Une mesure de la distance intertextuelle : la connexion lexicale

Étienne Brunet

Une mesure de la distance intertextuelle : la connexion lexicale

Étienne BRUNET

Les travaux qu'on range sous la rubrique *linguistique quantitative* ou *lexicométrie* ou *lexicologie statistique* associent toujours le nombre au discours, une méthode à un objet. Toutes les dénominations qu'on a proposées font référence au lexique, parfois plus largement à la linguistique, jamais à la littérature ni à la critique littéraire. Même quand l'objet d'étude est l'oeuvre d'un écrivain, il arrive que la spécificité de l'écriture littéraire ne soit guère prise en compte. C'est que les recherches de ce genre portent sur la structure du discours plus souvent que sur le contenu. Les mots de l'écrivain constituent bien la matière première de ces travaux, mais, immédiatement soumis à des transformations et à des reclassements, ils perdent dès l'entrée leur individualité, leur sens et leur saveur. Car la structure lexicale met en jeu des classes, des catégories, des effectifs, c'est-à-dire des nombres.

Quant au contenu lexical, on s'est contenté jusqu'ici de relever les spécificités, c'est-à-dire de souligner les unités linguistiques dont la fréquence manifestait un excédent ou un déficit significatif, compte tenu d'une norme préalablement établie (très souvent c'est l'ensemble qui constitue la norme pour les parties.) Le filtrage du vocabulaire spécifique a fait appel à des méthodes plus ou moins sophistiquées et plus ou moins exactes et il a pu prendre pour objet des formes, des vocables, des syntagmes, mais toujours il aboutissait à des listes interminables que l'ordinateur cédait aimablement au chercheur, après un dernier tri hiérarchique ou alphabétique. Dans certains cas, les écarts relevés étaient repris en entrée par des programmes de corrélation, de dispersion ou d'analyse factorielle et des perspectives plus synthétiques étaient ainsi ouvertes, qui restaient cependant partielles. Car ces méthodes ne peuvent être appliquées que là où les effectifs ont une ampleur suffisante, ce qui a pour effet d'exclure tous les mots de fréquence rare ou même moyenne.

Or c'est une prise en compte globale du contenu lexical que, dès 1959, Pierre Guiraud proposait dans *Problèmes et méthodes de la statistique linguistique* (p. 129) : "On pourrait établir un tableau de corrélations lexicales entre les différentes oeuvres en les prenant deux à deux pour voir les mots qu'elles ont en commun et ceux qu'elles ont en propre ; mais c'est un travail énorme." Et Charles Muller, en 1967, dans son *Etude de statistique lexicale* (pp. 169-173) se livrait précisément à ce type de calcul, qu'il désigne sous le nom de *connexion* lexicale et dont il donne la théorie détaillée dans son *Initiation à la statistique lexicale* (pp. 210-215)[1]. On conçoit que Guiraud, démuni, à l'époque, de moyens de calcul, ait renoncé à entreprendre un "travail énorme", mais on peut s'étonner que la tentative de Muller ait été si rarement reprise, alors que la méthode était clairement définie et les outils de traitement disponibles. Il faut certes mettre en oeuvre des programmes spécifiques qui sont assez complexes et qui requièrent une puissance suffisante. Mais la programmation est devenue un exercice banal et les machines ont tellement progressé qu'on ne devrait plus reculer devant des recherches de ce genre. Et c'est ce que nous avons tenté il y a dix ans sur le corpus de Giraudoux et tout récemment sur celui de Hugo — dont nous nous proposons de rendre compte.

Ce vaste dépouillement vient, comme tous ceux que nous avons traités, des dépouilles du *Trésor de la langue française*. Il contient plus de deux millions d'occurrences et vingt textes complets, dont trois romans (y compris l'intégralité des *Misérables),* quatre pièces de théâtre, huit textes poétiques[2], et quatre recueils de correspondance. S'y ajoute un récit de voyage, le *Rhin,* dont le genre est hybride. Comme le poids écrasant des *Misérables* risquait de détruire l'équilibre, on a réparti en trois sous-ensembles leur masse imposante. C'est donc 22 textes (ou sous-textes) qui vont être comparés deux à deux, ce qui porte le nombre de combinaisons à : (22 x 21)/2 = 231. Pour chacune de ces 231 confrontations, on a calculé :

— l'étendue du vocabulaire du texte a et celle du texte *b* (et aussi l'étendue de l'un et de l'autre en nombre d'occurrences)
— l'étendue du vocabulaire des deux textes réunis dans le même ensemble (en vocables et en occurrences)
— la part du vocabulaire commune aux deux textes et la part privative de chacun (là aussi en considérant *N* et V).

[1] Ce chapitre est repris — avec des développements moindres — dans *Principes et méthodes de statistique lexicale,* Hachette, 1977, pp. 145-154.

[2] A vrai dire, la *Légende des siècles* compte à elle seule pour trois textes, eu égard à l'échelonnement des dates de publication.

Tableau 1

Connexion lexicale. Données brutes

Travailleurs de la mer									
	ab	*a + b*	connex	*a*	*b*	*a − b*	*b − a*	indep *a*	indep *b*
CO3	3 732	10 575	0,352	8 824	5 483	5 092	1 751	0,577	0,319
	223 473	19 576	11,415	134 135	108 914	14 796	4 780	0,110	0,043
LS2	4 234	10 648	0,397	8 824	6 058	4 590	1 824	0,520	0,301
	211 655	15 243	13,885	134 135	92 763	11 622	3 621	0,086	0,039
LS3	3 125	9 888	0,316	8 824	4 189	5 699	1 064	0,645	0,253
	152 252	18 170	8,379	134 135	36 287	16 604	1 566	0,123	0,043
Correspondances 3									
	ab	*a + b*	connex	*a*	*b*	*a − b*	*b − a*	indep *a*	indep *b*
LS2	2 939	8 602	0,341	5 483	6 058	2 544	3 119	0,463	0,514
	183 841	17 836	10,307	108 914	92 763	9 032	8 804	0,082	0,094
LS3	2 326	7 346	0,316	5 483	4 189	3 157	1 863	0,575	0,444
	129 689	15 512	8,360	108 914	36 287	12 200	3 312	0,112	0,091
Légendes des siècles 2									
	ab	*a + b*	connex	*a*	*b*	*a − b*	*b − a*	indep *a*	indep *b*
LS3	3 236	7 011	0,461	6 058	4 189	2 822	953	0,465	0,227
	122 812	6 238	19,687	92 763	36 287	5 115	1 123	0,055	0,030

A titre d'exemple l'extrait du tableau 1 fournit ces données pour les derniers textes de la série : les *Travailleurs de la mer*, le troisième recueil de *Correspondance* et les dernières livraisons de la *Légende des siècles.* Ainsi si l'on rapproche les *Travailleurs de la mer* (texte a) de la *Correspondance (1867-1873)* (texte *b),* on obtient le tableau 2.

Tableau 2

	V	*N*	Notation
texte a	8 824	134 135	(a)
texte *b*	5 483	108 914	*(b)*
part commune	3 732	223 473	*(ab)*
ensemble	10 575	243 049	(a + *b)*
part exclusive a	5 092	14 796	(a - *b)*
part exclusive *b*	1 751	4 780	*(b - a)*

Un simple rapport permet de calculer la connexion lexicale de ces deux textes :

$$\frac{\text{vocabulaire commun}}{\text{vocabulaire de l'ensemble}} \quad \frac{3\,732}{10\,575} = 0,352.$$

Un autre rapport rend compte de l'indépendance du texte a :

$$\frac{\text{part exclusive de a}}{\text{vocabulaire a}} = \frac{5\,092}{8\,824} = 0,577$$

et un calcul semblable mesure l'indépendance du texte *b :*

$$\frac{\text{part exclusive de } b}{\text{vocabulaire } b} = \frac{1\,751}{5\,483} = 0,319$$

Tous ces rapports sont appliqués de la même façon aux occurrences, quoique le calcul y soit sans doute un peu moins légitime. Car une occurrence saurait difficilement être qualifiée de "commune", puisqu'elle doit être nécessairement comptabilisée dans un camp ou dans l'autre. Au reste le calcul de la connexion a été légèrement modifié pour les occurrences : il désigne cette fois le rapport entre les mots communs et les mots exclusifs, soit pour le même exemple que précédemment :

$$\frac{\text{part commune}}{\text{part exclusive}} = \frac{223\,473}{19\,576} = 11,415$$

On se trouve donc devant six séries de 231 coefficients, devant six tableaux triangulaires où chacun des 22 textes est rapproché successivement des 21 autres. Le premier (tableau 3), qui représente parallèlement les vocables (ligne 1) et les occurrences (ligne 2), concerne la connexion lexicale proprement dite. On s'intéressera à trois lignes de ce tableau, qui reproduisent le profil d'un texte poétique *(Les Feuilles d'automne),* d'une pièce de théâtre *(Hernani),* et d'un roman (première partie des *Misérables).* La représentation graphique de la figure 1 permet de constater que les *Feuilles d'automne* partagent le même vocabulaire que les *Rayons et les ombres,* les *Contemplations,* la *Légende des siècles,* la *Fin de Satan* et les *Chansons des rues et des bois.* Et cette loi du genre, qui est souveraine en poésie, règne aussi au théâtre : c'est Ruy *Blas,* autre drame en vers, qui se rapproche le plus *d'Hernani,* suivi des deux pièces en prose : *Lucrèce Borgia* et *Marie Tudor.* Quant à l'oeuvre romanesque, son unité lexicale se manifeste dans le cas des *Misérables,* dont les trois parties, fortement soudées entre elles, sont aussi liées à *Notre-Dame* et aux *Travailleurs de la mer.* Ces clivages reconnus parmi les vocables (histogrammes en noir sur le graphique) se retrouvent pareillement lorsqu'on considère les occurrences (histogrammes en gris).

L'**indépendance** lexicale n'est pas purement l'envers de la connexion. Car il y a lieu de dédoubler l'indice, selon qu'on considère la part exclusive du texte a ou celle du texte *b.* Si le vocabulaire de l'un est tout entier contenu dans le vocabulaire de l'autre, l'indice sera nul (au moins pour le plus petit). Il sera égal à 1 dans le cas, plus improbable encore, où les deux textes ne partagent aucun mot du lexique. La réalité fournit naturellement des valeurs intermédiaires.

Tableau 3

Connexion lexicale (indice brut)

Ligne 1 : vocables Ligne 2 : occurrences

HER	*AUT*	*DAM*	*BOR*	*TUD*	*RUY*	*RAY*	*RHI*	*CO1*	*CON*	*LS1*	*MI1*	*MI2*	*MI3*	*FIN*	*RUE*	*CO2*	*MER*	*CO3*	*LS2*	*LS3*	
0,31	0,31	0,29	0,30	0,32	0,34	0,31	0,28	0,43	0,31	0,29	0,31	0,29	0,30	0,30	0,27	0,37	0,29	0,38	0,30	0,29	*LET*
9,63	8,67	8,64	9,51	10,24	10,83	8,76	7,25	24,04	9,57	7,99	10,42	10,05	10,00	8,25	8,42	16,28	8,11	15,67	8,59	9,10	
	0,37	0,23	0,36	0,38	0,38	0,35	0,22	0,29	0,30	0,30	0,22	0,21	0,22	0,32	0,29	0,25	0,21	0,28	0,29	0,32	*HER*
	8,36	6,47	10,52	10,93	11,21	8,01	5,60	8,57	7,57	7,16	6,46	6,13	6,23	7,31	6,70	7,62	5,05	7,49	7,33	7,66	
		0,26	0,30	0,29	0,34	0,45	0,25	0,30	0,39	0,36	0,26	0,24	0,25	0,40	0,35	0,28	0,25	0,28	0,36	0,37	*AUT*
		6,63	6,24	6,59	7,66	14,05	6,40	7,83	12,71	9,52	6,78	6,45	6,58	11,02	8,95	7,63	5,88	7,24	10,46	9,97	
			0,21	0,21	0,29	0,28	0,49	0,37	0,38	0,39	0,48	0,47	0,47	0,32	0,30	0,39	0,44	0,35	0,40	0,32	*DAM*
			5,92	6,09	8,19	7,44	25,25	13,21	13,35	13,72	23,25	23,50	21,25	8,95	8,17	15,93	16,27	11,33	14,59	8,47	
				0,38	0,35	0,29	0,19	0,27	0,25	0,26	0,21	0,19	0,21	0,27	0,26	0,24	0,20	0,26	0,25	0,29	*BOR*
				11,94	9,90	5,98	5,11	8,42	5,57	5,60	6,30	5,89	5,92	5,52	5,60	7,30	4,71	7,21	5,46	6,23	
					0,36	0,29	0,19	0,28	0,26	0,26	0,21	0,19	0,20	0,28	0,26	0,24	0,20	0,27	0,26	0,28	*TUD*
					10,91	6,50	5,05	8,32	5,81	5,67	6,55	6,04	6,16	5,97	5,78	7,49	4,95	7,31	5,68	6,71	
						0,36	0,27	0,33	0,35	0,34	0,29	0,27	0,29	0,35	0,32	0,31	0,27	0,32	0,34	0,35	*RUY*
						8,50	7,17	10,06	9,35	8,89	9,02	8,47	8,46	8,29	7,71	9,47	6,60	8,93	8,79	8,48	
							0,28	0,31	0,43	0,38	0,28	0,27	0,28	0,41	0,39	0,29	0,27	0,30	0,39	0,40	*RAY*
							7,26	8,27	16,29	11,07	7,65	7,31	7,31	12,37	10,75	8,17	6,49	7,59	12,30	11,50	
								0,37	0,38	0,40	0,49	0,49	0,48	0,31	0,30	0,42	0,45	0,35	0,40	0,31	*RHI*
								13,40	12,79	13,98	27,41	26,07	23,49	8,26	7,54	18,51	18,81	11,86	14,84	7,98	
									0,34	0,33	0,40	0,38	0,39	0,31	0,30	0,47	0,35	0,45	0,34	0,31	*CO1*
									11,25	9,53	16,95	16,12	15,47	7,83	8,28	34,02	11,71	27,79	10,09	8,42	
										0,50	0,39	0,37	0,39	0,49	0,44	0,37	0,37	0,34	0,51	0,45	*CON*
										24,46	14,84	14,23	14,73	23,59	16,93	13,18	12,22	11,20	29,55	19,36	
											0,40	0,38	0,39	0,47	0,40	0,36	0,39	0,33	0,53	0,45	*LS1*
											14,96	13,48	14,36	19,13	12,27	11,38	12,70	9,46	30,75	17,14	
												0,53	0,52	0,33	0,32	0,44	0,48	0,39	0,41	0,32	*MI1*
												36,51	34,44	9,70	9,12	23,22	22,84	15,82	16,16	9,66	
													0,53	0,31	0,30	0,44	0,46	0,38	0,39	0,31	*MI2*
													34,30	9,02	8,46	23,06	20,03	15,08	14,67	8,98	
														0,33	0,32	0,44	0,48	0,38	0,41	0,32	*MI3*
														9,66	8,78	21,71	21,82	14,58	15,94	9,33	
															0,41	0,32	0,33	0,31	0,48	0,47	*FIN*
															12,45	8,56	8,68	7,61	22,34	17,96	
																0,30	0,30	0,30	0,42	0,41	*RUE*
																8,83	7,47	8,29	14,40	11,83	
																	0,40	0,50	0,37	0,32	*CO2*
																	16,45	42,05	12,32	9,00	
																		0,35	0,40	0,32	*MER*
																		11,42	13,89	8,38	
																			0,34	0,32	*CO3*
																			10,31	8,36	
																				0,46	*LS2*
																				19,69	

Tableau 4

L'indépendance lexicale. Indices bruts

Indépendance lexicale de $a \longrightarrow b$

HER	*AUT*	*DAM*	*BOR*	*TUD*	*RUY*	*RAY*	*RHI*	*CO1*	*CON*	*LS1*	*MI1*	*MI2*	*MI3*	*FIN*	*RUE*	*CO2*	*MER*	*CO3*	*LS2*	*LS3*	
0,60	0,57	0,21	0,63	0,62	0,52	0,55	0,21	0,25	0,42	0,43	0,19	0,17	0,20	0,51	0,54	0,23	0,25	0,32	0,39	0,52	*LET*
0,09	0,10	0,02	0,10	0,09	0,08	0,09	0,02	0,02	0,05	0,07	0,02	0,02	0,02	0,09	0,08	0,02	0,03	0,03	0,06	0,08	
	0,41	0,12	0,49	0,48	0,36	0,39	0,11	0,28	0,26	0,22	0,15	0,13	0,14	0,34	0,39	0,23	0,20	0,31	0,22	0,35	*HER*
	0,09	0,03	0,10	0,10	0,06	0,09	0,02	0,06	0,05	0,04	0,03	0,02	0,03	0,08	0,08	0,05	0,04	0,06	0,04	0,08	
		0,16	0,60	0,61	0,47	0,34	0,13	0,35	0,21	0,21	0,17	0,16	0,18	0,31	0,37	0,28	0,22	0,38	0,20	0,35	*AUT*
		0,03	0,16	0,17	0,11	0,06	0,02	0,08	0,03	0,04	0,03	0,03	0,03	0,06	0,07	0,06	0,04	0,08	0,03	0,06	
			0,78	0,79	0,70	0,70	0,31	0,57	0,56	0,54	0,36	0,33	0,36	0,65	0,66	0,50	0,40	0,59	0,53	0,65	*DAM*
			0,16	0,16	0,12	0,13	0,03	0,10	0,09	0,08	0,04	0,04	0,05	0,12	0,12	0,08	0,06	0,10	0,08	0,12	
				0,46	0,35	0,44	0,11	0,27	0,30	0,26	0,13	0,10	0,13	0,38	0,40	0,21	0,19	0,29	0,26	0,36	*BOR*
				0,08	0,06	0,09	0,02	0,05	0,06	0,05	0,02	0,02	0,02	0,09	0,08	0,04	0,04	0,05	0,05	0,08	
					0,33	0,43	0,10	0,23	0,27	0,24	0,12	0,10	0,12	0,36	0,39	0,18	0,16	0,26	0,23	0,35	*TUD*
					0,05	0,08	0,01	0,05	0,05	0,04	0,02	0,01	0,02	0,08	0,07	0,03	0,02	0,04	0,03	0,07	
						0,47	0,15	0,34	0,31	0,29	0,16	0,15	0,16	0,42	0,44	0,27	0,21	0,36	0,29	0,41	*RUY*
						0,10	0,02	0,07	0,06	0,05	0,03	0,02	0,03	0,10	0,09	0,05	0,04	0,07	0,05	0,09	
							0,15	0,37	0,22	0,24	0,18	0,17	0,18	0,35	0,37	0,30	0,23	0,39	0,22	0,36	*RAY*
							0,02	0,08	0,03	0,04	0,03	0,03	0,03	0,06	0,07	0,06	0,04	0,09	0,03	0,06	
								0,58	0,58	0,55	0,37	0,34	0,38	0,66	0,67	0,50	0,42	0,59	0,54	0,67	*RHI*
								0,09	0,09	0,08	0,04	0,04	0,05	0,13	0,13	0,07	0,05	0,10	0,08	0,13	
									0,50	0,49	0,25	0,22	0,26	0,59	0,60	0,28	0,33	0,38	0,47	0,58	*CO1*
									0,07	0,08	0,03	0,02	0,03	0,12	0,10	0,02	0,04	0,03	0,08	0,11	
										0,31	0,26	0,23	0,25	0,42	0,46	0,39	0,30	0,49	0,29	0,45	*CON*
										0,03	0,03	0,03	0,03	0,05	0,06	0,06	0,04	0,08	0,03	0,05	
											0,27	0,26	0,28	0,45	0,51	0,43	0,30	0,52	0,30	0,47	*LS1*
											0,04	0,04	0,04	0,06	0,08	0,08	0,05	0,11	0,03	0,07	
												0,27	0,31	0,64	0,65	0,46	0,36	0,55	0,51	0,64	*MI1*
												0,02	0,03	0,11	0,11	0,06	0,04	0,08	0,07	0,11	
													0,34	0,67	0,67	0,48	0,40	0,58	0,55	0,67	*MI2*
													0,03	0,12	0,12	0,06	0,05	0,08	0,08	0,11	
														0,64	0,65	0,47	0,36	0,56	0,52	0,65	*MI3*
														0,11	0,11	0,06	0,04	0,09	0,07	0,11	
															0,42	0,36	0,24	0,45	0,20	0,36	*FIN*
															0,07	0,07	0,04	0,10	0,02	0,05	
																0,38	0,26	0,47	0,28	0,42	*RUE*
																0,08	0,06	0,11	0,05	0,08	
																	0,35	0,40	0,50	0,61	*CO2*
																	0,04	0,03	0,08	0,11	
																		0,58	0,52	0,65	*MER*
																		0,11	0,09	0,12	
																			0,46	0,58	*CO3*
																			0,08	0,11	
																				0,47	*LS2*
																				0,06	

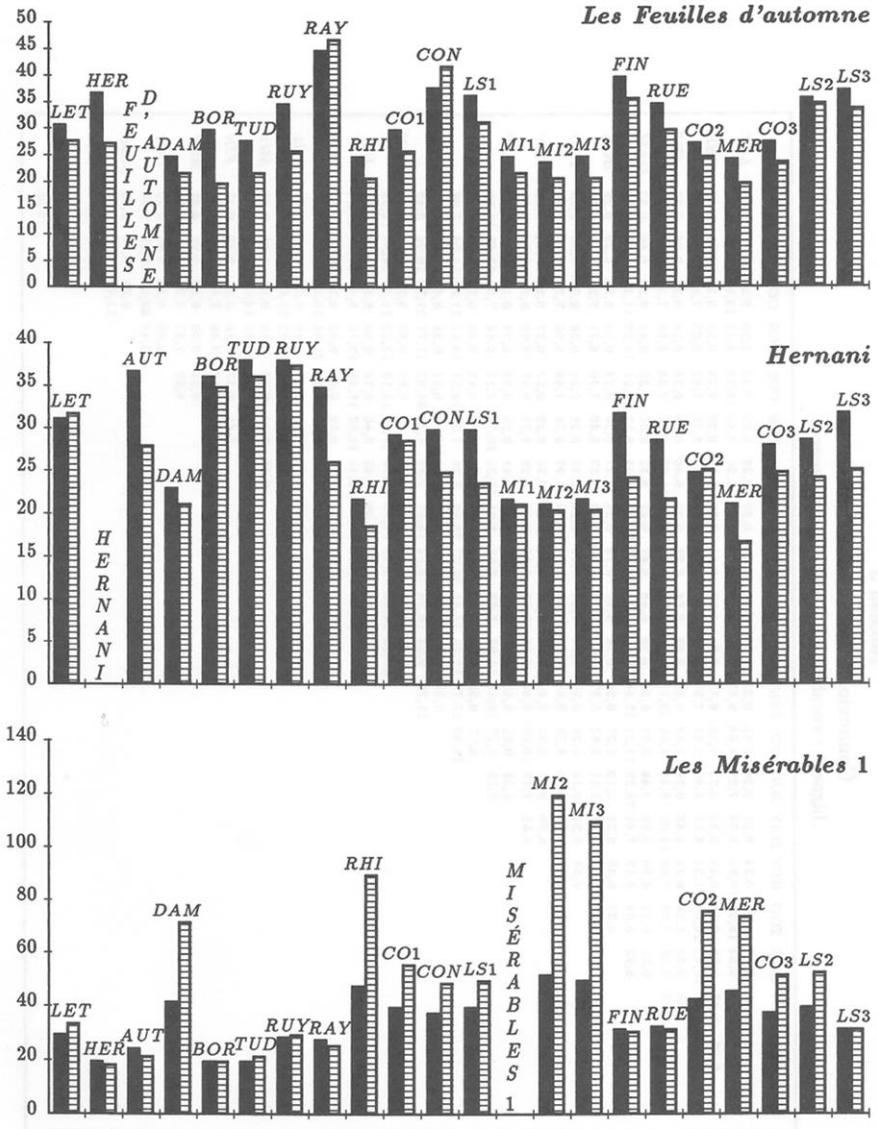

Figure 1

Courbes établies sur l'indice brut

En noir, les vocables; en hachures, les occurrences.

Mais ces valeurs pour un même couple divergent fortement, comme si la relation n'était pas symétrique et que les sentiments n'étaient pas payés de retour. Ainsi l'indépendance des *Travailleurs de la mer* par rapport à la *Légende des siècles* de 1877 est de 0,52, alors que celle de la *Légende* atteint seulement 0,301. Il ne faut pas conclure de là, trop vite, que la richesse lexicale est en cause. Car le rapprochement des *Travailleurs* avec la *Légende des siècles* de 1883 fait apparaître une dissymétrie plus grande encore : 0,645 et 0,253. L'explication est à chercher dans l'influence de l'étendue. Un texte très court est nécessairement inclus en grande partie dans un texte très long et le coefficient aura dans ce cas une valeur proche de 0, le texte long ayant au contraire un indice proche de 1. Il suffit de comparer les deux séries des tableaux 4 et 5. Pour une même ligne, par exemple celle qui concerne *Hernani* (ligne 2), tous les coefficients sont inférieurs à 0,50 lorsque l'indépendance de *Hernani* est envisagée (Tableau 4), alors qu'ils sont supérieurs à cette valeur lorsque l'indépendance des autres textes est envisagée vis-à-vis de *Hernani* (Tableau 5).

Une telle distorsion devrait conduire à l'abandon de cet indice d'indépendance. Mais un moyen très simple s'offre de la corriger : observons en effet que pour un couple désaccordé les deux indices s'écartent de la valeur centrale en sens inverse, mais d'un même pas. Quand l'un tend vers 0, l'autre tend vers 1. Il suffit donc pour une même paire d'ajouter l'un à l'autre les deux indices pour obtenir un compromis qui mesure l'indépendance réciproque des deux textes et qui a une signification opposée à celle de la connexion lexicale. Le résultat se lit dans le tableau 6. Si l'on s'intéresse par exemple à la dernière ligne, qui concerne la *Légende 3,* l'indépendance lexicale est la plus faible (et la connexion la plus forte) lorsque le texte mis en regard appartient à la même veine poétique : *Légende 1* (0,72) et *Légende 2* (0,70). L'indépendance (c'est-à-dire la distance) est la plus grande quand le même texte est confronté à la correspondance (respectivement 1,09, 1,03, 0,97 et 1,02) ou au théâtre en prose *(Lucrèce Borgia* 1,02 et *Marie Tudor* 1,01). Le tableau 6 propose une illustration graphique de l'exemple du *Rhin,* dont le genre est a *priori* mal défini, puisqu'il s'agit d'un récit de voyage, sous forme de lettres, et qui, dans les faits, se range délibérément du côté des romans, aussi bien dans la courbe de l'indépendance (partie médiane du tableau) que dans celle de la connexion (partie inférieure). Les deux graphiques (Figure 2) s'emboîtant parfaitement, la symétrie des notions d'indépendance et de connexion se trouve vérifiée.

Le tableau 6, qui est un tableau de distances, se prête naturellement à l'analyse factorielle. Le fait qu'on ait multiplié par 100 toutes les valeurs, pour supprimer les décimales et gagner de la place, ne modifie en rien les résultats. Mais les valeurs nulles qu'on lit sur la diagonale descendante ne pouvaient être maintenues sans nuire au traitement. Au lieu d'estimer que l'indépendance

Tableau 5

L'indépendance lexicale. Indices bruts

Indépendance lexicale de $b \longrightarrow a$

HER	*AUT*	*DAM*	*BOR*	*TUD*	*RUY*	*RAY*	*RHI*	*CO1*	*CON*	*LS1*	*MI1*	*MI2*	*MI3*	*FIN*	*RUE*	*CO2*	*MER*	*CO3*	*LS2*	*LS3*	
0,40	0,47	0,68	0,38	0,34	0,46	0,51	0,70	0,50	0,60	0,64	0,67	0,69	0,68	0,56	0,60	0,59	0,68	0,54	0,63	0,57	*LET*
0,10	0,13	0,15	0,09	0,08	0,11	0,14	0,17	0,06	0,14	0,17	0,12	0,13	0,13	0,15	0,17	0,08	0,17	0,08	0,15	0,15	
	0,51	0,76	0,44	0,41	0,52	0,55	0,78	0,68	0,66	0,67	0,77	0,79	0,77	0,61	0,64	0,73	0,77	0,69	0,68	0,61	*HER*
	0,12	0,15	0,08	0,07	0,10	0,13	0,17	0,11	0,13	0,15	0,15	0,15	0,15	0,14	0,17	0,12	0,19	0,13	0,14	0,14	
		0,72	0,46	0,46	0,52	0,42	0,74	0,65	0,57	0,60	0,73	0,75	0,73	0,51	0,55	0,69	0,73	0,66	0,60	0,53	*AUT*
		0,15	0,11	0,09	0,12	0,07	0,15	0,12	0,08	0,11	0,14	0,15	0,15	0,09	0,12	0,12	0,16	0,13	0,10	0,11	
			0,11	0,11	0,16	0,18	0,37	0,29	0,27	0,28	0,35	0,39	0,36	0,22	0,27	0,35	0,38	0,31	0,28	0,23	*DAM*
			0,02	0,02	0,03	0,03	0,04	0,03	0,03	0,04	0,04	0,04	0,04	0,03	0,06	0,04	0,06	0,04	0,04	0,04	
				0,44	0,56	0,63	0,80	0,70	0,72	0,72	0,79	0,80	0,79	0,67	0,68	0,75	0,79	0,71	0,73	0,66	*BOR*
				0,07	0,11	0,18	0,18	0,12	0,17	0,18	0,15	0,16	0,16	0,18	0,20	0,13	0,20	0,13	0,18	0,17	
					0,56	0,63	0,80	0,70	0,71	0,72	0,79	0,80	0,79	0,67	0,68	0,74	0,79	0,70	0,72	0,66	*TUD*
					0,11	0,18	0,18	0,12	0,18	0,19	0,15	0,16	0,16	0,18	0,21	0,13	0,19	0,14	0,18	0,17	
						0,48	0,72	0,61	0,59	0,60	0,69	0,72	0,70	0,54	0,56	0,65	0,70	0,61	0,61	0,53	*RUY*
						0,11	0,14	0,10	0,11	0,12	0,11	0,12	0,12	0,11	0,14	0,10	0,15	0,11	0,12	0,11	
							0,71	0,61	0,52	0,56	0,70	0,72	0,70	0,48	0,50	0,66	0,71	0,63	0,56	0,49	*RAY*
							0,13	0,11	0,07	0,10	0,13	0,13	0,14	0,08	0,10	0,12	0,15	0,12	0,09	0,09	
								0,24	0,23	0,24	0,31	0,35	0,32	0,19	0,22	0,29	0,34	0,26	0,24	0,21	*RHI*
								0,03	0,03	0,03	0,03	0,03	0,04	0,03	0,04	0,03	0,05	0,03	0,03	0,03	
									0,49	0,52	0,54	0,57	0,55	0,45	0,47	0,43	0,58	0,37	0,51	0,45	*CO1*
									0,09	0,11	0,08	0,08	0,08	0,10	0,12	0,03	0,11	0,04	0,10	0,10	
										0,36	0,55	0,58	0,55	0,24	0,30	0,52	0,57	0,50	0,36	0,28	*CON*
										0,05	0,08	0,08	0,08	0,03	0,05	0,08	0,10	0,08	0,04	0,04	
											0,53	0,57	0,54	0,23	0,31	0,52	0,53	0,49	0,31	0,25	*LS1*
											0,07	0,08	0,08	0,03	0,06	0,08	0,09	0,09	0,03	0,03	
												0,34	0,32	0,21	0,24	0,29	0,35	0,25	0,27	0,23	*MI1*
												0,03	0,03	0,03	0,05	0,03	0,05	0,03	0,03	0,04	
													0,28	0,20	0,22	0,25	0,32	0,23	0,26	0,20	*MI2*
													0,03	0,03	0,04	0,02	0,05	0,03	0,03	0,03	
														0,21	0,23	0,30	0,34	0,26	0,27	0,23	*MI3*
														0,03	0,05	0,03	0,05	0,03	0,03	0,03	
															0,42	0,61	0,64	0,58	0,45	0,37	*FIN*
															0,09	0,11	0,13	0,13	0,05	0,06	
																0,63	0,66	0,59	0,50	0,42	*RUE*
																0,10	0,13	0,11	0,07	0,07	
																	0,49	0,24	0,42	0,36	*CO2*
																	0,08	0,02	0,07	0,07	
																		0,32	0,30	0,25	*MER*
																		0,04	0,04	0,04	
																			0,51	0,44	*CO3*
																			0,09	0,09	
																				0,23	*LS2*
																				0,03	

Tableau 6

L'indépendance lexicale. Indice global

	LET	*HER*	*AUT*	*DAM*	*BOR*	*TUD*	*RUY*	*RAY*	*RHI*	*CO1*	*CON*	*LS1*	*MI1*	*MI2*	*MI3*	*FIN*	*RUE*	*CO2*	*MER*	*CO3*	*LS2*	*LS3*
LET	0	100	104	89	101	96	98	106	91	75	102	107	86	86	88	107	114	82	93	86	102	109
HER	100	0	92	88	93	89	88	94	89	96	92	89	92	92	91	95	103	96	97	100	90	96
AUT	104	92	0	88	106	107	99	76	87	100	78	81	90	91	91	82	92	97	95	104	80	88
DAM	89	88	88	0	89	90	86	88	68	86	83	82	71	72	72	87	93	85	78	90	81	88
BOR	101	93	106	89	0	90	91	107	91	97	102	98	92	90	92	105	108	96	98	100	99	102
TUD	96	89	107	90	90	0	89	106	90	93	98	96	91	90	91	103	107	92	95	96	95	101
RUY	98	88	99	86	91	89	0	95	87	95	90	89	85	87	86	96	100	92	91	97	90	94
RAY	106	94	76	88	107	106	95	0	86	98	74	80	88	89	88	83	87	96	94	102	78	85
RHI	91	89	87	68	91	90	87	86	0	82	81	79	68	69	70	85	89	79	76	85	78	88
COI	75	96	100	86	97	93	95	98	82	0	99	101	79	79	81	104	107	71	91	75	98	103
CON	102	92	78	83	102	98	90	74	81	99	0	67	81	81	80	66	76	91	87	99	65	73
LSI	107	89	81	82	98	96	89	80	79	101	67	0	80	83	82	68	82	95	83	101	61	72
MI1	86	92	90	71	92	91	85	88	68	79	81	80	0	61	63	85	89	75	71	80	78	87
MI2	86	92	91	72	90	90	87	89	69	79	81	83	61	0	62	87	89	73	72	81	81	87
MI3	88	91	91	72	92	91	86	88	70	81	80	82	63	62	0	85	88	77	70	82	79	88
FIN	107	95	82	87	105	103	96	83	85	104	66	68	85	87	85	0	84	97	88	103	65	73
RUE	114	103	92	93	108	107	100	87	89	107	76	82	89	89	88	84	0	101	94	106	78	84
CO2	82	96	97	85	96	92	92	96	79	71	91	95	75	73	77	97	101	0	84	64	92	97
MER	93	97	95	78	98	95	91	94	76	91	87	83	71	72	70	88	94	84	0	90	82	90
CO3	86	100	104	90	100	96	97	102	85	75	99	101	80	81	82	103	106	64	90	0	97	102
LS2	102	90	80	81	99	95	90	78	78	98	65	61	78	81	79	65	78	92	82	97	0	70
LS3	109	96	88	88	102	101	94	85	88	103	73	72	87	87	88	73	84	97	90	102	70	0

d'un texte à l'égard de lui-même est nulle, on a convenu que cette distance de soi à soi est nécessairement la plus faible de chaque série. On a donc cherché dans chaque ligne la valeur minimum et on l'a communiquée à l'emplacement de la diagonale qui marque le croisement d'un texte avec lui-même. Ainsi au zéro de la première ligne on a substitué le nombre 75 qui est le plus faible de la série et qui établit la distance des *Lettres à la fiancée* au premier recueil de *Correspondance.* Ainsi disposées les données sont soumises au programme d'analyse de correspondance[3]. Ce qu'on obtient à la fin des calculs est représenté dans les figures 3 et 4. Les deux premiers facteurs, dont rend compte la figure 3

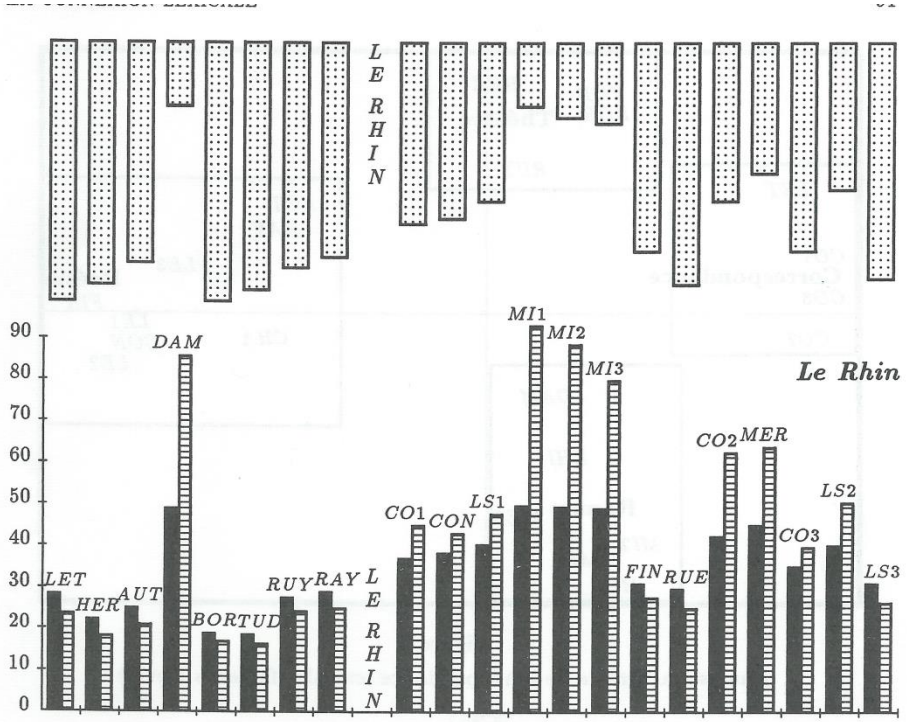

Figure 2
Au-dessus : indépendance lexicale
En dessous : connexion lexicale brute sans considération de fréquence
En noir, les vocables; en hachures, les occurrences.

[3] Ce type d'analyse, dont l'auteur est **J.P.** Benzécri, permet la représentation simultanée des lignes et des colonnes du tableau de données. Mais comme il s'agit ici d'un tableau carré, constitué de deux tableaux triangulaires symétriques, on n'a représenté que les variables (les colonnes), qui se superposent aux individus (aux lignes).

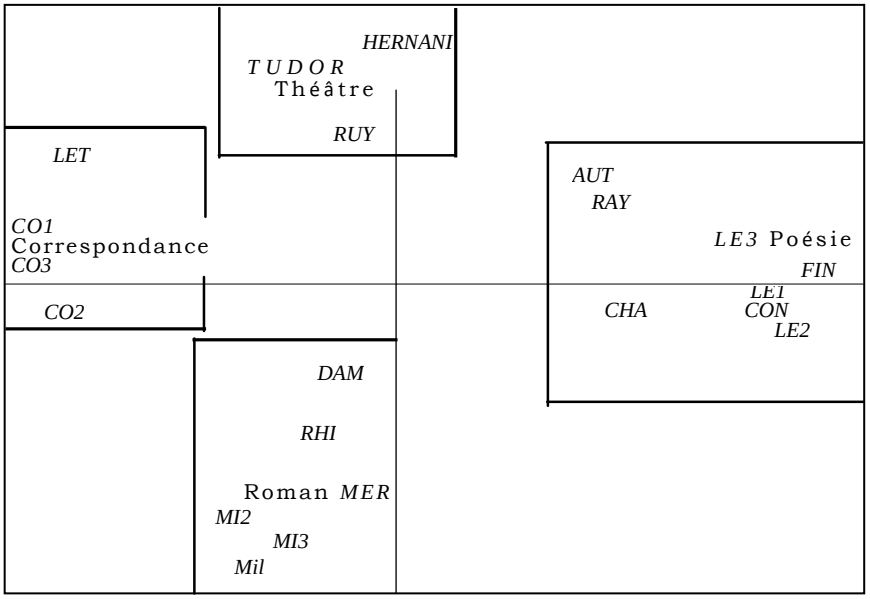

Figure 3

Analyse factorielle de l'indépendance lexicale (facteurs 1 et 2)

sont faciles à identifier. Le premier, qui oppose la gauche et la droite, met en regard la prose et les vers. Le second, qui n'intéresse guère que la prose, distingue le théâtre, en haut, et le roman, en bas. Soumis à ce système de forces, les textes se répartissent en quatre groupes, qui correspondent à un genre bien défini : la correspondance, le roman, le théâtre et la poésie. Hugo semble donc avoir oublié la *Préface de Cromwell,* et les genres littéraires, qu'il prétendait confondre, maintiennent dans son oeuvre leurs distances et leurs spécificités. Le détail même des agglomérations du graphique mérite attention : par exemple, la situation ambiguë de *Hernani* et de *Ruy Blas,* qui appartiennent au théâtre, mais aussi au corpus en vers, explique que ces deux textes hésitent à prendre parti sur le premier facteur et se tiennent à cheval sur l'axe des y . Dans le nuage de points où évolue la poésie, on note aussi la proximité des deux premiers recueils et le regroupement serré qui unit les *Contemplations,* la *Légende* et la *Fin de Satan.* Quant au *Rhin,* sa place aux côtés des romans est confirmée.

L'analyse propose en outre d'autres facteurs, dont le pouvoir discriminant, quoique nettement inférieur (7 % de la variance pour le facteur 3, 2 % pour le facteur 4, alors que le facteur 1 accapare 75 % et le facteur 2, 12 %), n'est pas tout à fait négligeable. Le facteur 3 qui oppose le haut et le bas dans la figure 4

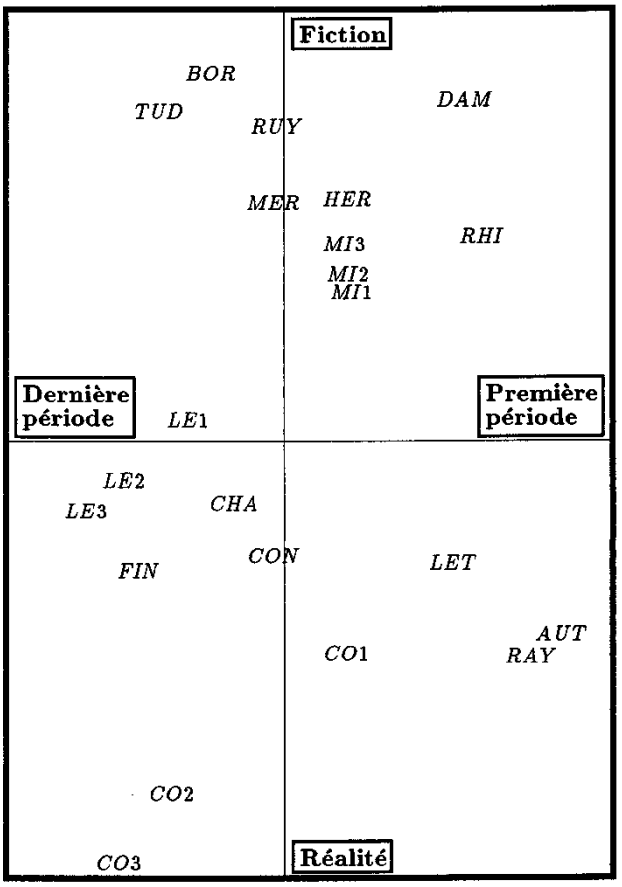

Figure 4
Analyse factorielle de l'indépendance lexicale (facteurs 3 et 4)

est ce qui sépare la fiction (théâtre et roman) de la réalité (correspondance et poésie). Les textes qui se reconnaissent le moins dans cette division sont ceux qui se rapprochent de l'axe horizontal, c'est-à-dire d'un côté le *Rhin* et les *Misérables,* qui contiennent de larges développements où le narrateur s'efface devant l'essayiste et la fiction devant la réalité, et de l'autre la *Légende des siècles* qui "écoute l'histoire aux portes de la légende" et où l'épopée se rapproche de la narration romanesque. Enfin le facteur 4, qui oriente le regard de la droite à la gauche, suit la perspective chronologique. Les premiers textes *(Lettres, Notre-Dame, Feuilles d'automne)* campent sur la marge droite, les derniers*(Légende, Correspondance 2 et 3, Fin de Satan)* occupent la partie gauche. Le temps, même chez un écrivain dont l'empan est aussi large (plus de 60 ans entre le premier et le dernier texte de notre corpus) et qui n'a pas de préjugé hostile au changement, exerce donc moins d'influence sur le vocabulaire que la distinction des genres. L'extrême variété de la production hugolienne ne tient pas à l'évolution

du siècle, non plus qu'au changement personnel, mais à une volonté de toucher à tous les genres et au besoin de les créer.

Les développements qui précèdent ont évalué la connexion (et l'indépendance) lexicale en adoptant pour chaque mot un principe simple mais grossier : pour qu'un vocable figure dans le vocabulaire commun à deux textes, il suffit qu'on le rencontre dans l'un et l'autre. Mais ces rencontres peuvent être insistantes ou fugitives, et plus gravement elles peuvent être fréquentes dans un texte et rares dans l'autre, et dans ce dernier cas la connexion devrait être considérée comme faiblement assurée. Supposons qu'un mot ait 10 occurrences dans un ensemble de deux textes. Dans la logique binaire du tout ou rien qu'on a suivie jusqu'ici, deux cas seulement étaient envisagés : ou bien le mot appartient aux deux textes ou bien il est particulier à l'un des deux. Or il est de multiples façons de partager ces 10 occurrences : de la répartition égale (5 et 5) si les textes sont de même étendue ou proportionnelle si l'étendue est différente, jusqu'au partage léonin où le même texte s'approprie toutes les occurrences. Entre ces deux répartitions opposées, la diversité des combinaisons (1 et 9, 2 et 8, etc.) devrait être prise en compte dans le calcul de la connexion lexicale, celle-ci croissant avec l'équité des partages, et l'indépendance s'accordant avec leur irrégularité. Pour voir ce qui se passe dans la réalité, nous avons relevé la répartition des 4 729 vocables qui appartiennent à l'ensemble formé par les deux premiers textes, *Lettres à la fiancée* et *Hernani* (soit 3 731 unités dans le premier, 2 476 dans le second, et 1 478 à l'intersection). Chacun de ces mots ayant une fréquence dans le premier texte (qui peut être nulle) et une fréquence dans le second, les effectifs observés sont consignés dans un tableau à deux dimensions (tableau 7) où l'abscisse est réservée à la fréquence dans les *Lettres* et l'ordonnée à celle relevée dans *Hernani.*

La ligne et la colonne 20 regroupent les effectifs des vocables dont la fréquence dépasse cette limite dans l'un ou l'autre texte. Il y en a 137 seulement (des mots grammaticaux certainement) qui ont 20 occurrences ou plus dans les deux textes. De ce tableau, les indices bruts de connexion ne considéraient que le total général (4 729), le total de la première colonne (c'est le vocabulaire exclusif des *Lettres,* soit 2 253), le total de la première ligne (ou vocabulaire exclusif de *Hernani,* soit 998), et le vocabulaire commun aux deux textes (tout ce qui n'appartient ni à la première colonne, ni à la première ligne, soit 1 478 vocables). Ici les informations sont beaucoup plus riches, mais leur

traitement est plus complexe. Car les observations n'ont de sens que si on peut les rapporter à un modèle. La règle de trois suffit pour apprécier le cas des hapax (on en dénombre 1 009 dans les *Lettres,* et 640 dans *Hernani,* ces deux effectifs étant répartis sur la première diagonale montante). Les hapax devraient en effet se distribuer proportionnellement à l'étendue de chacun des deux textes (qui comptent respectivement 93 890 et 22 833 occurrences)[4]. Mais dès la deuxième diagonale, qui reproduit la répartition de la classe 2, les choses se compliquent et le recours à la loi binomiale devient inévitable. Les trois effectifs théoriques de la classe 2 répondent alors à la formule :

$$(p+q)^2 = p^2 + 2pq + q^2,$$

p étant le rapport d'étendue du premier texte à l'ensemble et *q* le rapport du second au même ensemble. Pour les classes de fréquence suivantes (en progressant dans la série des diagonales montantes), le binome doit être élevé à la puissance adéquate, en respectant la formulation générale :

$$(p+q)^n = p^n + np^{n-1}q + C_n^{n-2}p^{n-2}q^2 + \cdots + C_n^2 p^2 q^{n-2} + npq^{n-1} + q^n.$$

Les tableaux 8 rend compte d'un autre couple qui croise les *Contemplations* et les *Travailleurs de la mer* et où le détail des calculs est poussé jusqu'à la fréquence 9, dans les deux textes, ce qui génère 44 cellules. Les effectifs théoriques sont alors rapprochés des observations et les écarts transformés en x^2, du moins aussi longtemps que le calcul reste valide, c'est-à-dire tant que la fréquence théorique reste supérieure à 5. Il y a regroupement au-delà de cette limite, une cellule résiduelle accueillant les effectifs trop faibles. On a plusieurs fois constaté que la spécialisation lexicale provoquait des surplus sur les marges, là où le vocabulaire est exclusif, et des déficits dans la zone commune, c'est-à-dire au voisinage de la diagonale descendante. C'est ce qu'on observe une nouvelle fois : la première ligne qui est dévolue aux *Travailleurs de la mer* (et qui correspond à la fréquence 0 dans les *Contemplations)* est uniformément constituée d'écarts positifs : 414, 396, 256, 171, 111, 86, 79, 51, 41. Et il en est de même de la première colonne, réservée aux *Contemplations'* . Quant aux écarts relevés sur la diagonale, ils sont en revanche négatifs : —431, —108, —60, —27 et il en est ainsi pour tout le secteur médian, tandis que les écarts positifs gagnent les pointes du triangle, là où la répartition entre les deux textes est déséquilibrée. Cette

4 L'avantage est à *Hernani* qui ne représente que 1/5 de l'ensemble et qui obtient 2/5 des hapax.

5 Le cas des hapax est à mettre à part, car il n'y a pas de cellule commune qui puisse les recevoir sur la diagonale, et l'excédent dans un des deux textes produit mécaniquement un déficit de même valeur dans l'autre texte. L'avantage (+ 414) est naturellement au roman.

Tableau 7

La connexion lexicale

Croisement *Lettres – Hernani*

(Les effectifs de la classe 20 englobent les classes supérieures)

Fréq.	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
0	0	640	161	76	41	17	15	11	7	4	3	3	3	1	0	0	3	1	0	2	10
1	1 009	145	44	16	10	6	8	6	4	5	2	0	0	0	0	0	0	1	0	0	4
2	391	68	31	19	7	9	2	0	0	3	1	1	2	1	0	0	0	0	0	0	4
3	199	54	18	17	8	5	3	4	2	1	3	0	0	0	0	0	0	0	0	0	4
4	131	34	17	9	6	3	3	2	0	1	0	1	0	0	0	0	0	0	0	0	2
5	105	34	12	11	6	5	3	2	2	2	1	0	0	1	1	0	0	0	0	0	0
6	88	23	12	7	3	4	2	0	2	1	0	2	2	0	1	0	0	1	1	0	1
7	54	21	4	8	7	6	2	1	1	1	0	2	1	0	0	0	0	1	0	0	0
8	41	13	6	6	8	3	0	1	1	1	0	0	0	0	1	0	1	0	0	0	0
9	33	11	8	5	4	1	1	0	1	0	0	0	0	0	1	0	0	0	1	0	0
10	30	4	10	3	2	1	0	1	2	2	0	0	0	1	0	0	1	0	0	1	1
11	17	9	4	2	4	1	1	2	0	1	0	0	1	0	0	0	0	0	0	0	0
12	20	7	4	3	4	0	1	1	0	1	0	0	0	1	0	0	1	0	0	0	1
13	14	4	7	5	2	3	3	0	2	0	0	0	0	0	0	0	0	0	0	0	1
14	17	8	4	0	1	2	1	0	0	0	0	1	0	0	0	0	0	0	0	0	2
15	10	3	3	6	2	2	0	0	3	0	0	0	0	1	0	0	0	0	0	0	0
16	15	4	4	2	3	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	2
17	5	4	3	0	1	1	1	0	1	0	0	1	0	1	1	0	0	0	0	0	0
18	8	1	1	1	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	2
19	6	2	1	2	2	1	0	0	0	0	1	0	0	0	0	0	0	1	0	0	2
20	60	44	36	29	28	17	13	18	11	7	14	8	8	7	10	8	6	8	1	8	137

Tableaux 8

La connexion affinée. Détail des calculs

Croisement *Contemplations – Travailleurs de la Mer*

8.1.– Tableau réel

Fréq.	0	1	2	3	4	5	6	7	8	9
0	0	2 651	1 000	468	253	149	101	88	55	43
1	1 051	360	205	147	95	52	64	26	16	
2	295	167	103	81	50	38	31	20		
3	120	65	52	26	27	27	25			
4	48	53	28	21	25	12				
5	39	33	22	16	18					
6	26	14	9	19						
7	17	18	17							
8	19	8								
9	5									

8.2.– Tableau théorique

Fréq.	0	1	2	3	4	5	6	7	8	9
0	0	2 237	604	212	82	38	15	9	4	2
1	1 465	791	416	215	124	60	39	21	12	
2	259	272	211	162	99	77	48	30		
3	59	92	106	86	84	63	46			
4	15	35	42	55	52	4	6			
5	5	11	22	27	30					
6	1	5	9	13						
7	0	2	4							
8	0	1								
9	0									

observation a été renouvelée dans les 231 tableaux analogues à celui-ci et le fait de la spécialisation lexicale n'est donc pas douteux. On l'explique aisément par des raisons thématiques, stylistiques, parfois chronologiques, par tout ce qui constitue les contraintes de la situation.

Notre objet n'est pourtant pas de vérifier ce fait d'expérience, mais de le mesurer, et de comparer les mesures d'un couple à l'autre. Or le test du x^2 , qui est cumulable, permet d'obtenir une valeur globale qui établit une distance dans le contenu lexical de deux textes. Cette valeur globale n'est pas exhaustive, puisqu'on n'a considéré que les classes de fréquence inférieures à 10. Ce filtre imposé par le caractère exponentiel des calculs retient pourtant 80 % du vocabulaire (8 417 sur 10 471 dans l'exemple du tableau 8). Et il ne laisse à l'écart que les fréquences hautes, qui sont aussi les moins intéressantes pour le point de vue qui nous occupe. La valeur du x^2 obtenue (6115) est donc largement représentative. Mais que signifie-t-elle? Pour l'apprécier, il faut tenir compte des degrés de liberté (ici 38) et convertir en écart réduit une valeur si monstrueuse qu'elle dépasse les tables courantes. La formule de transformation :

$$z = \sqrt{(2\ X^2)} - \sqrt{(ddl - 1)}$$

a été appliquée aux 231 distances, comme le montrent les tableaux 9 à 12.

Tableaux 8 (suite)
La connexion affinée. Détail des calculs
Croisement *Contemplations – Travailleurs de la Mer*

8.3.– Tableau des écarts

Fréq.	0	1	2	3	4	5	6	7	8	9
0	0	414	396	256	171	111	86	79	51	41
1	−414	−431	−211	−68	−29	−8	25	5	4	
2	36	−105	−108	−81	−49	−39	−17	−10		
3	61	−27	−54	−60	−57	−36	−21			
4	33	18	−14	−34	−27	−34				
5	34	22	0	−11	−12					
6	25	9	0	6						
7	17	16	13							
8	19	7								
9	5									

8.4.– Tableau des χ^2

Fréq.	0	1	2	3	4	5	6	7	8	9
0	0,0	76,5	259,6	309,6	355,3	327,1	476,2	736,7	0,0	0,0
1	116,9	234,8	107,1	21,6	6,7	1,2	15,6	1,2	1,7	
2	5,0	40,8	55,5	40,5	24,2	19,8	6,2	3,5		
3	61,6	8,0	27,6	42,2	38,8	20,7	9,9			
4	71,7	9,6	4,9	21,1	13,8	24,7				
5	0,0	43,1	0,0	4,5	4,7					
6	0,0	0,0	0,0	2,7						
7	0,0	0,0	0,0							
8	0,0	0,0								
9	0,0									

8.5.– Données relatives au couple (totaux des diagonales)

Fréq.	1	2	3	4	5	6	7	8	9
	3 702	1 654	960	616	469	316	291	226	183
								Total :	8 417

Mots considérés dans le couple : 8 417
Mots considérés dans *Contemplations* : 3 609 (sur 5 892)
Mots considérés dans *Travailleurs de la mer* : 6 797 (sur 8 824)
χ^2 (regroupement des effectifs faibles) : 2 462
χ^2 (total) : 6 115
Degrés de liberté : 38

On comparera la valeur obtenue (101) dans l'exemple choisi, qui est hétérogène puisqu'il associe un roman et un texte poétique, à celle d'un couple uni, par exemple celui que forment les *Contemplations* et la *Légende 1* (32). Il faut prendre garde que ces valeurs représentent des distances et que la connexion est d'autant plus forte que la distance est faible. Les *Contemplations* se rapprochent aussi des *Feuilles d'automne* (22), des *Rayons et des ombres* (15), de la *Fin de Satan* (15), et de façon générale de tous les textes poétiques, alors que la distance augmente lorsque le terme de référence est un roman (respectivement 90, 86, 86, 89 et 101), un recueil de lettres (151, 136, 144, 134), ou une pièce de théâtre (52, 51, 56, 51).

On a tout lieu d'être satisfait, d'autant que la séparation des genres, et plus discrètement l'évolution chronologique, produisent partout des effets aussi remarquables, ce qui peut se vérifier en suivant n'importe quelle ligne ou n'importe quelle colonne — n'importe quel texte — du tableau 10. Mais si on compare les lignes entre elles, on ne peut pas être aveugle à l'inégalité de traitement qui frappe les textes longs et les textes courts. Les premiers ont en général des valeurs plus fortes. La raison tient à l'effet de taille auquel les mesures probabilistes sont toujours sensibles. Pour des écarts proportionnellement semblables — par exemple un excédent ou un déficit de 10 % — le X^2 et l'écart réduit auront des valeurs très différentes selon que le calcul porte sur des milliers ou des millions d'observations. Ces tests mesurent la probabilité ou plus souvent l'improbabilité des faits observés, si le hasard était seul en cause, et non pas directement l'importance de l'écart[6]. Et l'on peut rencontrer d'amples écarts aléatoires (dans les petits nombres), et, aussi bien, des écarts faibles en pourcentage, où l'hypothèse nulle doit pourtant être rejetée (dans les grands nombres). C'est pourquoi ces tests trouvent parfois des détracteurs, déçus de n'y point trouver l'invariabilité d'un mètre-étalon. Cela est gênant lorsqu'on cherche moins à établir avec certitude la réalité d'un écart fonctionnel (et ici le phénomène de la spécialisation n'a plus besoin d'être démontré), qu'à classer des écarts qui tous échappent au hasard. Il faut donc éliminer dans ces écarts ce qui tient aux variations de taille. Pour remplir cet office de pondération, on avait proposé jadis la racine carrée du vocabulaire[7]. Nous avons appliqué ce coefficient à toutes les valeurs du tableau 10, le résultat étant reproduit au bas du même tableau. Par exemple l'écart 79,1 qu'on trouve en première ligne et deuxième colonne (c'est le croisement des *Lettres à la fiancée* avec *Hernani)* est divisé par V2 476 , soit 49,75 (2 476 étant le vocabulaire de la pièce).

---

[6] L'importance de l'écart compte aussi dans la valeur de ces tests, mais conjuguée à l'étendue des observations.

[7] Rien ne justifie en théorie ce choix plutôt qu'un autre. L'expérience montre pourtant — et pas seulement en cette occasion — que -VVI est souvent le meilleur coefficient de pondération.

Tableau 9

La connexion lexicale. Le χ^2 pour chaque couple

HER	*AUT*	*DAM*	*BOR*	*TUD*	*RUY*	*RAY*	*RHI*	*CO1*	*CON*	*LS1*	*MI1*	*MI2*	*MI3*	*FIN*	*RUE*	*CO2*	*MER*	*CO3*	*LS2*	*LS3*	
3752	8981	9827	2654	2075	6817	9858	9972	2353	12753	20151	8124	8565	9789	13236	15571	4267	13818	5920	14749	13890	*LET*
	1428	1135	591	742	593	1606	1271	2652	1809	1317	1422	1561	1595	1984	2017	2998	1899	2709	1321	1536	*HER*
		2058	2148	2687	1992	142	1937	5776	476	670	2569	2258	2190	792	968	7818	2660	8046	538	675	*AUT*
			983	1308	975	1562	4055	7601	4964	3930	3213	3480	3778	3580	3126	13710	6591	8774	4932	2530	*DAM*
				468	734	2475	970	1886	1750	1568	955	992	1111	2017	2188	1633	1592	1730	1274	1774	*BOR*
					1121	2932	1338	1256	2051	2126	1271	1302	1288	2783	2939	1339	1942	1322	1645	2505	*TUD*
						1746	1188	3641	1794	1663	965	905	913	2517	2077	3725	1840	3826	1609	1589	*RUY*
							1273	7254	288	619	1882	1679	1810	803	677	7768	2739	6701	465	607	*RAY*
								7289	4065	3618	3099	3979	3766	2918	2603	12106	6068	7072	4282	1632	*RHI*
									10473	14921	5189	5735	6201	12131	11584	1329	10804	1494	12794	9969	*CO1*
										1016	4501	4575	4786	293	903	11654	6115	10276	435	482	*CON*
											4448	4278	3677	454	744	19696	4907	12969	153	81	*LS1*
												752	603	3363	3385	7237	3382	4961	4648	2522	*MI1*
													687	3395	2419	8040	3151	5437	4884	2129	*MI2*
														2684	2213	8801	2444	5557	4080	1963	*MI3*
															1231	10570	3646	10847	270	279	*FIN*
																11552	3247	10563	735	399	*RUE*
																	15952	328	11653	9893	*CO2*
																		9970	5201	2298	*MER*
																			10772	7473	*CO3*
																				89	*LS2*

Tableau 10

La connexion lexicale. Les degrés de liberté

HER	*AUT*	*DAM*	*BOR*	*TUD*	*RUY*	*RAY*	*RHI*	*CO1*	*CON*	*LS1*	*MI1*	*MI2*	*MI3*	*FIN*	*RUE*	*CO2*	*MER*	*CO3*	*LS2*	*LS3*	
29	31	38	28	30	33	32	38	37	37	39	38	38	39	36	33	36	39	38	38	35	*LET*
	34	27	32	32	34	35	27	28	32	33	27	28	28	34	35	26	29	29	31	35	*HER*
		28	33	32	34	34	28	29	32	33	28	28	28	34	35	27	30	31	32	35	*AUT*
			27	28	29	29	40	39	39	37	39	40	40	33	29	39	40	38	38	31	*DAM*
				31	33	35	26	28	30	32	26	26	27	33	34	25	28	28	29	33	*BOR*
					34	34	28	29	32	33	28	28	28	35	33	27	30	29	32	35	*TUD*
						35	28	31	33	36	29	29	29	36	35	28	31	32	33	34	*RUY*
							28	31	33	35	29	28	29	34	35	28	31	31	33	35	*RAY*
								39	37	37	40	40	41	33	29	39	41	39	39	30	*RHI*
									38	38	38	39	39	36	32	38	38	37	39	33	*CO1*
										38	39	39	39	35	34	38	38	38	37	36	*CON*
											37	37	37	38	37	37	39	38	39	36	*LS1*
												39	40	33	30	39	40	39	38	31	*MI1*
													40	33	29	39	40	39	39	31	*MI2*
														33	30	40	40	38	39	32	*MI3*
															36	32	36	36	36	36	*FIN*
																29	32	33	33	34	*RUE*
																	40	39	37	30	*CO2*
																		39	39	33	*MER*
																			38	33	*CO3*
																				36	*LS2*

Tableau 11

Transformation des χ^2 en écarts réduits

	LET	*HER*	*AUT*	*DAM*	*BOR*	*TUD*	*RUY*	*RAY*	*RHI*	*CO1*	*CON*	*LS1*	*MI1*	*MI2*	*MI3*	*FIN*	*RUE*	*CO2*	*MER*	*CO3*	*LS2*	*LS3*
LET		79,1	126,2	131,5	65,4	56,7	108,7	132,5	132,6	60,1	151,2	192,0	118,8	122,2	131,1	154,3	168,4	84,0	157,5	100,1	163,1	158,9
HER	79,1		45,3	40,4	26,4	30,6	26,3	48,4	43,1	65,4	52,2	43,3	46,0	48,5	49,1	54,8	55,2	70,3	54,1	66,1	43,6	47,1
AUT	126,2	45,3		56,7	57,5	65,4	54,9	8,7	54,8	99,9	22,9	28,6	64,3	59,8	58,8	31,6	35,7	117,8	65,3	119,0	24,9	28,4
DAM	131,5	40,4	56,7		37,1	43,7	36,6	48,4	81,2	114,5	90,9	80,1	71,4	74,5	78,0	76,6	71,5	156,8	105,9	123,8	90,7	63,3
BOR	65,4	26,4	57,5	37,1		22,8	30,2	62,0	36,9	54,0	51,5	48,1	36,6	37,4	39,9	55,5	58,0	50,1	49,0	51,4	42,9	51,5
TUD	56,7	30,6	65,4	43,7	22,8		39,2	68,4	44,3	42,6	56,1	57,1	43,0	43,6	43,3	66,3	68,6	44,5	54,6	43,9	49,4	62,5
RUY	108,7	26,3	54,9	36,6	30,2	39,2		50,8	41,3	77,5	51,8	49,2	36,4	35,0	35,2	62,5	56,1	78,9	52,8	79,5	48,7	48,2
RAY	132,5	48,4	8,7	48,4	62,0	68,4	50,8		43,0	112,6	15,9	26,9	53,8	50,5	52,6	31,9	28,5	117,2	66,2	108,0	22,4	26,5
RHI	132,6	43,1	54,8	81,2	36,9	44,3	41,3	43,0		112,0	81,6	76,5	69,8	80,3	77,8	68,3	64,6	146,8	101,2	110,2	83,8	49,4
CO1	60,1	65,4	99,9	114,5	54,0	42,6	77,5	112,6	112,0		136,1	164,1	93,2	98,3	102,6	147,3	144,3	42,9	138,3	46,1	151,2	133,1
CON	151,2	52,2	22,9	90,9	51,5	56,1	51,8	15,9	81,6	136,1		36,4	86,1	86,9	89,1	15,9	34,3	144,0	101,9	134,7	21,0	22,6
LS1	192,0	43,3	28,6	80,1	48,1	57,1	49,2	26,9	76,5	164,1	36,4		85,8	84,0	77,2	21,5	30,0	189,9	90,3	152,4	8,7	4,3
MI1	118,8	46,0	64,3	71,4	36,6	43,0	36,4	53,8	69,8	93,2	86,1	85,8		30,0	25,8	74,0	74,6	111,5	73,4	90,8	87,8	63,2
MI2	122,2	48,5	59,8	74,5	37,4	43,6	35,0	50,5	80,3	98,3	86,9	84,0	30,0		28,2	74,3	62,0	118,0	70,5	95,5	90,1	57,4
MI3	131,1	49,1	58,8	78,0	39,9	43,3	35,2	52,6	77,8	102,6	89,1	77,2	25,8	28,2		65,2	58,8	123,8	61,0	96,8	81,6	54,7
FIN	154,3	54,8	31,6	76,6	55,5	66,3	62,5	31,9	68,3	147,3	15,9	21,5	74,0	74,3	65,2		41,2	137,5	77,0	138,9	14,8	15,2
RUE	168,4	55,2	35,7	71,5	58,0	68,6	56,1	28,5	64,6	144,3	34,3	30,0	74,6	62,0	58,8	41,2		144,4	72,6	137,3	30,3	20,0
CO2	84,0	70,3	117,8	156,8	50,1	44,5	78,9	117,2	146,8	42,9	144,0	189,9	111,5	118,0	123,8	137,5	144,4		169,7	16,8	144,1	133,0
MER	157,5	54,1	65,3	105,9	49,0	54,6	52,8	66,2	101,2	138,3	101,9	90,3	73,4	70,5	61,0	77,0	72,6	169,7		132,4	93,2	59,7
CO3	100,1	66,1	119,0	123,8	51,4	43,9	79,5	108,0	110,2	46,1	134,7	152,4	90,8	95,5	96,8	138,9	137,3	16,8	132,4		138,1	114,2
LS2	163,1	43,6	24,9	90,7	42,9	49,4	48,7	22,4	83,8	151,2	21,0	8,7	87,8	90,1	81,6	14,8	30,3	144,1	93,2	138,1		4,9
LS3	158,9	47,1	28,4	63,3	51,5	62,5	48,2	26,5	49,4	133,1	22,6	4,3	63,2	57,4	54,7	15,2	20,0	133,0	59,7	114,2	4,9	

Tableau 12
Transformation des χ^2 en écarts réduits
Pondération par $\sqrt{V}$ [V appartenant à la colonne]
(Valeurs multipliées par 100)

	LET	HER	AUT	DAM	BOR	TUD	RUY	RAY	RHI	CO1	CON	LS1	MI1	MI2	MI3	FIN	RUE	CO2	MER	CO3	LS2	LS3
LET		159	230	137	139	122	189	228	133	81	204	250	125	122	137	239	259	101	168	135	210	246
HER	129		83	42	56	66	46	83	43	88	71	56	48	48	51	85	85	84	58	89	56	73
AUT	207	91		59	122	140	96	15	55	134	31	37	67	60	61	49	55	141	69	161	32	44
DAM	215	81	104		78	94	64	83	81	154	123	104	75	75	81	118	110	188	113	167	116	98
BOR	107	53	105	39		49	53	107	37	73	70	63	38	37	42	86	89	60	52	69	55	80
TUD	93	61	119	46	48		68	118	44	57	76	74	45	44	45	103	106	53	58	59	63	97
RUY	178	53	100	38	64	84		87	41	104	70	64	38	35	37	97	86	95	56	107	63	74
RAY	217	97	16	50	131	147	88		43	152	22	35	56	51	55	49	44	141	70	146	29	41
RHI	217	87	100	85	78	95	72	74		151	110	100	73	80	81	106	100	176	108	149	108	76
CO1	98	131	182	120	114	91	135	194	112		184	214	98	98	107	228	222	51	147	62	194	206
CON	247	105	42	95	109	120	90	27	82	183		47	90	87	93	25	53	173	109	182	27	35
LS1	314	87	52	84	102	122	86	46	77	221	49		90	84	81	33	46	228	96	206	11	7
MI1	195	93	117	75	77	92	63	92	70	125	116	112		30	27	114	115	134	78	123	113	98
MI2	200	97	109	78	79	93	61	87	81	132	117	109	31		29	115	96	141	75	129	116	89
MI3	215	99	107	82	84	93	61	90	78	138	120	101	27	28		101	91	148	65	131	105	85
FIN	253	110	58	80	117	142	109	55	69	198	21	28	78	74	68		63	165	82	188	19	23
RUE	276	111	65	75	123	147	98	49	65	194	46	39	78	62	61	64		173	77	185	39	31
CO2	137	141	215	164	106	95	137	202	147	58	195	247	117	118	129	213	223		181	23	185	205
MER	258	109	119	111	104	117	92	114	101	186	138	118	77	70	64	119	112	203		179	120	92
CO3	164	133	217	129	109	94	138	186	110	62	182	199	95	95	101	215	212	20	141		177	176
LS2	267	88	45	95	91	106	85	39	84	203	28	11	92	90	85	23	47	173	99	187		8
LS3	260	95	52	66	109	134	84	46	50	179	31	6	66	57	57	24	31	159	64	154	6	

On obtient 1,59 (ou 159 dans le tableau, après application d'un facteur 100). Cette valeur est alors comparable à celle qui résulte du croisement des *Lettres* avec les *Feuilles d'automne,* soit 126, 2/N/3 001 -= 2, 30 , et avec toutes celles qu'on lit sur la première ligne et qui mesurent les distances — pondérées — du premier texte avec tous les autres[8]. Les profils des 22 textes se lisent donc horizontalement, alors que dans le sens vertical, les valeurs d'origine ont été conservées, à une constante près[9].

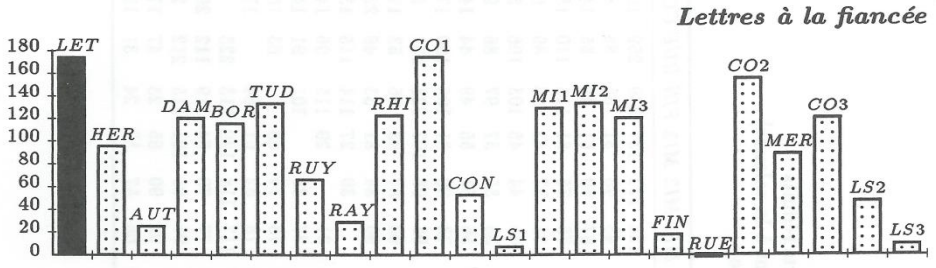

Figure 5

Le résultat, concernant les *Lettres à la fiancée,* est rendu plus lisible dans la représentation graphique de la figure 5. Comme on veut souligner la connexion, et qu'un histogramme rend compte habituellement d'une relation de voisinage et non d'une distance, on a inversé la perspective, les "bâtons" les plus longs se rapportant aux textes dont le contenu lexical est le plus proche de celui des *Lettres.* Or ce sont ceux qui appartiennent au genre épistolaire, et surtout le premier recueil de correspondance (symbole *Col* sur le graphique). Viennent ensuite les pièces en prose et les romans (le *Rhin* y compris). Les textes les plus éloignés des *Lettres* sont ceux qui utilisent le vers. Ce qui est vrai des *Lettres* l'est aussi du genre épistolaire tout entier. Le graphique de la figure 6 superpose les trois recueils de correspondance du corpus, pour les périodes 1814-1848, 1849-1866 et 1867-1873. On sera sensible au parallélisme étroit des trois courbes : les trois textes considérés nouent entre eux des relations privilégiées et manifestent au genre poétique une hostilité déclarée, en même temps qu'une neutralité bienveillante à l'égard du roman et du théâtre en prose. On remarquera que le recouvrement est plus serré dans la partie centrale, qui

[8] En deuxième ligne, première colonne, le même croisement des *Lettres* avec *Hernani* est cette fois pondéré par N/3 731 , soit 61,08 (3 731 étant le vocabulaire des *Lettres),* car c'est le profil de la pièce qu'on veut alors établir.

[9] Par exemple toutes les valeurs de la colonne 1 ont été divisées par la constante 61,01 qui est la racine carrée du vocabulaire des *Lettres.*

représente un passage obligé de la chronologie, et que les courbes commencent à diverger aux deux extrémités. C'est là l'effet — d'amplitude modérée — de l'évolution : le premier recueil *(Col)* est plus proche des premiers textes (et notamment des *Lettres à la fiancée),* et les deux derniers (*Co2* et *Coi)* partagent plus de mots avec les titres de la dernière période.

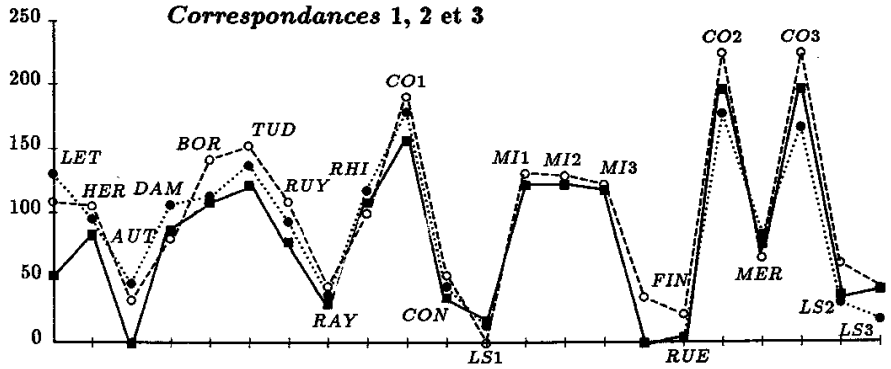

Figure 6

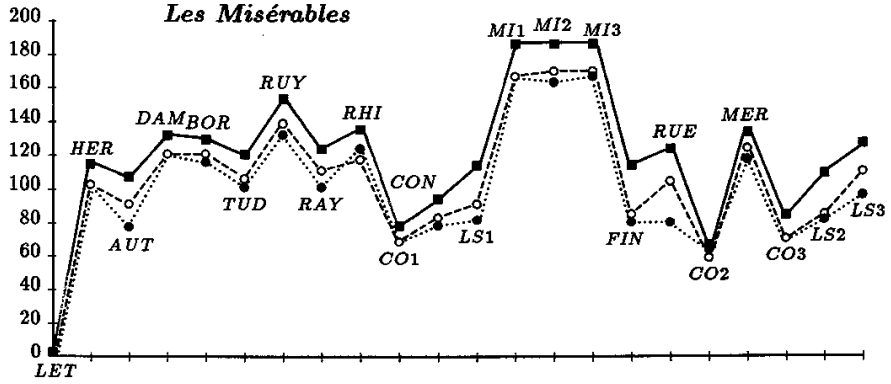

Figure 7

Nous retrouverons plus loin l'effet du temps sur le vocabulaire d'un auteur. Mais restons-en pour l'instant au genre littéraire dont l'influence est prépondérante dans le graphique de la figure 7. Il s'agit des trois sous-ensembles découpés dans les *Misérables.* Comment pourrait-on mieux marquer l'unité de ce roman? La connexion lexicale atteint des sommets quand ces trois parties sont mises en relation. Elle est nettement moins élevée lorsque la comparaison

est faite avec un autre roman, ou avec le théâtre (le *Rhin* se range ici). Elle est plus médiocre lorsqu'intervient la poésie, et plus faible encore à l'endroit de la correspondance.

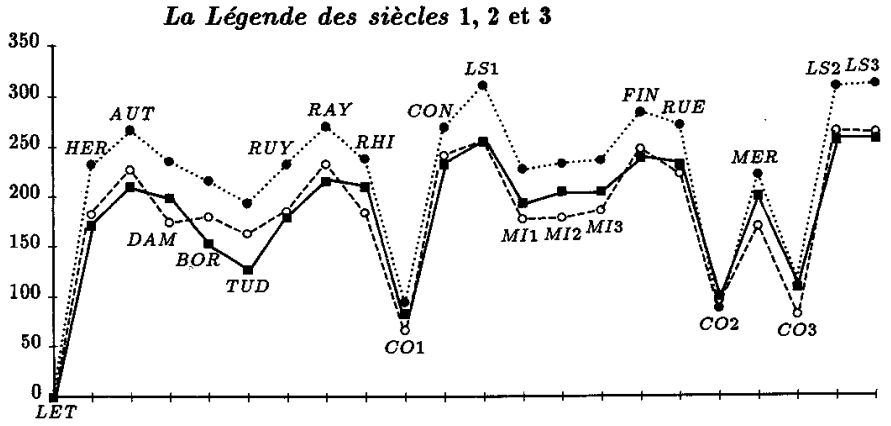

Figure 8

Un dernier exemple, emprunté cette fois à la poésie, suffira à établir la toute-puissance du genre. Les trois recueils de la *Légende des siècles* prennent place simultanément sur le graphique de la figure 8. Là encore le parallélisme est remarquable et les choix solidaires. Non seulement on assiste à une cooptation réciproque des trois recueils, mais aussi à la même préférence donnée plus largement aux textes poétiques : les *Feuilles,* les *Rayons,* les *Contemplations,* les *Chansons* et la *Fin de Satan.* Les pièces en vers *Hernani* et Ruy *Blas* trouvent grâce plus facilement que les autres, et les romans moins malaisément que la correspondance. Ce sont les oeuvres épistolaires qui dans ce graphique comme dans le précédent servent de repoussoir. Et la même observation est vérifiée quand on analyse les autres recueils poétiques, ou les autres romans, ou aussi le théâtre. Pourquoi les recueils de lettres sont-ils ainsi mis au ban de la société lexicale, comme s'il s'agissait d'un parti méprisable avec lequel les autres partis de l'hémicycle ne voudraient rien partager. Pourquoi les votes des grands genres littéraires, même quand ils se combattent, ne se mêlent-ils jamais à ceux de la correspondance? Cela tient, semble-t-il, à deux explications : d'une part le genre épistolaire — du moins lorsqu'il s'agit de vraies lettres, non destinées à la publication[10] — n'a ni le même référent, ni le même registre que les

[10] L'exemple du *Rhin* donne une bien intéressante contre-épreuve. Les lettres largement fictives qui constituent ce récit de voyage orientent l'ouvrage du côté de la littérature de fiction, dans la sphère d'influence romanesque.

genres proprement littéraires. On y désigne des choses différentes (par exemple les événements familiaux, les préoccupations intimes, les questions d'argent, les relations avec les éditeurs[11]) et, quand les objets sont les mêmes, on en parle sur un ton différent, en utilisant un autre registre. D'autre part, il faut bien constater que la correspondance utilise — au moins chez Hugo[12] un — vocabulaire plus pauvre et plus courant[13]. Quand on écrit une simple lettre, on ne consulte pas le dictionnaire. Le stock lexical que la correspondance peut partager avec les autres textes est donc plus réduit. Et faute de production suffisante, ce genre littéraire n'exporte pas et les échanges commerciaux avec les autres genres sont déséquilibrés.

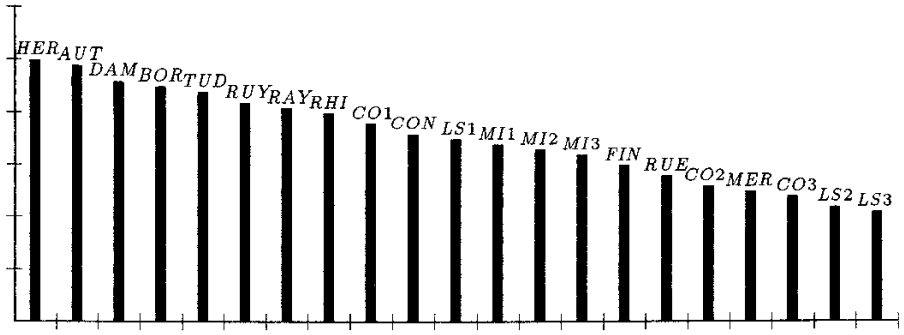

Figure 9
La gravitation temporelle
Droite de tendance à partir du premier texte

La place nous manque pour restituer visuellement 22 profils lexicaux de notre corpus. Il est impossible de dire que ces 22 portraits ont un air de famille Car cela supposerait qu'on puisse comparer cette famille aux autres,

[11] Hugo ne fait pas état de la littérature dans le texte même de ses oeuvres littéraires. De la même façon qu'on ne parle pas de corde chez un pendu, on n'utilise pas le mot *imagination* dans une oeuvre d'imagination, ni le mot *littéraire* dans un texte littéraire. Mais ces préoccupations - et les mots associés — qui sont constantes chez un écrivain, et notamment chez Hugo, alimentent les discours ordinaires et la correspondance. Il est vrai qu'après Proust l'écriture se prend parfois pour objet, et les barrières lexicales tombent.

[12] Cela peut être différent chez d'autres écrivains épistoliers, soit qu'ils voient la postérité à travers leur destinataire, soit que la liberté d'être soi-même permette d'échapper au carcan littéraire et donne libre cours à la truculence lexicale — et nous pensons ici à Flaubert.

[13] Notre étude de la richesse lexicale chez Hugo montre que la variété du lexique est la moins large dans les recueils de lettres, et particulièrement dans les *Lettres à la fiancée,* où l'amoureux reprend indéfiniment l'expression des mêmes sentiments exaltés, en se souciant fort peu d'éviter les répétitions.

et confronter Hugo aux autres auteurs[14]. Mais on peut assurer que dans la production littéraire de Hugo le genre joue un rôle analogue à celui du sexe dans la cellule familiale. Les garçons et les filles n'ont pas les mêmes préoccupations et n'utilisent pas les mêmes mots. Mais n'y aurait-il pas aussi quelque différence entre les aînés et les benjamins? A côté du genre, dont l'influence reste prépondérante, le temps n'a-t-il pas son mot à dire? Déjà le tableau 6, consacré à l'indépendance lexicale, nous proposait un facteur 4 qui coïncidait avec la chronologie. Et l'étude des *Lettres à la fiancée,* premier texte du corpus, laissait deviner une pente descendante, par dessus les accidents du terrain. Nous avons repris cette distribution, en la soumettant au calcul classique de la droite de tendance. En nivelant les creux et les bosses, cette technique permet de faire apparaître l'orientation générale du corpus (figure 9). C'est comme si l'on étudiait l'écoulement des eaux dans un relief tourmenté. Le graphique relatif aux *Lettres à la fiancée* montre bien que les eaux ne sont pas prisonnières et qu'une perspective chronologique se fait jour dans le contenu lexical. L'autre bout de la chaîne est également un lieu privilégié pour mettre en évidence la perspective du temps : en installant son objectif sur le dernier recueil de la *Légence des siècles,* on aperçoit en enfilade tous les textes du corpus (figure 10). Ceux qui sont les plus proches chronologiquement sont aussi ceux où la connexion lexicale est la plus forte, du moins si l'on ne considère que les textes qui appartiennent au même genre littéraire. En s'en tenant aux oeuvres poétiques, *Légende 3* exerce une attraction plus forte sur *Légende 2,* la *Fin de Satan* et *Légende 1,* tandis que s'affaiblit le lien avec les *Feuilles d'automne* et les *Rayons et les ombres.* La connexion est naturellement très basse avec les textes de correspondance, mais là encore la droite de tendance montre bien que la part commune du vocabulaire se réduit quand à l'opposition de genre s'ajoute l'éloignement chronologique.

Il est moins aisé de mettre en relief l'influence du temps, si l'on se place dans la zone centrale de la série, la droite de tendance ayant alors deux tronçons contrariés. On a donc imaginé un subterfuge qui met en évidence l'influence du temps pour tous les éléments, centraux ou périphériques, de la série. Il suffit de relever dans le tableau 10 les paires où la connexion est maximale, en retenant pour chaque texte l'élément privilégié. C'est *Correspondance 1* pour *Lettres à la fiancée, Légende 2* pour *Légende 3,* la *Fin de Satan* pour les *Contemplations,* etc.[15] Reste à transcrire sur un graphique l'ensemble de

[14] Si l'on entreprenait cette recherche — coûteuse — on aborderait un problème intéressant : y a-t-il plus de différences dans le contenu lexical entre des écrivains différents qu'entre plusieurs textes d'un même auteur? Autrement dit les distances *inter* et *intra* sont-elles du même ordre?

[15] Les relations sont souvent symétriques dans ces couples, la préférence de l'un s'accordant à celle de l'autre : Corres*pondance 2* et *3, Légende 2* et *3, Misérables 1* et *3.* Mais il y a aussi des couples mal assortis, où le partenaire choisi porte ailleurs son regard : ainsi les *Contemplations* s'orientent vers la *Fin de Satan,* dont la faveur va *à Légende 2,* qui pour sa part préfère *Légende 3.*

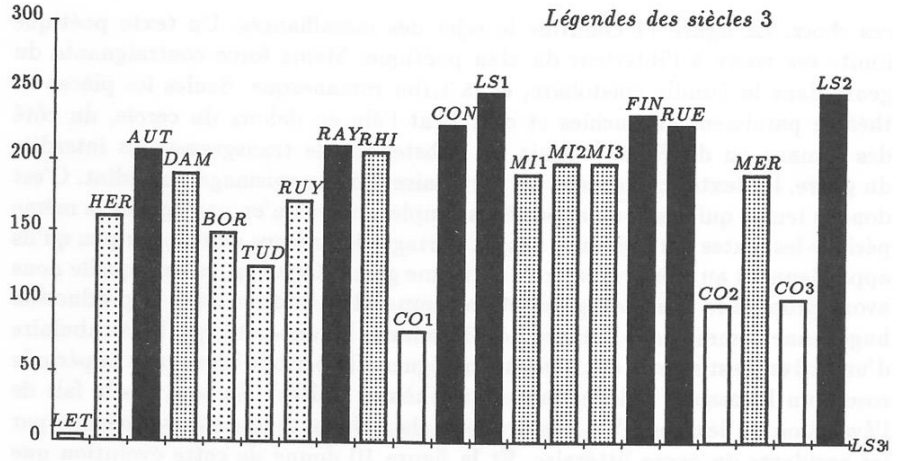

Figure 10

La gravitation temporelle
Droite de tendance à partir du dernier texte

En noir, le genre poétique; en blanc, la correspondance;
En gris clair, le théâtre; en gris foncé, le roman.

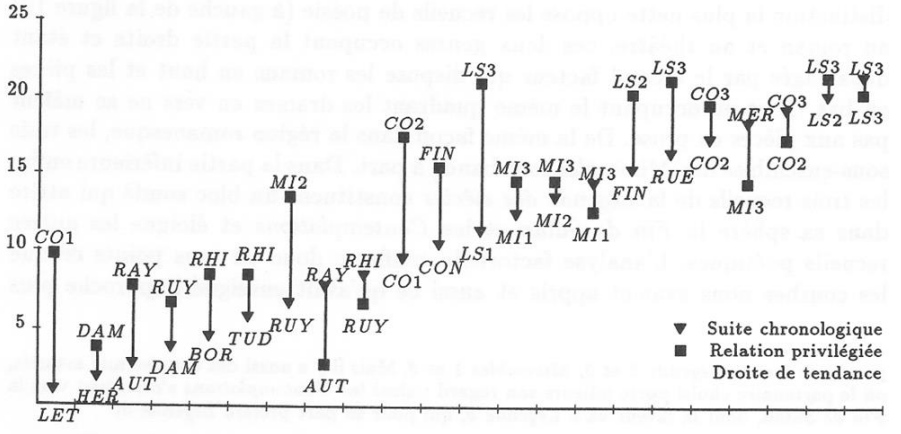

Figure 11

La gravitation temporelle
Graphique de la connexion maximale des couples de textes

ces choix. La figure 11 confirme le rejet des mésalliances. Un texte poétique limite ses voeux à l'intérieur du clan poétique. Même force contraignante du genre dans la famille épistolaire, et la tribu romanesque. Seules les pièces de théâtre paraissent affranchies et cherchent l'élu en dehors du cercle, du côté des romans ou du *Rhin'*. Mais en s'abstenant de transgresser les interdits du genre, les textes choisissent un partenaire dans le voisinage immédiat. C'est donc le temps qui marie ou apparie les couples, parce qu'en partageant la même période les textes ont tendance aussi à partager les mêmes mots, pour peu qu'ils appartiennent au même monde, au même genre. Cette loi, pour laquelle nous avons proposé le terme de gravitation temporelle, s'exerce dans la production hugolienne comme dans l'oeuvre de Giraudoux. Dirons-nous que le vocabulaire d'un écrivain est sujet à des phases, analogues à la période bleue ou à la période rose d'un Picasso et indépendantes des thèmes traités? En tout cas le fait de l'évolution du lexique n'est pas douteux chez Hugo, même s'il est masqué par les accidents du genre littéraire. Et la figure 10 donne de cette évolution une illustration très claire.

Si l'on veut retrouver la même image synthétique que nous avaient fournie les figures 3 et 4 à partir des indices bruts, il convient de donner en pâture à l'analyse factorielle les données plus élaborées du tableau 10. Cette fois il ne s'agit plus tout à fait d'un tableau symétrique, puisque les valeurs ont été pondérées dans le sens horizontal, mais non vertical. On a dû aussi renoncer à intégrer les textes épistolaires qui, trop excentriques, accaparaient l'attention sur eux et monopolisaient le premier facteur[16]. Il ne s'agit donc que des textes du corpus où l'intention littéraire est avouée. Comme il fallait s'y attendre la distinction la plus nette oppose les recueils de poésie (à gauche de la figure 12) au roman et au théâtre, ces deux genres occupant la partie droite et étant départagés par le second facteur qui dispose les romans en haut et les pièces en bas. Tout en occupant le même quadrant les drames en vers ne se mêlent pas aux pièces en prose. De la même façon dans la région romanesque, les trois sous-ensembles des *Misérables* font bande à part. Dans la partie inférieure enfin, les trois recueils de la *Légende des siècles* constituent un bloc soudé qui attire dans sa sphère la *Fin de Satan* et les *Contemplations* et éloigne les autres recueils poétiques. L'analyse factorielle confirme donc en tous points ce que les courbes nous avaient appris et aussi ce qu'avait enseigné l'approche plus

[16] Le *Rhin,* n'appartenant pas à un genre caractérisé, est un élément plus disponible.

[17] Les recueils de lettres réapparaissent pourtant sur le graphique, mais en "variables supplémentaires", c'est-à-dire sans avoir eu la moindre part aux calculs et à la décision.

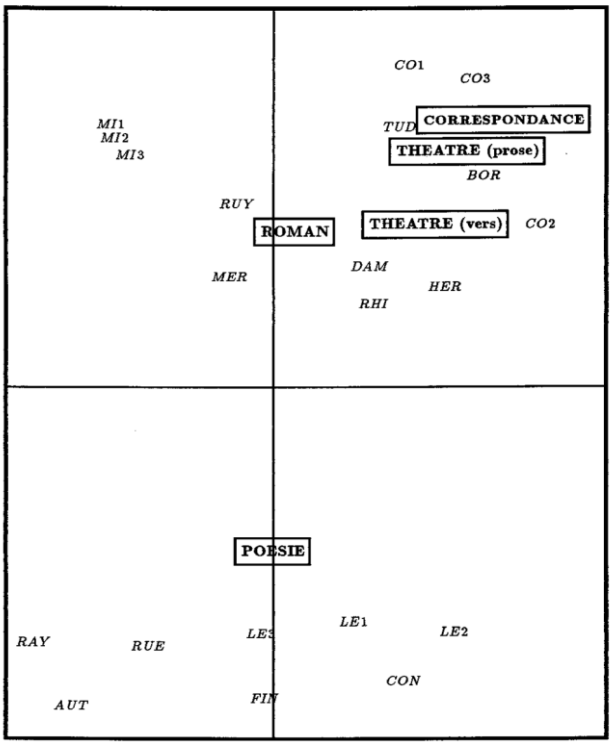

Figure 12
Analyse factorielle de la connexion lexicale
Facteurs 1 et 2

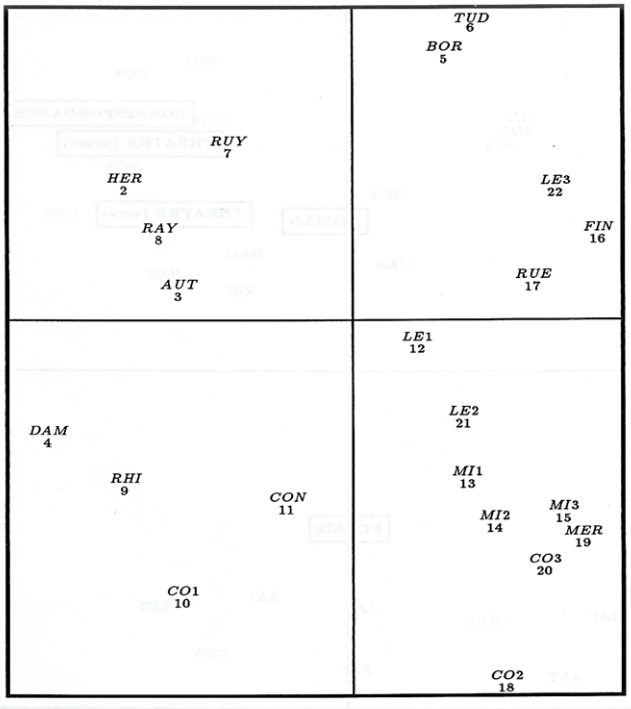

Figure 13
Analyse factorielle de la connexion lexicale
Facteurs 3 et 4. Mise en évidence du temps
(les numéros indiquent le rang chronologique).

grossière des indices bruts[18]. Il n'est pas jusqu'à l'influence du temps qui ne se retrouve dans les résultats, si l'on examine les facteurs 3 et 4 de la même analyse (figure 13). Tous les textes de la première période occupent la partie gauche (sauf les deux pièces en prose), et tous ceux de la seconde période la partie droite. Cette influence de la chronologie passe toutefois loin derrière celle du genre, puisqu'elle n'est sensible qu'au niveau du facteur 4, et que ce dernier ne rend compte que de 3 % de la variance, contre respectivement 76 %, 11 % et 6 % aux trois premiers.

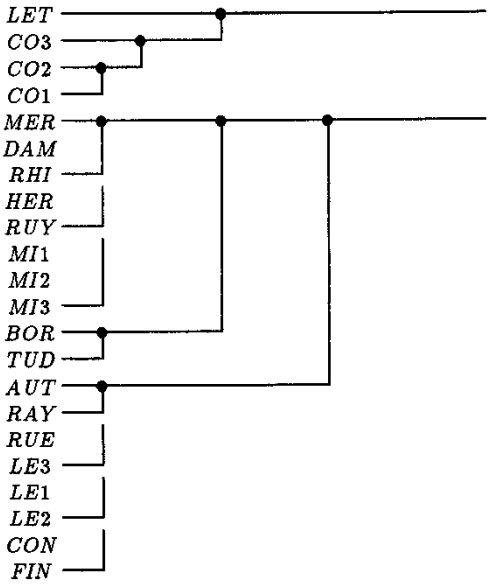

Figure 14
Classement hiérarchique ascendant

Deux autres procédures expérimentales ont été appliquées aux distances établies dans le tableau 10. La première est connue depuis longtemps sous le nom de "classement hiérarchique ascendant". Les enseignements qu'elle nous délivre ne sont pas d'une grande finesse : certes on reconnaît bien l'exterritorialité de la correspondance dont les quatre sous-ensembles font sécession dans la figure 14. A l'opposé les recueils poétiques se détachent du reste, tout en montrant des liaisons particulières : les *Feuilles* avec les *Rayons,*

[18] On ajoute que la qualité de représentation est toujours au-delà de 900 (sur 1 000) et qu'elle dépasse généralement 950.

les *Contemplations* avec la *Fin de Satan.* Mais au milieu, la hiérarchie est un peu hésitante et elle mêle le théâtre et le roman. On doit préférer une méthode nouvelle proposée par Xuan Luong sous le nom d'analyse arborée[19].

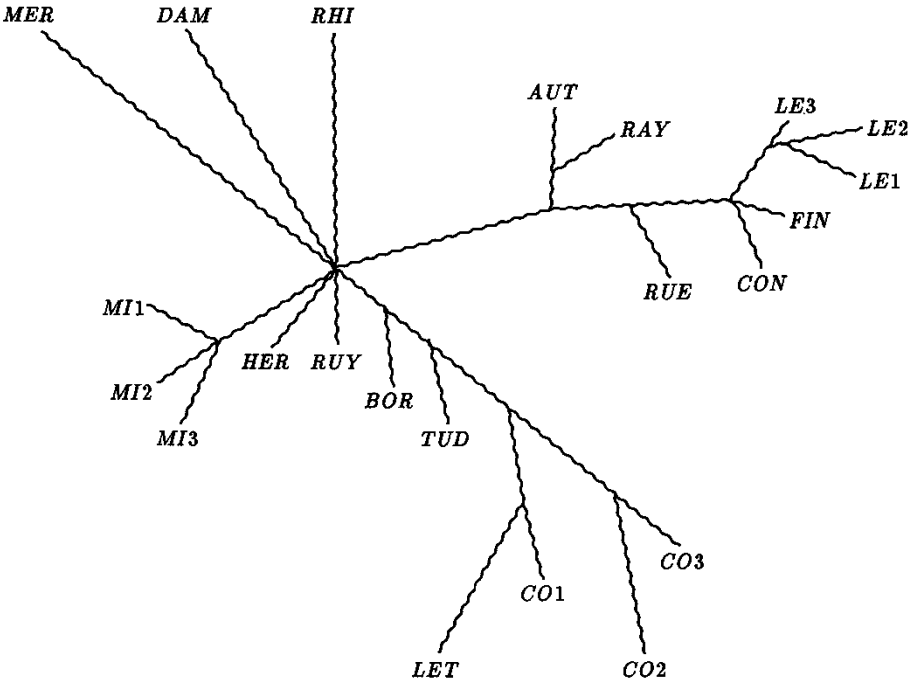

Figure 15
Analyse arborée
Méthode de X. Luong

L'objet en est de représenter des distances (ou des dissimilarités) sous la forme d'un arbre dont les feuilles (ou éléments terminaux) constituent les variables dont il faut rendre compte (ici les textes). Tout un réseau de noeuds structure cet arbre, depuis les ramifications périphériques jusqu'aux branches maîtresses, si bien que de tout texte on peut se rendre à un autre, en suivant un chemin plus ou moins direct (avec peu ou beaucoup de bifurcations) et en

[19] Pour de plus amples détails, nous renvoyons à deux articles du créateur de la méthode : "Représentations arborées de mesures de dissimilarités", *in Statistique et Analyse de données,* 1986, vol. 11, n° 1, pp. 20-41 et "Représentation arborée des données textuelles", in *Méthodes quantitatives et informatiques dans l'étude des textes,* 1986, Slatkine–Champion, vol. 2, pp. 575-586.

parcourant une distance plus ou moins longue. L'information donnée porte donc sur la structure, et, quand celle-ci est découverte, sur la mesure des arêtes ou distances. Le résultat reproduit à la figure 15 est d'une parfaite lisibilité : on sera sensible à la netteté des groupements, notamment à celui qui se constitue à droite et qui réunit les textes poétiques et à celui qui fédère les textes épistolaires (en bas). Dans les deux cas les sous-groupements s'imposent avec la même clarté : c'est la chronologie qui explique la relation privilégiée entre les *Lettres* et *Correspondance 1,* entre *Correspondance 2* et *Correspondance 3,* entre les *Feuilles* et les *Rayons.* L'unité des trois sous-ensembles de la *Légende des siècles* est nettement affirmée (c'est le triplet ultime du graphique), comme le rapport étroit qui lie la *Fin de Satan* aux *Contemplations* et que Hugo a explicitement reconnu. Reste à interpréter la partie médiane du graphe[20]. Un groupement y est clairement dessiné, qui manifeste la solidarité des trois parties distinguées (arbitrairement) dans les *Misérables.*

Par contre le graphe souligne l'autonomie des autres éléments romanesques qui ont un lien lâche avec la structure. C'est qu'une grande distance est établie entre le premier et le dernier roman de notre corpus, entre *Notre Dame* et les *Travailleurs de la mer,* qui divergent par le thème, l'écriture et la date de composition. Quant au *Rhin,* il est unique en son genre. Le théâtre n'est pas non plus solidement rattaché à la structure, surtout les pièces en vers dont le statut propre implique une division intérieure et qui convergent vers la place publique (ou point de ralliement général pour les indécis). Les pièces en prose, elles, se sont engagées sur le versant de la prose, à mi-chemin entre la correspondance et le roman. Mais là même où la structure est imprécise, les indications de distance gardent leur valeur. Quoique non direct, le chemin est très court qui conduit de *Marie Tudor à Lucrèce Borgia.* Et il en est ainsi des deux pièces en vers, très voisines sur le graphique, même si elles ne sont pas mitoyennes : *Hernani* et Ruy *Blas.* Remarquons enfin que ces deux textes versifiés, qui ne font pas partie du groupement poétique, sont les seuls à tenter de s'en rapprocher.

La connexion lexicale, ainsi mise en lumière, mérite donc mieux que l'accueil révérencieux mais silencieux que cette méthode a rencontré jusqu'ici. Il est peu d'instruments aussi puissants, capables de moissonner toute la matière lexicale, tâche presque aussi complexe que celle qui consisterait à mesurer la ressemblance de deux civilisations, l'opposition de deux systèmes de pensée, ou le degré précis de haine ou de tendresse entre deux individus. Il en est peu qui donnent naissance à des produits aussi raffinés et aussi compacts,

[20] Cette partie se trouve à gauche par simple commodité. A chaque noeud, l'angle d'incidence des branches n'a aucune signification, et à chaque bifurcation on peut orienter le graphe comme l'on veut, comme s'il s'agissait d'un pantin désarticulé. Les seules contraintes à respecter sont constituées par la longueur des arêtes et l'ordre de succession.

puisqu'un simple nombre finit par évaluer et résumer la distance globale de deux vocabulaires. Il reste cependant que cette mesure, une fois établie sur le contenu concret du discours, devient elle-même assez abstraite. Tout n'est pas dit quand on a relevé le dosage exact de la haine, ou établi le degré alcoolique d'un vin, ou mesuré la connexion lexicale de deux textes. Encore faut-il comprendre, goûter, lire. Mesurer ne suffit pas, il faut arpenter, c'est-à-dire parcourir.