence 1, dont 1051 dans les *Contemplations* et 2651 dans les *Travailleurs.* Une simple règle de trois, fondée sur l'étendue respective des deux textes, laissait prévoir la répartition:1465 d'un côté et 2237 de l'autre. L'écart est donc de 414 en faveur des *Travailleurs.* Le calcul, répété autant de fois qu'on trouve de couples, soit 231 fois, engendre le tableau 10. Chaque colonne y est consacrée à un texte particulier et enregistre ses victoires ou ses défaites dans les 21 duels où il est engagé. Les *Lettres à la fiancée* ont toujours le dessous, avec un passif de -12216 points. A l'opposé les *Chansons des rues et des bois* l'emportent sur tous les autres concurrents, avec un score de +7408. Convertissons ces résultats en classement et opérons le rapprochement avec le classement de la richesse lexicale; on obtient le graphique du tableau 11 où la superposition des deux courbes est remarquable (le coefficient de corrélation est de 0,95).

Tableau 10. Effectifs des *hapax de croisement*

	Let	*Her*	*Aut*	*Dam*	*Bor*	*Tud*	*Ruy*	*Ray*	*Rhi*	*Co1*	*Con*	*LS1*	*Mi1*	*Mi2*	*Mi3*	*Fin*	*Rue*	*Co2*	*Mer*	*Co3*	*LS2*	*LS3*
LET	0	317	471	658	260	154	545	588	691	408	497	711	718	871	818	528	953	367	889	398	585	789
HER	-317	0	120	146	-23	-138	144	156	157	-94	10	169	123	174	168	57	357	-108	159	-90	87	261
AUT	-471	-120	0	61	-116	-225	37	65	78	-250	-16	119	50	91	74	-18	279	-232	93	-263	42	167
DAM	-658	-146	-61	0	-165	-240	-58	-5	72	-406	-217	3	4	230	110	-94	286	-538	262	-354	-148	133
BOR	-260	23	116	165	0	-91	178	147	135	-68	-32	126	142	198	177	40	339	-85	160	-72	44	244
TUD	-154	138	225	240	91	0	298	275	221	65	102	259	232	268	275	156	481	30	301	57	165	392
RUY	-545	-144	-37	58	-178	-298	0	26	56	-301	-139	29	68	132	115	-97	247	-279	110	-298	-54	159
RAY	-588	-156	-65	5	-147	-275	-26	0	58	-333	-39	71	6	76	56	-95	244	-328	58	-321	2	113
RHI	-691	-157	-78	-72	-135	-221	-56	-58	0	-509	-266	-62	-61	144	67	-134	210	-606	177	-417	-218	100
CO1	-408	94	250	406	68	-65	301	333	509	0	208	457	464	648	582	277	691	6	670	43	307	554
CON	-497	-10	16	217	32	-102	139	39	266	-208	0	234	250	406	349	-18	350	-258	414	-166	143	225
LS1	-711	-169	-119	-3	-126	-259	-29	-71	62	-457	-234	0	14	126	98	-231	193	-526	192	-409	-97	13
MI1	-718	-123	-50	-4	-142	-232	-68	-6	61	-464	-250	-14	0	260	129	-119	233	-591	238	-421	-156	116
MI2	-871	-174	-91	-230	-198	-268	-132	-76	-144	-648	-406	-126	-260	0	-127	-202	130	-791	36	-581	-310	23
MI3	-818	-168	-74	-110	-177	-275	-115	-56	-67	-582	-349	-98	-129	127	0	-188	159	-709	123	-519	-281	83
FIN	-528	-57	18	94	-40	-156	97	95	134	-277	18	231	119	202	188	0	353	-309	251	-247	159	230
RUE	-953	-357	-279	-286	-339	-481	-247	-244	-210	-691	-350	-193	-233	-130	-159	-353	0	-685	-178	-652	-258	-130
CO2	-367	108	232	538	85	-30	279	328	606	-6	258	526	591	791	709	309	685	0	824	-12	392	547
MER	-889	-159	-93	-262	-160	-301	-110	-58	-177	-670	-414	-192	-238	-36	-123	-251	178	-824	0	-592	-350	14
CO3	-398	90	263	354	72	-57	298	321	417	-43	166	409	421	581	519	247	652	12	592	0	255	514
LS2	-585	-87	-42	148	-44	-165	54	-22	18	-307	-143	97	156	310	281	-159	258	-392	350	-255	0	83
LS3	-789	-261	-167	-133	-244	-392	-159	-113	-100	-554	-225	-13	-116	-23	-83	-230	130	-547	-14	-514	-83	0
somme	-12216	-1518	555	1990	-1626	-4117	1370	1684	3043	-6395	-1821	2743	2321	5446	4223	-575	7408	-7393	5707	-5685	22	4630
rang	22	16	13	9	17	21	10	11	6	20	15	8	7	4	5	14	1	18	3	19	12	2
riches.	22	18	13	6	20	21	12	10	5	19	15	9	7	2	4	14	3	17	1	16	11	8

Figure 10. Les hapax de croisement dans le corpus Hugo.



- IV -
Les groupes de fréquence

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 127, de 128 à 255, de 256 à 511, de 512 à 1023, etc.), 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). On trouvera dans le tableau 11 les effectifs observés dans le corpus Hugo. Précisons que ces jalons (128, 256, etc..) sont posés dans le grand corpus du TLF. Le premier groupe par exemple concerne les fréquences très rares; car la fréquence 128 dans une masse de 70 millions de mots est l'équivalent de la fréquence 3 dans le corpus Hugo ou de la fréquence 0 dans un texte isolé. Mais avant d'exploiter l'ensemble des groupes, on peut se faire une idée de leur répartition en les regroupant en quatre lots: fréquences hautes, moyennes, basses et très basses. Le niveau choisi est celui des occurrences. Si l'on regroupe en même temps les histogrammes qui relèvent du même genre littéraire, on obtient la représentation très claire du graphique 12.

Considérons la première série, dévolue au roman[1]. Ce qui saute aux yeux c'est le déficit constant des fréquences moyennes, et l'excédent dans les fréquences basses et surtout très basses, mais aussi dans les fréquences hautes. Cet afflux des fréquences basses s'accorde avec ce que nous savons des hapax. La distribution des fréquences en poésie offre moins de cohésion, même si l'excédent des fréquences basses et le déficit des fréquences hautes sont observés dans tous les recueils. On voit en effet s'opposer deux périodes, les trois premiers recueils *(Feuilles d'automne, Les rayons et les ombres, Contemplations)* se distinguant des quatre derniers *(Chansons des rues* et *Légende des siècles),* tandis que la *Fin de Satan* hésite entre les deux camps. Le centre de gravité est placé plus haut dans les premiers recueils, au bénéfice des fréquences moyennes et aux dépens des fréquences extrêmes. Il s'abaisse dans les derniers recueils, avec un affaissement des fréquences moyennes et la préférence donnée aux

fréquences basses et très basses. Les histogrammes consacrés au théâtre et au genre épistolaire basculent de l'autre côté de la gamme des fréquences et offrent de grandes ressemblances: la rareté des classes basses et très basses, l'afflux des classes moyennes et, à un moindre degré, des hautes fréquences.

Tableau 11. Les groupes de fréquences. Effectifs

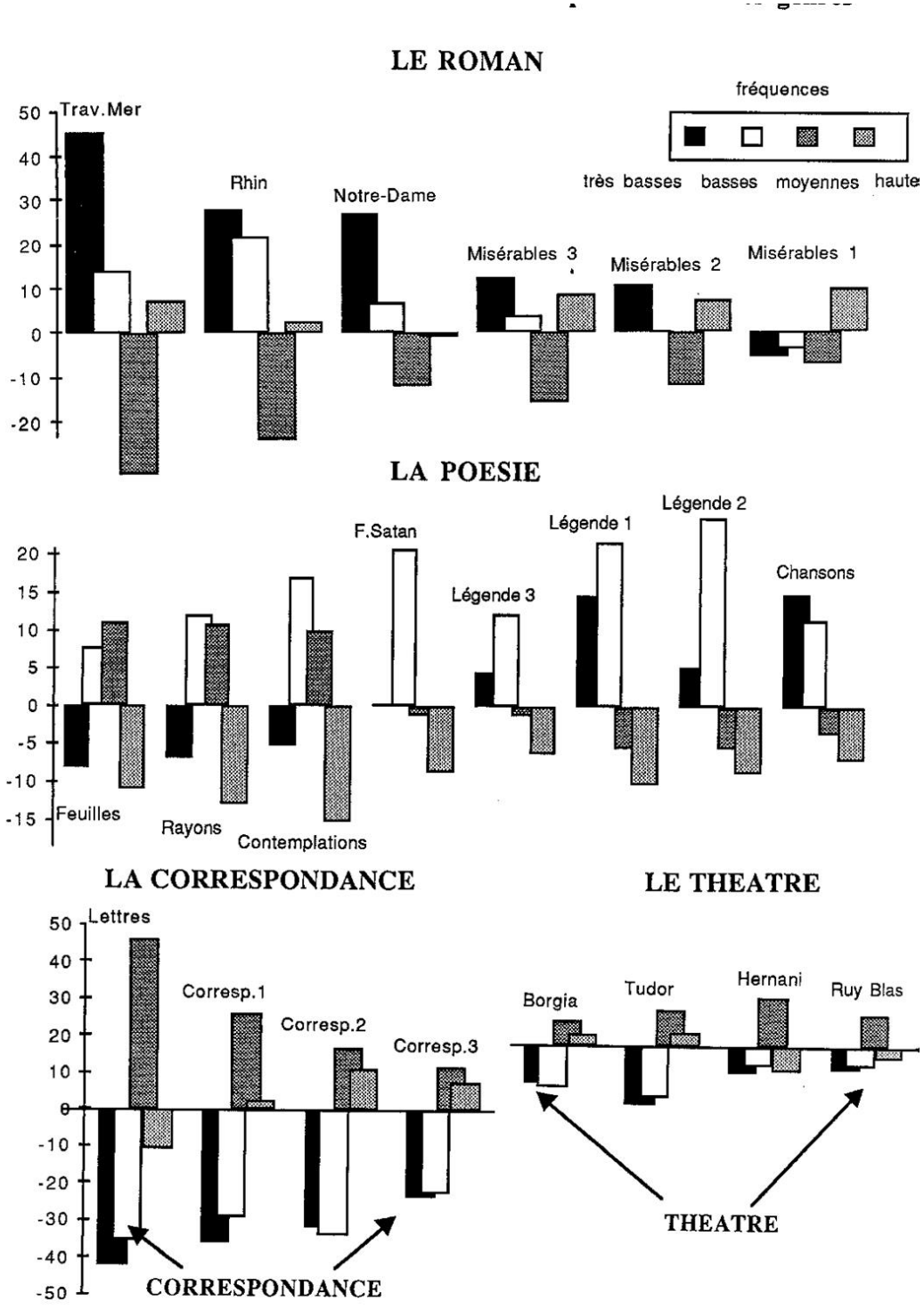
OCCURR.	*Lettres*	*Hern*	*Feuil*	*N.Dame*	*Borgia*	*Tudo*	*Ruy B*	*Rayons*	*Rhin*	*Corr.1*	*Contemp*
fréq <128	607	246	274	4166	264	293	535	364	4726	1147	1315
fréq< 256	263	141	220	2227	128	117	187	268	2684	530	870
fréq< 512	639	420	358	4007	232	182	392	410	4285	1006	1467
fréq< 1024	1090	436	718	4584	344	339	604	862	5746	1814	2490
fréq< 2048	2292	660	975	6884	524	589	906	1216	8064	2973	3843
fréq< 4096	3097	1069	1364	8282	804	1002	1381	1636	10707	4466	4682
fréq< 8192	5222	1265	1711	9566	999	1144	2027	1956	12360	6664	6103

OCCURR.	*Lettres*	*Hern*	*Feuil*	*N.Dame*	*Borgia*	*Tudo*	*Ruy B*	*Rayons*	*Rhin*	*Corr.1*	*Contemp*
fréq< 16384	5900	1457	1608	9585	1097	1326	1785	2028	11489	7251	5501
fréq< 32768	5585	1691	1871	11874	1407	1651	2303	2150	13277	7833	6160
fréq< 65536	8395	1936	1994	12297	1729	2138	2214	2291	10724	10062	6354
fréq< 1310728	695	1490	1297	8434	1426	1807	1622	1349	8630	8383	4048
fréq< 2621449	483	1834	1784	10496	1571	2085	2008	1925	10848	8623	6140
fréq> 2621442	62210188	10692	84113	10194	12274	13904	12142	96259	56650	38854	
total	93890	22833	24866	176515	20719	24947	29868	28597	199799	117402	87827

OCCURR.	*Légend1*	*Misér1*	*Misér2*	*Misér3*	*Satan*	*Chanson*	*Corr.2*	*T.Mer*	*Corr.3*	*Lég 2*	*Lég 3*	*total*
fréq <128	1567	3229	3985	3591	722	782	2461	4034	1523	1549	671	38051
fréq < 256	1019	1663	1990	1848	527	500	1094	2273	618	1095	415	20677
fréq < 512	1668	2794	3527	3168	950	723	1955	3309	1138	1899	744	35273
fréq< 1024	2430	4348	4589	4558	1504	993	3288	3756	2186	2839	1129	50647
fréq< 2048	3542	6057	6561	5842	2386	1356	5093	5265	3029	4360	1642	74059
fréq< 4096	3969	8309	8869	7790	2755	1696	6680	6970	3944	5184	1867	96523
fréq< 8192	4628	10356	10486	9043	3274	2096	10952	7578	6919	5855	2250	122454
fréq< 16384	4361	10262	10364	8603	2741	1894	11567	6703	6774	5235	2140	119671
fréq< 32768	5146	12164	11504	10753	3444	1974	12226	8030	7148	6543	2433	137167
fréq< 65536	4624	11717	12666	10261	3071	1977	13967	6651	8391	5649	2372	141480
fréq< 1310723	225	9102	9481	8255	2139	1450	11837	5912	6569	4128	1681	110960
fréq< 2621444	801	11364	11790	10494	3005	1911	12637	7698	6864	5919	2417	135697
fréq> 2621443	84089307	91912	81663	21719	14228	92265	65956	53811	42508	16526	991627	
total	74820	180672	187724	165869	48237	31580	186022	134135	108914	92763	36287	2074286

[1] Comme le *Rhin* montre un schéma de distribution identique à celui des romans, nous l'avons classé dans cette catégorie. La plupart du temps c'est la catégorie qui convient le mieux à ce récit de voyage.

Graphique 12. La distribution des fréquences dans les genres.



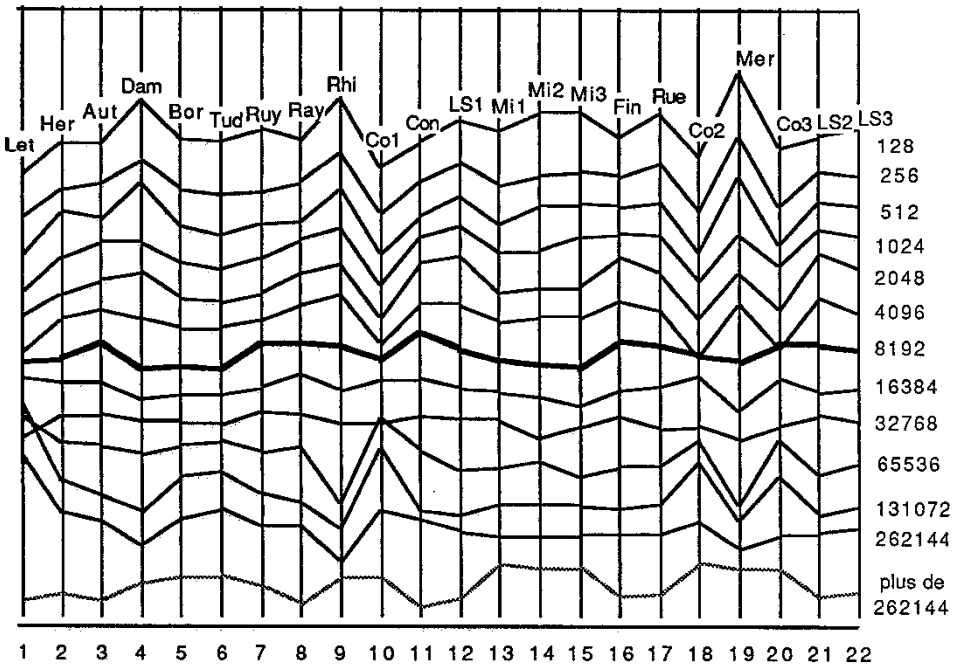
Mais on peut souhaiter rejoindre le tableau 11, avec ses 13 lignes et ses 22 colonnes, et l'étudier dans son intégralité, sans simplification. On commencera par considérer une 23e colonne qui totalise les 22 premières, tous textes confondus, et rend compte du choix global de Hugo dans la gamme des fréquences. Comme on peut le constater dans le tableau 13, ce choix est dirigé vers les mots rares, surtout ceux qui ont moins de 128 occurrences dans le TLF. L'écart réduit est ici très significatif: +79 et il le reste généralement dans les classes suivantes, jusqu'à la fréquence 2048. Au delà de cette limite les écarts tendent à se maintenir dans la zone négative (6 fois sur 8). Le goût de Hugo pour les mots rares se trouve donc confirmé.

Tableau 13 . Les groupes de fréquences.

fréquence	*hugo*	*corpus*	*théorique*	*écart*	*réduit*
<128	38051	866315	25439,08	12611,92	79,07
<256	20677	710948	20876,77	-199,77	- 1,3 8
<512	35273	1131272	33219,46	2053,54	11,27
<1024	50647	1669784	49032,70	1614,30	7,29
<2048	74059	2461202	72272,46	1786,54	6,65
<4096	96523	3481268	102226,39	-5703,39	-17,84
<8192	122454	4282559	125756,06	-3302,06	-9,31
<16384	119671	4459394	130948,76	-11277,76	-31,17
<32768	137167	4441477	130422,64	6744,36	18,68
<65536	141480	4884506	143432,05	-1952,05	-5,15
<131072	110960	3862455	113419,83	-2459,83	-7,30
<262144	135697	4395316	129067,13	6629,87	18,45
>262144	991627	33992266	998172,67	-6545,67	-6,55

A l'intérieur du corpus Hugo, le mouvement des fréquences s'inscrit dans les 22 colonnes du tableau 11. Là aussi il convient d'abord de neutraliser l'influence de l'étendue, et de convertir les effectifs bruts en écarts réduits, puis en courbes. La figure 14 superpose les 22 profils obtenus. Visiblement les groupes de fréquences, comme les couleurs du spectre, forment un continuum et les mouvements s'inversent progressivement en passant de la première image à la dernière. Dans ce mouvement décomposé où 22 images instantanées restituent une sorte de ralenti, ce qui est convexe devient concave et inversement. On observera particulièrement l'onde qui se propage au niveau des *Travailleurs de la mer* et qui est d'autant plus accusée que ce texte voisine et contraste, dans la chronologie, avec deux recueils de correspondance: la pyramide qui signale l'excédent des fréquences basses dans ce roman, fléchit et s'aplatit dans les fréquences moyennes, avant de se transformer en pyramide inversée dans les fréquences hautes. Des ondes semblables sont visibles dans *Notre-Dame* et le *Rhin* et des ondes contraires dans les *Lettres* et le premier recueil de correspondance.

Figure 14. Le mouvement progressif des groupes de fréquence



Il est un autre moyen de synthétiser un tableau de grandes dimensions afin d'en faire apparaître les lignes de force, c'est l'analyse factorielle. Cette procédure mathématique a rencontré un si grand succès, depuis que la longueur des calculs ne fait plus problème, qu'il n'est pas nécessaire d'expliquer le fonctionnement d'un outil si commun. Nous renvoyons au promoteur de la méthode, Jean-Paul Benzécri. Bornons-nous à présenter le résultat d'une telle analyse dans le graphique 15 .

L'interprétation ne fait pas difficulté dans le cas présent. Les groupes de fréquences 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 1024, 4096 et 2048), puis à droite (classes 8192, 32768, 16384). Ici, parmi les fréquences basses et 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 (262144, 65536, 131072), 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 la gamme intermédiaire. 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.

On aurait pu croire que des deux facteurs qui se disputent la souveraineté sur les mots, le temps et le genre, le premier l'emporterait sur le second chez Hugo, écrivain précoce, qui meurt octogénaire, dont les productions s'échelonnent sur plus de soixante ans, et dont l'évolution parcourt le long chemin qui va de Louis XVIII à Louis Blanc. Il n'en est rien. Hugo a beau prétendre secouer le joug des genres et mêler les registres, comme ils se mêlent dans le théâtre de Shakespeare et dans la vie, c'est pourtant le genre qui prévaut chez lui, dans la structure de son vocabulaire. Il prévaut aussi dans le contenu lexical, mais c'est là une autre histoire.

Figure 15. Analyse factorielle des classes de fréquences