Hugométrie

Étienne Brunet

▶ **To cite this version:**

Étienne Brunet. Hugométrie. Colloque Victor Hugo, Université de CLUJ (Roumanie), Apr 2002, Cluj Roumanie. hal-01562323

HAL Id: hal-01562323
https://hal.science/hal-01562323

Submitted on 13 Jul 2017

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Etienne BRUNET
Institut de Linguistique française (CNRS)
Université de Nice

Hugométrie

Nous empruntons ce titre à un livre de Charles Muller qui vient de paraître, à l'occasion du Bicentenaire de la naissance du poète, et où le promoteur de la statistique linguistique appelle de ses voeux la réalisation d'une base hugolienne enfin complète[1]. Pourquoi ne pas faire pour Hugo ce qu'on a fait pour Voltaire Balzac, Dumas, Chateaubriand et beaucoup d'autres? Exploitant les données disponibles en 1988 à l'Institut National de la langue française, nous avions certes réalisé une étude, dont Muller avait écrit la préface. Mais Muller y regrettait l'absence de textes majeurs, non encore dépouillés : "Que bientôt le corpus de Nancy devienne le corpus de Hugo! Que *les Chants du crépuscule* viennent rejoindre les *Contemplations*; qu'avant 1993 *Quatre-vingt-treize* et les *Châtiments* prennent la place qu'ils méritent; que *l'Art d'être grand-père* ne reste pas en pénitence; que *Cromwell* et sa *Préface* sortent de l'ombre; et, malgré les aspérités des textes inachevés, que *Dieu* rejoigne *Satan*!"

[1] "En 1902, au moment du centenaire, il y avait de belles empoignades entre hugophiles et hugophobes, sans parler des intraitables et quelque peu sectaires hugolâtres. Aujourd'hui, le bicentenaire nourrit un hugotropisme contagieux; on publie même certain récit romanesque qui, ressuscitant Javert et prolongeant la vie de Cosette, peut être qualifié de hugogénique; ou est-ce du pseudo-Hugo? du para- ou péri-Hugo? du simili-Hugo? Mais notre temps produisant moins de hugolâtrie ou de hugomanie, vacciné contre toute hugopathie, allergique aux vaines polémiques, aux pichrocholines hugomachies, préfère une scientifique hugologie; on ne manque pas de savants et fervents hugologues, concepteurs ou inspirateurs de thèses hugologiques, dignes d'occuper une chaire de hugologie générale, ou candidats à des maîtrises de hugologie comparée ou appliquée; et, avec le secours de l'ordinateur, l'époque se doit de développer la hugoscopie; on envisage, pour la versification hugolienne, de formaliser une hugométrie, et, pour l'étude analytique des manuscrits, de créer une section de hugographie; tous travaux dont les produits convergeront sur les rayons de la hugothèque, secteur que la Grande Bibliothèque consacrera à la hugoculture." Charles Muller, *Mes rencontres avec Hugo*, coll. *La nuée bleue*, Éditions de l'Est, 2002, 220 p.

On verra dans le tableau 1 que ces lacunes (mis à part *Dieu*) ont été comblées et beaucoup d'autres que Muller avait omises. Les titres en caractères italiques correspondent aux textes qui ne figuraient pas dans notre précédente étude[2].

Tableau 1 La composition du corpus Hugo

	TITRE	OCCURRENCES	GRAPHIES	P	Q	CODE
1	Lettres	109345	6457	.0368	.9632	Le
2	Islande	148998	11168	.0501	.9499	Is
3	Jargal	67606	8129	.0227	.9773	Ja
4	Odes	67063	7850	.0225	.9775	Od
5	Orientales	26130	4691	.0088	.9912	Or
6	Hernani	35981	4026	.0121	.9879	He
7	Dame	216527	16491	.0728	.9272	ND
8	Feuilles	27447	4357	.0092	.9908	Fe
9	Borgia	27966	3321	.0094	.9906	Bo
10	Tudor	33785	3231	.0114	.9886	Tu
11	Voix	30139	4761	.0101	.9899	Vo
12	Ruyblas	36690	4306	.0123	.9877	RB
13	Rayons	31886	4872	.0107	.9893	Ra
14	Rhin	245094	19931	.0824	.9176	Rh
15	1corresp	142434	9816	.0479	.9521	C1
16	Petit	87769	10062	.0295	.9705	NP
17	Châtiments	64708	8774	.0218	.9782	Ch
18	Contemplations	111570	9621	.0375	.9625	Co
19	Satan	57953	6866	.0195	.9805	Sa
20	Shakespeare	10515	2716	.0035	.9965	Sh
21	Chansons	39241	6418	.0132	.9868	CR
22	2corresp	233737	13223	.0786	.9214	C2
23	Travailleurs	172660	15614	.058	.942	TM
24	Légende	258075	17045	.0868	.9132	LS
25	Terrible	66118	7969	.0222	.9778	Te
26	3corresp	142852	9814	.048	.952	C3
27	Treize	147709	12824	.0497	.9503	Tr
28	Grandpère	37215	5499	.0125	.9875	Gr
29	Crime	174309	13355	.0586	.9414	Cr
30	Ane	28645	5716	.0096	.9904	An
31	Vents	94279	10047	.0317	.9683	QV
	TOTAL	2974446	61282			

Les titres ajoutés sont ceux qu'on a trouvés disponibles sur Internet, non seulement sur le site *Gallica* qui propose surtout les textes de l'Institut National de la langue française, déjà explorés dans notre étude de 1988, mais aussi sur les sites fournisseurs de textes comme *ABU*, *ATHENA*, la bibliothèque de Lisieux, et quelques sites voués précisément à Victor Hugo, en particulier ceux de l'*IUFM* d'Amiens et de l'académie de Strasbourg. On peut regretter toutefois l'absence persistante de certaines oeuvres qui n'ont fait l'objet d'aucun dépouillement ou qui n'ont pas été mises

[2] Etienne Brunet, *Le Vocabulaire de Hugo*, 3 tomes, Slatkine-Champion, 1988.

à la disposition du public. L'occasion du bicentenaire était belle pour entreprendre un recensement complet, ce qu'on fera sûrement à l'occasion d'un prochain anniversaire. En attendant, le présent corpus permet de faire le tour de l'oeuvre de Hugo, sans lacunes trop criantes, mis à part l'oeuvre maîtresse des *Misérables*. Curieusement ce roman n'est lisible que dans le format image à la *BNF*, alors que l'*INaLF* l'a livré dans le format texte, et toutes les autres entreprises de saisie ont reculé devant la longueur du roman, sauf un site américain qui propose une traduction anglaise. On ne trouvera donc pas les *Misérables* dans la présente base. On ne regrette qu'à moitié cette lacune, non qu'on mésestime l'importance de ce chef d'oeuvre, mais ce texte est si long qu'il déséquilibre les comparaisons statistiques, ce dont avait souffert notre étude antérieure, où les *Misérables* avaient pris place[3].

L'accès à un corpus hugolien plus large ou plus pur[4] n'est pourtant pas notre préoccupation première. L'objectif principal était d'assurer un traitement plus élaboré où les mots ne seraient plus seulement des graphies, mais des vocables ou des lemmes. Sans doute la lemmatisation n'était pas tout à fait absente de notre monographie précédente, mais elle était grossière et expéditive. Car le regroupement des formes derrière l'entrée du dictionnaire avait été réalisé sans désambiguïser les homographes. Quand un homographe se présentait dans la liste de correspondance graphie-lemme, c'est la première proposition qui était retenue, sans autre examen. L'examen en effet ne peut s'exercer que dans le texte même, et c'est ce à quoi nous nous sommes employé ici, en recourant à un véritable logiciel d'étiquetage et de lemmatisation, dérivé de *Cordial*. Tous les textes du corpus ont été soumis à ce lemmatiseur connu sous le nom d'*Analyseur*. Le résultat de l'analyse s'inscrit dans un fichier où chaque graphie du texte génère une ligne contenant respectivement la localisation, la graphie, le lemme, le code grammatical et la fonction.

---

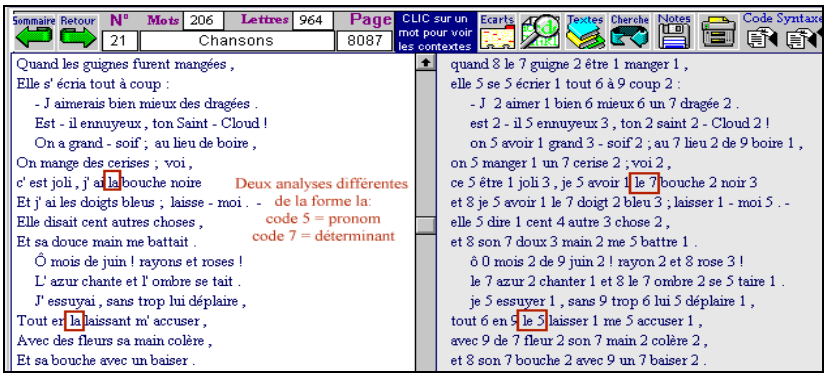
[3] Pour réduire les *Misérables* à une échelle commune, on avait divisé le texte en trois parties. Les *Misérables* sont de nouveau intégrés dans une base plus récente qu'on peut télécharger gratuitement sur internet sous le nom HUGOPROSE et à l'adresse http://logometrie.unice.fr/pages/bases/. À la même adresse une base HUGOVERS est aussi disponible, avec 14 recueils poétiques s'ajoutant à 30 textes de prose. Le logiciel HYPERBASE est également téléchargeable pour traiter n'importe quel ensemble de textes, et particulièrement ceux qui n'ont pas encore trouvé place dans les bases hugoliennes.

[4] Les sites où nous avons puisé les ressources textuelles ne précisent pas toujours les éditions qu'ils utilisent et ne garantissent pas la qualité du dépouillement. Comme le format "texte seulement" est encore le plus répandu dans le téléchargement, on perd en outre les informations relatives au hors texte, à la pagination et à la typographie. En particulier les italiques sont ignorés.

Figure 2. Le menu principal de la base hypertextuelle Hugo

Notre logiciel *Hyperbase* - dont le menu principal est reproduit dans la figure 2 - prend alors en compte les principaux éléments d'un tel fichier et les distribue séquentiellement dans quatre champs parallèles, voués respectivement aux formes, aux lemmes, aux codes et aux structures. La figure 3 met en correspondance les formes et les lemmes d'une même page des *Chansons des rues et des bois*.

Figure 3. L'alignement forme-lemme



On notera que les lemmes, dans la partie droite de l'écran, sont pourvus d'un indice numérique, afin de séparer les uns des autres les homographes. Ainsi *le 7* (dans *la bouche noire*) distingue l'article du pronom codé 5 (dans *la laissant*). Ces codes simplifiés qui reproduisent la classification de Muller et de Labbé (1 verbe, 2 substantif, 3 adjectif, 4 numéral, 5 pronom, 6 adverbe, 7 déterminant, 8 conjonction, 9 préposition) n'appartiennent pas en propre à *Cordial*, mais ont été dérivés de l'analyse complète fournie par *Cordial*.

Cette analyse complète est rendue visible, quoique peu lisible, pour peu qu'on sollicite le bouton *CODE* situé à droite de la barre de menu. Là aussi l'alignement est rigoureux, en sorte que l'on sait précisément à quel mot correspond telle ou telle analyse (voir figure 4). Les quatre champs sont sensibles au clic de la souris: tout objet que l'on désigne, qu'il s'agisse d'une forme, d'un lemme, d'un code ou d'une structure, renvoie aux autres occurrences où le même objet est rencontré, les relations hypertextuelles s'appliquant aux quatre champs. Mais ces relations lient aussi ces champs entre eux, en sorte qu'en cliquant sur un code grammatical dans le champ de droite (par exemple *_Vais3pV*, soit *Verbe, auxiliaire, indicatif, passé simple, 3e personne du pluriel, base de proposition*), on obtient successivement en vidéo inverse tous les verbes qui répondent strictement à ce codage (comme dans la figure 5).

L'indexation et toutes les opérations subséquentes sont alors répétées quatre fois, au niveau des structures syntaxiques (c'est-à-dire des séquences ordonnées des parties du discours), puis des codes grammaticaux, puis des lemmes, puis des formes. À l'issue de ce traitement, on obtient quatre index (figure 5) qui réagissent pareillement au clic de la souris. La forme (ou le lemme ou le code ou la structure) qu'on désigne montre le détail de ses occurrences, parmi lesquelles l'utilisateur fait son choix pour se référer au texte.

Figure 4. L'alignement forme-code

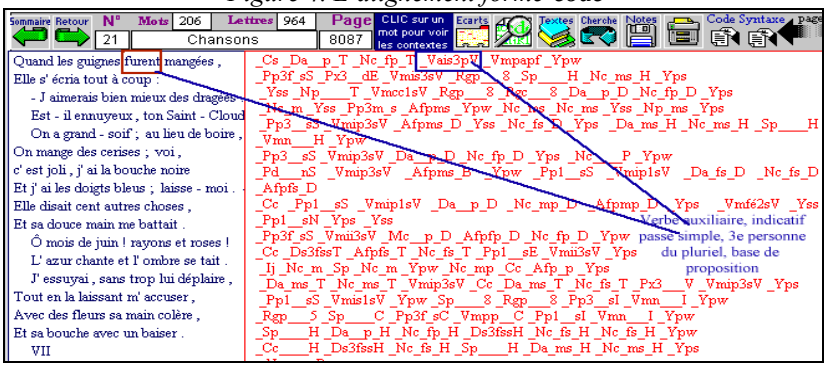
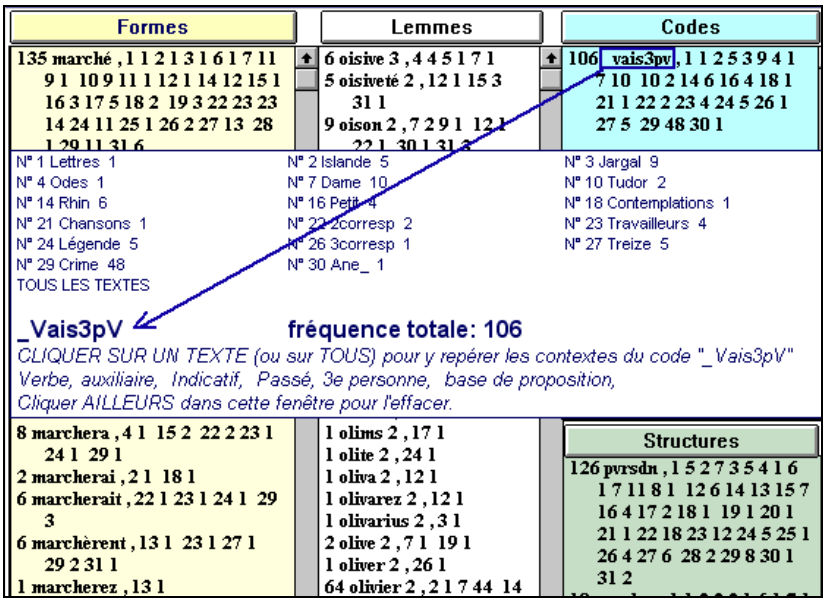
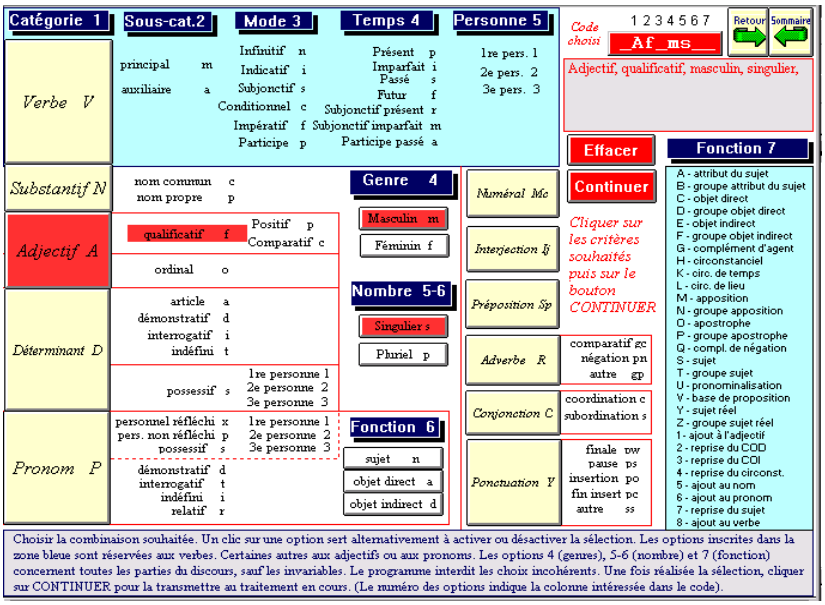


Figure 5. Les quatre index issus de Cordial



S'il s'agit d'un code, dont la signification peut être opaque, le décryptage est assuré et traduit en clair, comme dans l'exemple ci-dessus. Lorsqu'on veut, non plus reconnaître un code particulier, mais rassembler les mots qui ont reçu le même codage (par exemple pour constituer une concordance ou un histogramme), on est renvoyé à une page spéciale (figure 6) qui dénombre toutes les combinaisons possibles. Car *Cordial* pousse loin l'analyse, en relevant pour chaque mot la catégorie, la sous-catégorie, le genre, le nombre, la fonction et s'il s'agit d'un verbe le temps, le mode et la personne. Un clic dans une option provoque alternativement l'activation ou la désactivation correspondante. Certaines options sont impliquées ou exclues automatiquement, dès qu'une autre est choisie, de telle façon qu'il y ait toujours cohérence. Car il serait absurde de sélectionner le futur d'un substantif ou le féminin d'un verbe à l'infinitif. Chaque clic modifie le filtre dont l'affichage apparaît dans une fenêtre, en haut et à droite de l'écran, avec sa traduction en clair. Toute colonne non intéressée par la sélection est remplie par défaut par un joker, dont l'effet est d'admettre tout code qu'on rencontre à cet endroit. Ainsi dans l'exemple choisi la colonne 3 n'ayant pas été sélectionnée dans le code *Af_ms___*, tous les adjectifs seront retenus, quel que soit le degré, positif ou comparatif. De même le vide rencontré dans la colonne 7 laissera la sélection indifférente à la fonction dans la phrase.

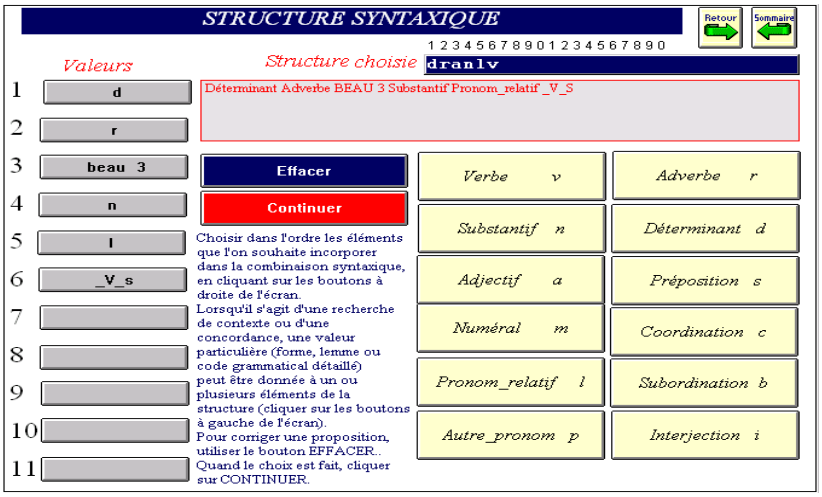
Figure 6. Le choix d'un code grammatical



Une fois que la sélection est faite, elle est transmise à la fonction appelante, qui délivre un contexte, une concordance, ou une liste, c'est-à-dire un tableau à deux dimensions dont chaque ligne dresse le profil d'une sélection grammaticale à travers le corpus et chaque colonne celui d'un texte du corpus à travers les codes grammaticaux.

Le choix des structures syntaxiques se fait de la même façon en choisissant dans l'ordre les catégories qui forment la séquence souhaitée, par exemple déterminant + adverbe + adjectif + substantif + pronom relatif, à chaque endroit de la séquence une constante (forme, lemme ou code grammatical détaillé) pouvant se substituer à la variable. Ainsi dans la figure 8 deux contraintes sont imposées à la sélection: l'adjectif doit appartenir au lemme *beau* et le verbe qui suit le relatif doit être au subjonctif (code *_V_s*). Si l'on ajoute que la fonction de chaque mot peut être isolée dans le code, et qu'un balisage sémantique est appliqué aussi au texte, on mesure combien l'étude, traditionnelle ou statistique, du texte s'enrichit de mille possibilités.

Figure 8. Choix d'une structure syntaxique



Ce n'est pas le lieu d'entreprendre l'exploitation d'un tel corpus lemmatisé, que nous préférons livrer dans sa pureté virginale aux spécialistes du poète. Là où l'étude des formes permet une première approche, celle des lemmes apporte le plus souvent une confirmation. Ainsi les 31 textes de notre corpus se répartissent de la même façon lorsqu'est mesurée la distance entre leurs vocabulaires, lemmatisés ou non. C'est le genre qui visiblement impose sa loi dans la figure 9, laquelle rend compte des lemmes, et la figure 10, qui s'en tient aux graphies. Précisons que le calcul, pour chaque paire de textes, prend en considération tous les mots rencontrés, en tenant compte de leurs fréquences respectives dans les deux textes comparés. La coupure y est brutale qui sépare la poésie, à droite, de la prose, à gauche[5], et dans la direction qui conduit à la prose, les chemins divergent selon qu'il s'agit d'essais, de romans, de théâtre ou de correspondance. Cependant la chronologie n'est pas complètement effacée par la domination du genre: elle peut réapparaître à l'intérieur d'un même genre, et se manifester dans la production poétique où trois périodes se succèdent.

[5] On notera l'exception du théâtre en vers où se combattent deux exigences: celle qui tient au théâtre et qui est dominante, et celle qui tient à la versification.

Figure 9. La distance lexicale calculée sur les lemmes

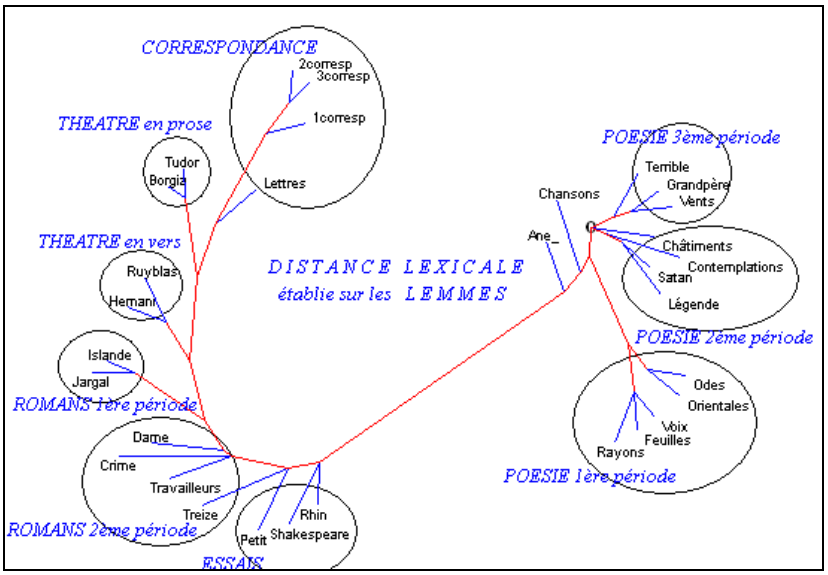


Figure 10. La distance lexicale établie sur les graphies

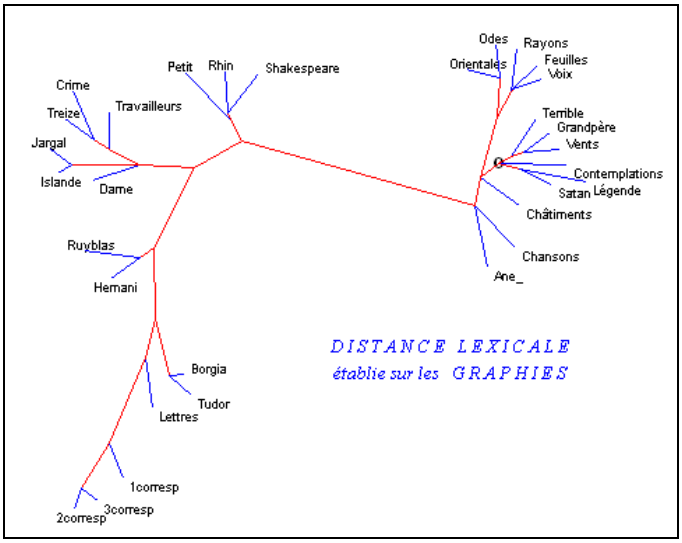
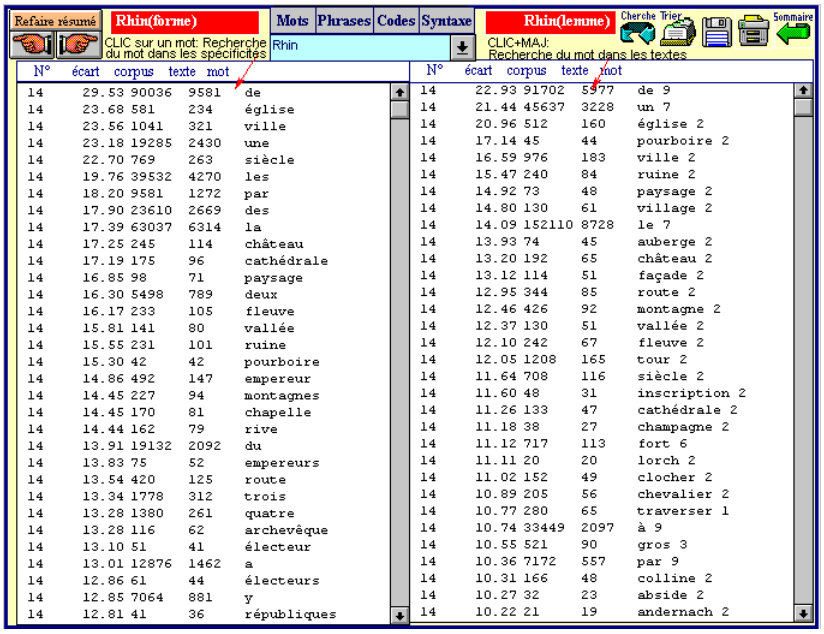


Figure 11. Les spécificités du Rhin (graphies à gauche, lemmes à droite)

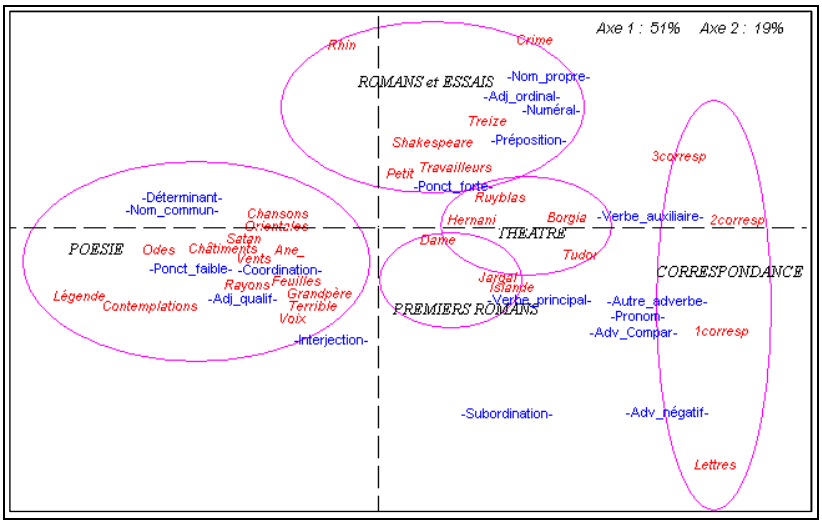


De même les listes de spécificités obtenues pour chaque texte à partir des formes et des lemmes ont beaucoup d'éléments communs (comme dans la figure 11, consacrée au *Rhin*), quoique les informations qu'elles donnent pour les verbes soient beaucoup plus sûres si la lemmatisation a opéré le regroupement des formes fléchies. Mais dès qu'on aborde la syntaxe et les faits de style, l'étude des formes montre vite ses limites et ses faiblesses: l'approche biaisée des mots grammaticaux ne permet qu'une approximation timide, car beaucoup de mots-outils sont homographes et servent à plusieurs usages. Sans lemmatisation préalable, bien des aspects du texte restent inaccessibles: le genre grammatical, le nombre, la fonction dans la phrase, le temps verbal, le mode, la personne, les parties du discours et leurs multiples combinaisons réalisées dans les syntagmes et les structures syntaxiques, des plus simples aux plus complexes, tout cela échappe aux formes brutes.

Ainsi les codes grammaticaux, dûment reconnus dans le texte même, peuvent donner lieu à une typologie des textes où, là encore, le genre fait la loi. Les camps rivaux sont toujours les mêmes : ils opposent, ici comme dans beaucoup de corpus précédemment analysés, le clan du verbe et

celui du nom. À droite, le verbe, qui annexe les catégories associées: pronoms, adverbes et subordonnants, établit son règne sur la correspondance, sur la scène et sur les premiers romans, jusqu'à *Notre-dame de Paris*. Les autres romans s'écartent vers le haut du graphique, en compagnie des essais.

Figure 12. Les catégories grammaticales. Analyse factorielle



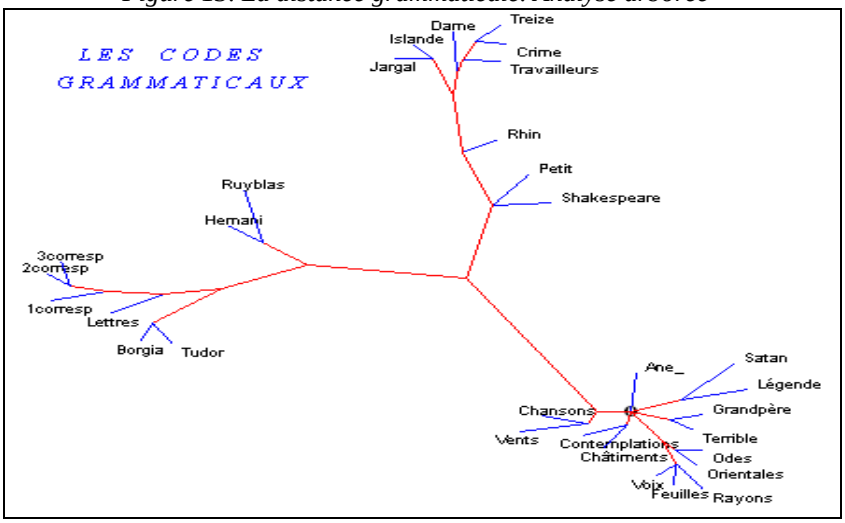
Là le paysage grammatical change, l'influence du verbe s'affaiblit, laissant la place aux noms propres, aux prépositions et aux circonstances qui font appel aux numéraux. Dès qu'on franchit la barrière verticale, on pénètre dans le territoire poétique, où tous les recueils sont concentrés et où les classes nominales (nom commun et adjectif) sont en faveur, ainsi que leurs acolytes (déterminants). Les conjonctions de coordination s'y trouvent aussi, pour des raisons de rhétorique ornementale. Quant à la ponctuation, elle est plus légère en poésie (virgule dominante) que dans la prose, qui accueille plus volontiers des signes de ponctuation forte.

Comme les codes grammaticaux ont été indexés dans leur grande variété (plusieurs milliers de codes différents), on a le moyen d'établir entre les textes une distance grammaticale, qui ignorant absolument le sens des mots n'est attentive qu'à leurs marques grammaticales. Seront rapprochés par le calcul les textes qui ont tendance à utiliser les mêmes codes. Rien n'oblige le graphique 13 à reproduire celui qu'on obtient à partir des lemmes (graphique 9), puisque la mesure porte sur des objets différents et en principe

indépendants, sémantique dans un cas, grammatical dans l'autre. Et pourtant les graphes que réalise l'analyse arborée sont presque superposables. On y lit la même opposition violente entre poésie et prose, mais la prose se diversifie plus nettement, une branche conduisant à l'essai, puis au roman, l'autre au théâtre et à la correspondance. Une fois de plus, c'est la carte des genres qui est reproduite. La révolution que Hugo prônait dans la *Préface* de *Cromwell* n'a donc en rien aboli la barrière stylistique et grammaticale qui sépare les genres. Hugo a beau mettre un "bonnet rouge au dictionnaire" et prétendre "appeler le cochon par son nom". Il est rare qu'il mêle en poésie les mots nobles et les mots bas. Ses audaces se vérifient en matière de versification, sans aller toutefois jusqu'au vers libre, mais elles sont timides en matière de lexique, et nulles dans le domaine grammatical, où les spécificités des genres établis sont par lui respectées et exploitées. "Paix à la syntaxe" proclame-t-il dans *Réponse à un acte d'accusation*.

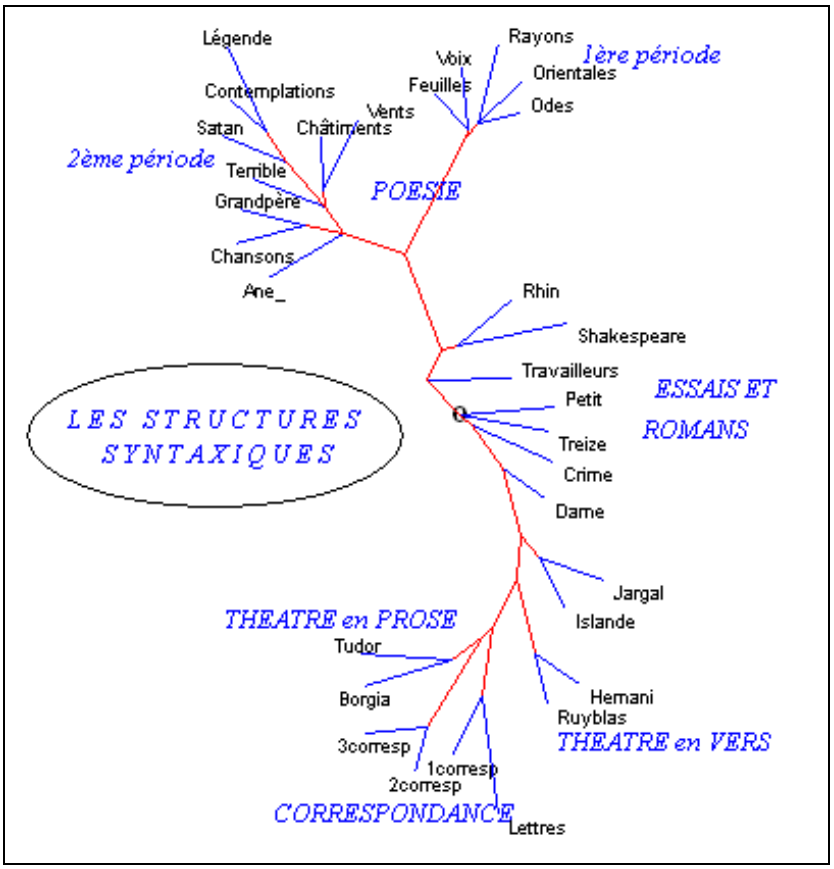
On peut aller plus loin dans l'exploration de la syntaxe hugolienne, en proposant au même calcul de distance, non plus seulement les codes isolés, mais leurs assemblages dans la chaîne syntaxique. Toutes les séquences des parties du discours, dans l'espace compris entre deux ponctuations, ont été répertoriées et indexées. Elles peuvent individuellement faire l'objet d'une recherche documentaire ou d'un histogramme, comme indiqué dans la figure 8. Leur ensemble constitue aussi un matériau propre à dessiner la carte des textes (voir graphique 14), si l'on emploie les mêmes méthodes que précédemment pour mesurer la distance.

Figure 13. La distance grammaticale. Analyse arborée



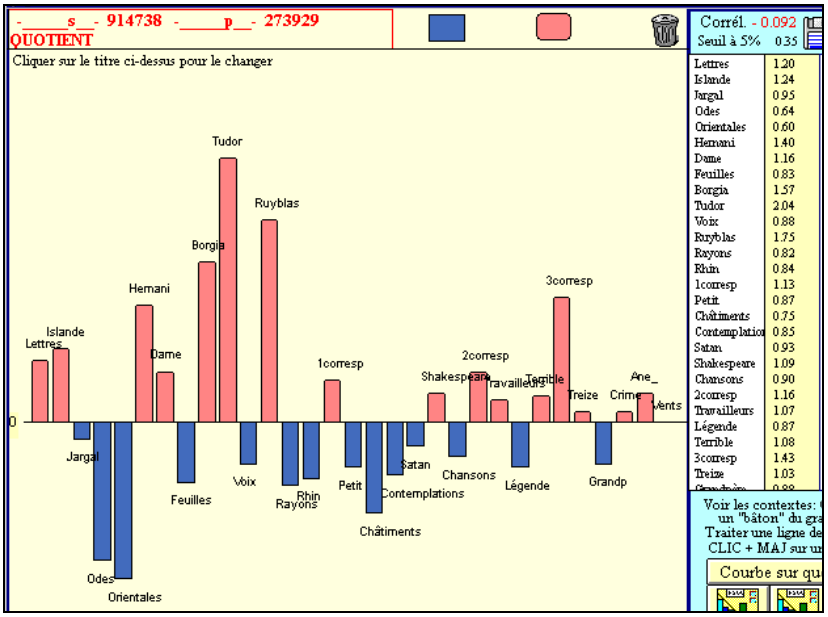
Beaucoup de ces structures syntaxiques sont longues et de ce fait peu souvent représentées dans leur intégralité. Mais les structures partielles à deux ou trois éléments ont aussi leur intérêt, par exemple les deux variantes du couple adjectif + substantif, c'est-à-dire l'antéposition ou la postposition de l'adjectif. Tableaux de contingence, histogrammes, analyses multidimensionnelles et calcul de spécificités s'appliquent aussi à de telles données. Le logiciel HYPERBASE fournit les fonctions et les outils nécessaires, qu'il serait trop long d'expliquer et d'illustrer ici.

Figure 14. La distance syntaxique. Analyse arborée.



La place nous manque aussi pour exposer les multiples approches que permet un corpus lemmatisé. Certaines sont relatives au rythme, comme celles qui apprécient la longueur des mots, ou des phrases. D'autres prennent pour objet le registre ou niveau de langue dont rendent compte les classes de fréquence. Beaucoup considèrent le verbe, où les critères de différenciation sont particulièrement discriminants: le temps, le mode, la personne. D'autres encore envisagent l'infra lexical, c'est-à-dire les composants du mot, et entre autres les préfixes et suffixes, les désinences, voire même les lettres. Toutes les distinctions, même les plus fines, et tous les regroupements, même les plus larges, sont à la disposition du chercheur. À titre d'exemple, le graphique ci-dessous (figure 15) rend compte du quotient singulier/pluriel dans les catégories nominales. Cette super catégorie du nombre relève aussi du genre: tous les recueils de poésie, et eux seuls, privilégient le pluriel (partie basse du graphique), comme si l'expression poétique était celle de la "multiple splendeur", selon l'expression de Verhaeren. Car cette particularité n'est pas propre à Hugo, non plus qu'à Verhaeren: elle est partagée par la plupart des poètes.

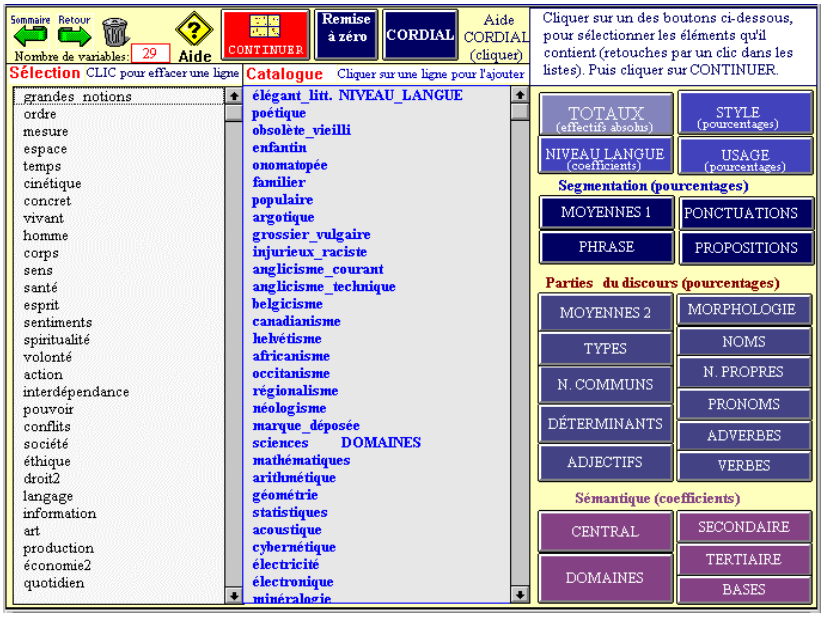
Figure 15. Le rapport singulier/pluriel dans les substantifs



Osera-t-on enfin aborder la thématique? Plusieurs méthodes s'y emploient, soit qu'on cherche à cerner un thème particulier en examinant les contextes où il se manifeste (une fonction, appelée précisément THEME, permet d'isoler les corrélats du thème choisi), soit qu'on aide le corpus, par un calcul de spécificité, à se détacher d'une référence externe (dans le cas présent Frantext représente cette toile de fond), soit qu'un thesaurus des objets conceptuels soit invoqué pour apprécier ce dont parle le texte (*Cordial* fournit cette référence extérieure), soit enfin qu'on examine en interne, page après page, les associations lexicales qui dessinent d'un pôle à l'autre les constellations zodiacales du texte. Nous nous limiterons à ces deux dernières approches, même si les premières sont aussi mises en oeuvre par le logiciel.

Commençons par l'approche de *Cordial* qui ressemble à un calcul de spécificité. *Cordial* fournit en effet des statistiques dans le domaine de la sémantique et de la thématique. Un immense tableau est ainsi constitué qui comporte autant de lignes que de variables ou concepts répertoriés (il y en a plus de 1000) et autant de colonnes que l'on compte de textes dans le corpus. Une page spéciale (reproduite ci-dessous) donne accès à ces indices et en assure la gestion, par séries homogènes, de type sémantique et, parfois, rythmique ou stylistique.

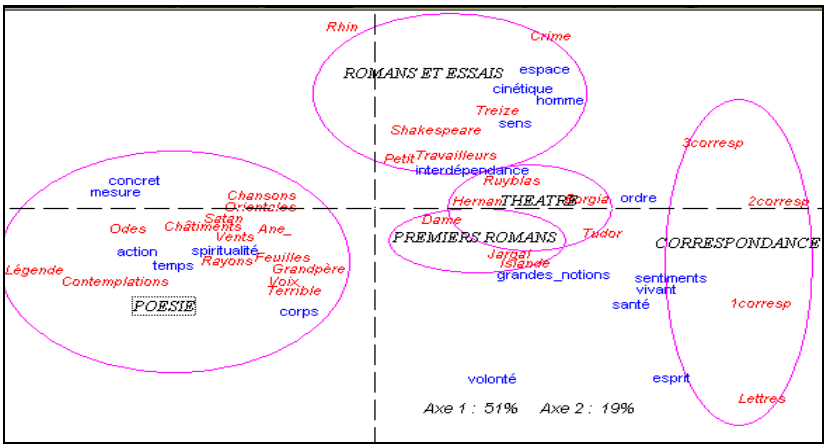
Figure 16. Les 1345 indices statistiques relevés par Cordial



Le résultat de l'analyse thématique confirme ce que nous avaient enseigné les études qui précèdent sur la prédominance du genre, même là où on l'attend moins naturellement. Même si les étiquettes sont floues et souvent peu aptes à représenter les sociétés du passé (on aurait préféré par exemple *mouvement* à *cinétique*), on voit à travers la figure 17 que la poésie hugolienne est liée au concret, au corps et au temps, tandis qu'espace et mouvement se déploient dans le roman, la correspondance étant le refuge des sentiments et de l'abstraction.

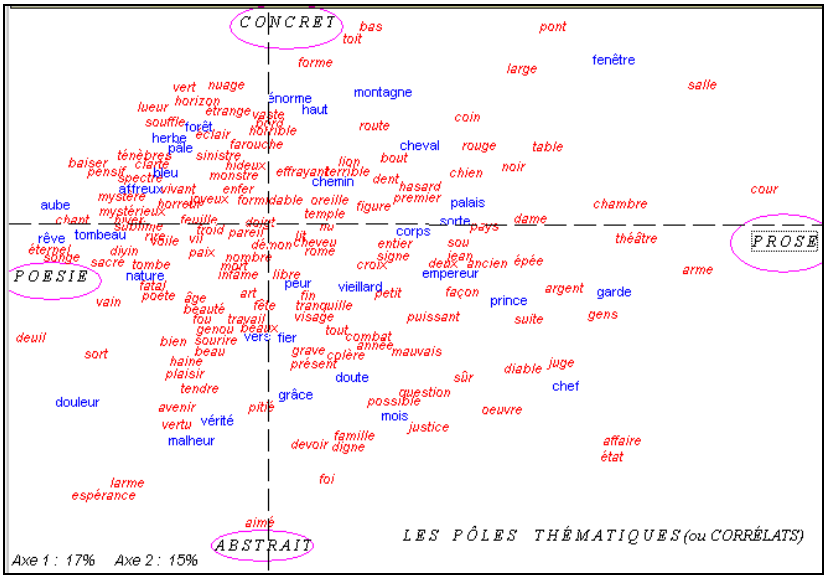
Indépendamment de *Cordial*, on a essayé une autre approche fondée sur les corrélats et apparentée à la démarche du logiciel *Alceste*. Au lieu de recevoir leur étiquette de l'extérieur, d'un thesaurus préalablement constitué, on voudrait que les mots reconnaissent d'eux-mêmes leur parenté, du seul fait de leur voisinage dans les mêmes contextes. Pour ce faire, le programme comptabilise tous les mots-pleins (au moins ceux qui sont des substantifs ou des adjectifs ni trop rares, ni trop fréquents) et enregistre toutes les rencontres, occasionnelles ou insistantes, avec les autres mots. Un lien est établi entre deux lemmes quand ils ont tendance à se rencontrer dans la même page. Le choix de la page permet d'échapper en partie aux contraintes syntaxiques qu'imposerait le choix d'une unité linguistique plus courte (syntagme, phrase ou paragraphe). L'élimination des mots fréquents et des mots-outils concourt aussi à privilégier les relations sémantiques ou thématiques plutôt que les rapports de dépendance syntaxique. On notera enfin que la division en textes est ignorée. La cohabitation à longue distance dans un même texte n'entre pas dans le calcul. Seule compte la proximité immédiate dans la même page, là où l'on a le plus de chances de relever les isotopies.

Figure 17. Analyse factorielle des thèmes relevés par Cordial



Le résultat est à interpréter dans l'analyse factorielle de la figure 18. Ne cherchons pas les textes ici, puisque le calcul les ignore (mais il ne serait pas impossible de les projeter dans l'analyse en éléments supplémentaires). Il n'y a là que des lemmes que des accointances ou répugnances particulières - et non le hasard - ont distribués aux quatre coins du graphique. En se laissant guider par le principe: qui s'assemble se ressemble, on devine que les éléments regroupés à l'extrême droite et à l'extrême gauche du graphe s'opposent comme la prose s'oppose à la poésie. D'un côté on relève des termes prosaïques liés au milieu, au cadre physique, aux occupations, aux intérêts (*état, affaire, argent, gens, épée, arme, garde, salle, chambre, cour, fenêtre, coin*, etc). De l'autre c'est la palette lexicale, habituelle aux poètes, dont certains éléments sont chers à Hugo (*pensif, affreux, mystère, mystérieux, sublime, divin, rêve, songe, poète, ténèbres, deuil, tombeau, tombe, spectre, voile, nature*, etc..).

Figure 18. La carte des associations lexicales

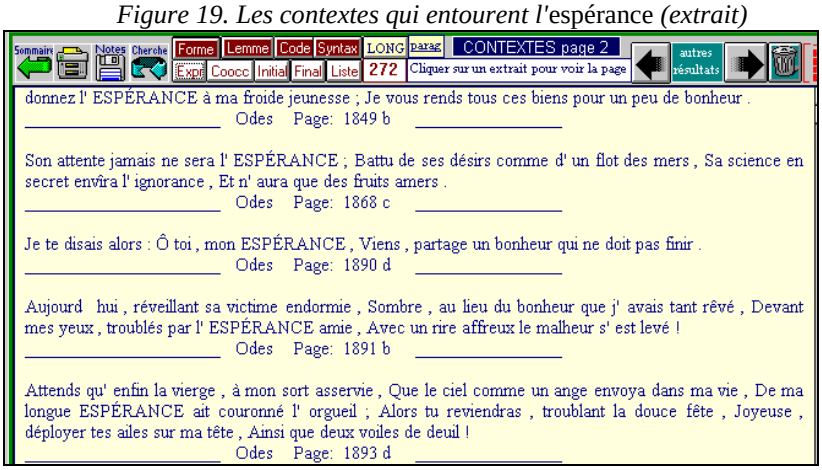


Ce premier facteur, assez attendu, est croisé avec un second, d'importance à peu près égale (15% contre 17% de la variance), qui oppose le haut et le bas du graphe et qui se rattache au couple concret/abstrait. La moitié inférieure fait évoluer des sentiments, des valeurs, des notions, des qualités immatérielles (*espérance, foi, vertu, vérité, pitié, grâce, devoir,*

digne, justice, doute, colère, haine, plaisir, avenir, malheur, etc.). En haut prennent place les substantifs qui désignent les réalités concrètes ou les adjectifs pour les décrire (*toit, pont, montagne, route, horizon, nuage, forêt, herbe, éclair, lueur, vert, bas, haut, énorme, vaste*, etc). On a souvent dit que les mots et les choses se présentaient de façon contrastée et antithétique dans l'esprit de Hugo. On voit ici l'effet des deux principales forces de polarisation. La troisième, qui n'est pas représentée dans la figure 18, est l'opposition manichéenne entre le bien et le mal, le beau et le laid, le positif et le négatif, la nature et l'homme. Parmi les réprouvés: *Rome* et l'*empereur*, les *princes*, les *puissants* et leurs *épées*, les *chiens* et leurs *dents*, les *spectres* et les *monstres*, les *ténèbres* et les *éclairs*, les *morts*, les *démons* et l'*enfer*, et tout ce qui est *vil, infâme, hideux, horrible, sinistre, fatal, pâle, froid, terrible, effrayant*. Parmi les élus: la *beauté*, la *nature*, la *forêt*, l'*herbe*, la *grâce*, le *plaisir*, le *sourire*, le *baiser*, le *chant*, l'*art*, le *travail*, les *vers*, le *poète* et tout ce qui est *vert, bleu, tendre, beau, tranquille, divin, digne et libre.*

La contre-épreuve est facile: isolons un mot, par exemple le lemme espérance, qui figure au bas de la figure ci-dessus, parmi les vertus ou notions abstraites, et recherchons toutes ses occurrences dans le corpus hugolien. On en trouve 272, dont certaines sont détaillées dans l'extrait de la figure 19.

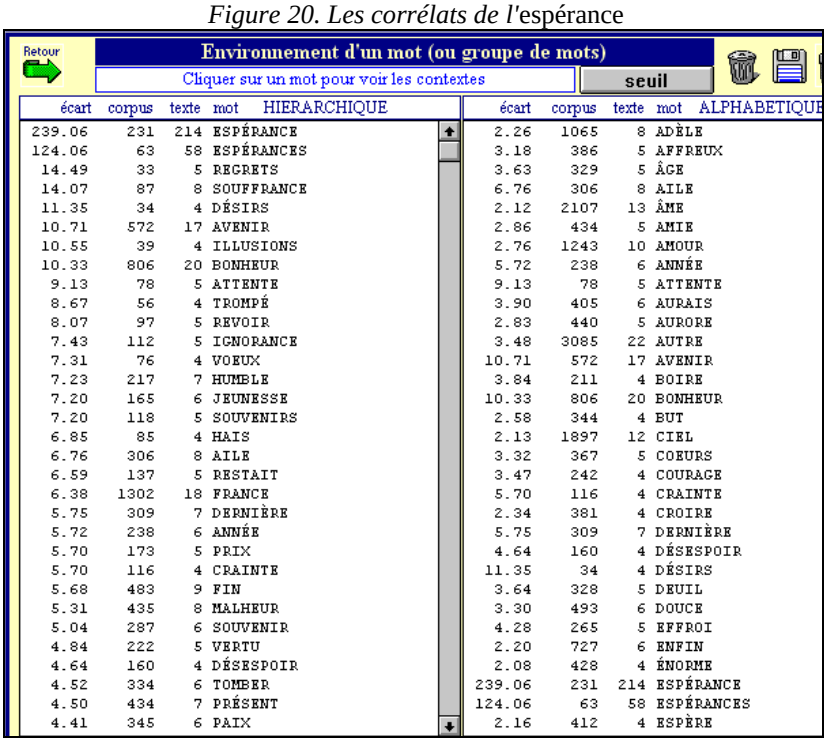
*Figure 19. Les contextes qui entourent l'*espérance *(extrait)*



Si l'on rassemble ces 272 contextes et qu'on sollicite la fonction *THEME*, les particularités de ces contextes seront mises en relief et en particulier les corrélats ou mots associés au pôle choisi. Les corrélats de l'*espérance* sont bien ceux que la figure 18 nous laissait présager: certes certains s'expliquent

par la versification: si la *France*, la *souffrance*, et l'*ignorance* se trouvent si souvent à proximité de l'*espérance*, c'est parfois pour les besoins de la rime. La phraséologie et la syntaxe peuvent aussi justifier certaines associations, notamment lorsqu'une expression figée est sous-jacente (*aile de l'espérance*, *tromper l'espérance)*. Mais la plupart des corrélats sont isotopiques. Ils partagent avec l'espérance des sèmes et des attributs communs. Beaucoup sont des vertus ou des sentiments proches, formant un camaïeu dégradé dont les nuances s'échelonnent entre les synonymes et les antonymes (*vertu, paix, bonheur, attente, voeux, désirs, avenir, souvenir, regrets, crainte, malheur, désespoir*).

*Figure 20. Les corrélats de l'*espérance



écart	corpus	texte	mot	HIERARCHIQUE
239.06	231	214	ESPÉRANCE	
124.06	63	58	ESPÉRANCES	
14.49	33	5	REGRETS	
14.07	87	8	SOUFFRANCE	
11.35	34	4	DÉSIRS	
10.71	572	17	AVENIR	
10.55	39	4	ILLUSIONS	
10.33	806	20	BONHEUR	
9.13	78	5	ATTENTE	
8.67	56	4	TROMPÉ	
8.07	97	5	REVOIR	
7.43	112	5	IGNORANCE	
7.31	76	4	VOEUX	
7.23	217	7	HUMBLE	
7.20	165	6	JEUNESSE	
7.20	118	5	SOUVENIRS	
6.85	85	4	HAIS	
6.76	306	8	AILE	
6.59	137	5	RESTAIT	
6.38	1302	18	FRANCE	
5.75	309	7	DERNIÈRE	
5.72	238	6	ANNÉE	
5.70	173	5	PRIX	
5.70	116	4	CRAINTE	
5.68	483	9	FIN	
5.31	435	8	MALHEUR	
5.04	287	6	SOUVENIR	
4.84	222	5	VERTU	
4.64	160	4	DÉSESPOIR	
4.52	334	6	TOMBER	
4.50	434	7	PRÉSENT	
4.41	345	6	PAIX	

écart	corpus	texte	mot	ALPHABETIQUE
2.26	1065	8	ADÈLE	
3.18	386	5	AFFREUX	
3.63	329	5	ÂGE	
6.76	306	8	AILE	
2.12	2107	13	ÂME	
2.86	434	5	AMIE	
2.76	1243	10	AMOUR	
5.72	238	6	ANNÉE	
9.13	78	5	ATTENTE	
3.90	405	6	AURAIS	
2.83	440	5	AURORE	
3.48	3085	22	AUTRE	
10.71	572	17	AVENIR	
3.84	211	4	BOIRE	
10.33	806	20	BONHEUR	
2.58	344	4	BUT	
2.13	1897	12	CIEL	
3.32	367	5	COEURS	
3.47	242	4	COURAGE	
5.70	116	4	CRAINTE	
2.34	381	4	CROIRE	
5.75	309	7	DERNIÈRE	
4.64	160	4	DÉSESPOIR	
11.35	34	4	DÉSIRS	
3.64	328	5	DEUIL	
3.30	493	6	DOUCE	
4.28	265	5	EFFROI	
2.20	727	6	ENFIN	
2.08	428	4	ÉNORME	
239.06	231	214	ESPÉRANCE	
124.06	63	58	ESPÉRANCES	
2.16	412	4	ESPÈRE	

Ainsi, la recherche hypertextuelle et statistique est un va-et-vient entre deux démarches qui se confortent l'une l'autre. L'une est probatoire; elle part du particulier, d'un mot ou d'un thème à rechercher (comme le mot *espérance*), ou d'une hypothèse à vérifier. L'autre est exploratoire et accueillante à la découverte, voire à la surprise; elle interroge le corpus dans son ensemble, en lui appliquant quelque filtre ou quelque traitement

prédéfini, dont le résultat, parfois imprévu, sollicite l'attention et génère à l'infini hypothèses et contrôles. Rien de bien nouveau dans ce balancement cyclique de l'esprit humain, si ce n'est que l'ordinateur accélère le mouvement et qu'il avertit le chercheur quand on tourne en rond.