Le traitement des faits linguistiques et stylistiques sur ordinateur. Texte d'application: Giraudoux

Étienne Brunet

Le traitement des faits linguistiques et stylistiques sur ordinateur. Texte d'application : Giraudoux[1]

Etienne Brunet

La recherche dont nous allons rendre compte est le produit croisé d'un matériau, d'une méthode et d'un outil. Le matériau a été fourni en majeure partie par le *Trésor de la Langue Française* : c'est le dictionnaire des fréquences de Giraudoux. La méthode a été empruntée pour l'essentiel aux pionniers et aux maîtres de la statistique linguistique, Herdan, Guiraud et Muller, et complétée par certaines démarches mathématiques familières aux sociologues et aux chercheurs des sciences humaines. L'outil enfin est constitué par le seul langage de programmation vraiment universel et réellement adéquat aux recherches linguistiques, le PL/1. Ce langage qui allie la souplesse à la puissance permet tout à la fois le traitement des chaînes de caractères, les calculs mathématiques les plus sophistiqués et notamment le calcul booléen, la constitution et la consultation de fichiers divers, la manipulation des adresses, la gestion de la mémoire dite « pointée », la réalisation de « piles », de chaînages et donc de grammaires formelles. A l'imitation des collègues des Facultés de Sciences qui sont très généralement les programmeurs de leurs propres recherches, nous avons donc appris à manier cet outil, devenant analyste, programmeur, opérateur, voire dépanneur. Au demeurant personne ne juge indigne d'apprendre à conduire une voiture et les chauffeurs sont devenus un luxe rare. L'informatique est un domaine en soi si coûteux que toute économie de personnel technique est

1. Article paru dans Jean David et Robert Martin (éds), *Statistique et Linguistique*, Paris, Klincksieck, 1974, p. 105-103. Il rend compte d'une thèse que l'auteur s'apprêtait à soutenir.

bienvenue. A procéder soi-même on gagne en outre du temps : il est plus expéditif de demander directement à la machine ce que l'on veut plutôt que de l'expliquer à un programmeur professionnel. De toute façon on y gagne en sécurité, en fiabilité : comment s'assurer des résultats si l'on ne sait pas contrôler leur élaboration. Le résultat dépend de l'algorithme utilisé, qu'il est imprudent de confier à un tiers.

Chacun étant le tiers d'autrui, nous pourrions éprouver quelque embarras au moment de proposer nos algorithmes, nos programmes et nos résultats[2], si nous ne songions qu'il s'agit là d'un exemple que nous soumettons à la critique et au contrôle, non d'un modèle dont l'application puisse être généralisée sans précaution.

1. Etablissement du corpus

1.1. Le corpus de Nancy

Nous avons puisé nos données à la mine de Nancy. Seize textes de Giraudoux avaient été retenus pour le *Trésor de la Langue Française* et mis sur bandes perforées. Et sur notre demande un dictionnaire des fréquences de Giraudoux avait été constitué – qui nous fut gracieusement communiqué sous forme de listing. Il a fallu perforer à nouveau les 13.000 entrées de ce dictionnaire, chaque carte précisant outre le lemme sa fréquence dans chacun des seize textes et dans l'ensemble. Mais on a profité de ce traitement pour résoudre les problèmes d'homographie laissés en suspens, corriger les fautes, éliminer les déchets (quelques noms propres notamment) et introduire des codes :

– codes grammaticaux (mots de relation 0, verbes 1, substantifs 2, adjectifs 3, adverbes 4, participes passés 6, participes présents 7),

– codes sémantiques dans le cas des substantifs (champ spatio-temporel 0, minéral 1, végétal 2, animal 3, concret 4, abstrait 5, noms de personne 6).

2. Nous avons publié une cinquantaine de programmes dans notre revue spécialisée *C.U.M.F.I.D.* (*Cahiers des Utilisateurs de Machines à des Fins d'Information et de Documentation*), U.E.R. Lettres et Sciences Humaines, Université de Nice.

Certes ces codes n'ont pas toujours un caractère très rigoureux. Par exemple la différence est parfois peu sensible entre le concret et l'abstrait; et l'arbitraire n'est pas absent de certains choix. Au moins est-ce un arbitraire impartial et constant, en ce sens que la même décision atteint tous les textes avec cohérence et qu'une compensation peut être espérée pour l'ensemble des cas douteux qui trouveraient un équilibre dans la loi des grands nombres. Dans le pire des cas, il suffit de s'abstenir de toute conclusion concernant les données qui auraient été trop mal distinguées, comme c'est le cas pour les mots « si » ou « en », dont l'ambiguïté n'empêche pourtant pas qu'ils ne puissent être traités au niveau général des mots de relation, qui vaut pour l'un et l'autre emplois.

De toute façon la plupart des options nous étaient imposées par l'état du corpus tel qu'il nous fut livré. Comme beaucoup de monographies ne manqueront pas d'être constituées à partir du *Trésor de la Langue Française*, il nous a semblé avantageux de suivre les décisions prises à Nancy. Car c'était là le moyen de maintenir une rigoureuse cohérence et de permettre la comparaison des auteurs entre eux et la confrontation de chacun à l'ensemble. Dans un domaine aussi coûteux, les fantaisies des francs-tireurs sont mal venues et la discipline vaut mieux que l'originalité.

1.2. Extension du corpus

Si nous nous sommes interdit de modifier le corpus, nous n'avons pas renoncé à le compléter. Cela ne s'imposait pas pour le théâtre de Giraudoux où douze titres avaient été retenus dans le corpus de Nancy *(Siegfried, Amphitryon 38, Judith, Intermezzo, La Guerre de Troie, Électre, Cantique des Cantiques, Ondine, l'Apollon de Bellac, Sodome et Gomorrhe, La Folle de Chaillot* et *Pour Lucrèce.)* Par contre l'oeuvre romanesque était fort mal représentée puisqu'elle ne comprenait que quatre textes : *Simon le Pathétique, Suzanne et le Pacifique, Siegfried et le Limousin* et *Bella,* et que leur extension chronologique était limitée. L'étude du roman giralducien supposait une assise élargie que nous avons constituée en ajoutant au corpus primitif sept textes nouveaux : *Provinciales, L'École des Indifférents, Juliette au Pays des hommes, Églantine,* les *Aventures de Jérome Bardini, Combat avec l'Ange* et *Choix des Élues.* A vrai dire les deux premiers et les deux derniers textes de cette série ne sont représentés que sur échantillons. Pour *Provinciales* et *l'École des Indifférents,* cette réduction s'imposait du fait

qu'il s'agit là de recueils de nouvelles, En ne retenant que l'une d'elles (respectivement *Sainte Estelle* et *Jacques l'Égoïste),* on conservait au récit une unité dont le recueil n'était pas pourvu au même degré. Par contre les deux derniers romans ont été réduits de manière aléatoire au cinquième de leur longueur (il a suffi d'extraire une page sur cinq). Ce faisant nous évitions d'avoir à comparer deux sous-ensembles – théâtre et roman – trop inégaux, et surtout en abaissant l'étendue de certains romans au-dessous de la moyenne du théâtre, on annulait l'influence de l'étendue relative – ce qui permettait l'étude des genres sans brouillage extérieur. Il s'agissait aussi d'expérimenter la méthode de l'échantillonnage, à côté de celle du dépouillement exhaustif. Enfin – faut-il l'avouer – des raisons d'économie nous ont inspiré, car les deux derniers romans sont longs et leur traitement intégral s'annonçait fort onéreux.

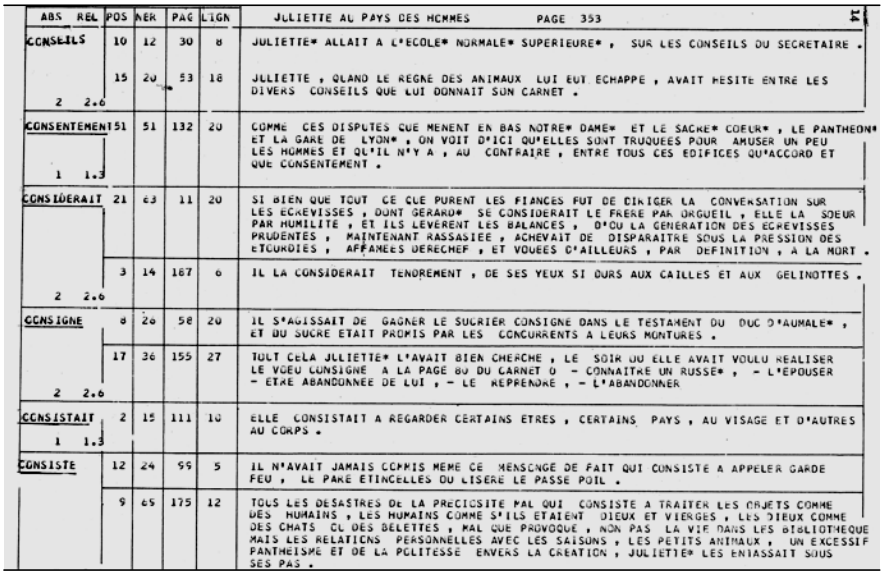
1.3. Les phases du traitement

Chacun des sept textes ajoutés au corpus a fait l'objet d'une préparation complexe dont les différentes phases peuvent être ainsi résumées:

– *perforation* du texte et vérification (30.000 cartes perforées pour l'ensemble).

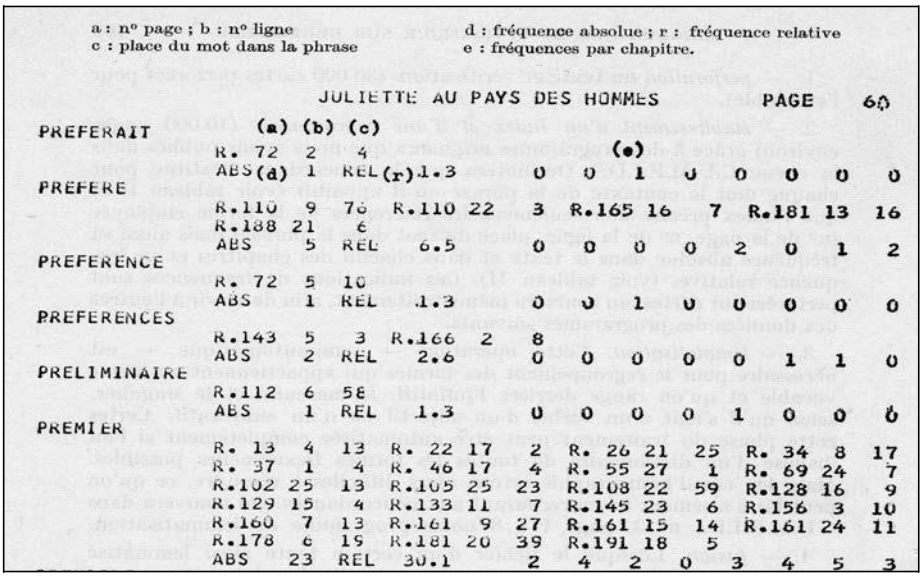
- *établissement d'un index et d'une concordance* (10.000 pages environ) grâce à des programmes originaux que nous avons publiés dans la revue *C.U.M.F.I.D*[3]. On notera que la concordance restitue pour chaque mot le contexte de la phrase où il apparaît (tableau 1) et que l'index précise non seulement les références de la forme employée (n° de la page, n° de la ligne, place du mot dans la phrase) mais aussi sa fréquence absolue dans le texte et dans chacun des chapitres et sa fréquence relative (tableau 2). Ces indications de fréquences sont perforées sur cartes, au cours du même traitement, afin de servir à l'entrée des données des programmes suivants.

3. *C.U.M.F.I.D.*, n° 1, 1970, p. 93 à 161.



ABS REL	POS	AER	PAG	LIGN	JULIETTE AU PAYS DES HOMMES PAGE 353
CONSEILS	10	12	30	8	JULIETTE* ALLAIT A L'ECOLE* NORMALE* SUPERIEURE* , SUR LES CONSEILS DU SECRETAIRE .
	15	20	53	18	JULIETTE , QUAND LE REGNE DES ANIMAUX LUI EUT ECHAPPE , AVAIT HESITE ENTRE LES DIVERS CONSEILS QUE LUI DONNAIT SON CARNET .
2 2.6					
CONSENTEMENT	51	51	132	20	COMME CES DISPUTES QUE MENENT EN BAS NOTRE* DAME* ET LE SACRE* COEUR* , LE PANTHEON* ET LA GARE DE LYON* , ON VOIT D'ICI QU'ELLES SONT TRUQUEES POUR AMUSER UN PEU LES HOMMES ET QU'IL N'Y A , AU CONTRAIRE , ENTRE TOUS CES EDIFICES QU'ACCORD ET QUE CONSENTEMENT .
1 1.3					
CONSIDERAIT	21	63	11	20	SI BIEN QUE TOUT CE QUE PURENT LES FIANCES FUT DE DIRIGER LA CONVERSATION SUR LES ECREVISSES , DONT GERARD* SE CONSIDERAIT LE FRERE PAR ORGUEIL , ELLE LA SOEUR PAR HUMILITE , ET ILS LEVERENT LES BALANCES , D'OU LA GENERATION DES ECREVISSES PRUDENTES , MAINTENANT RASSASIEE , ACHEVAIT DE DISPARAITRE SOUS LA PRESSION DES ETOURDIES , AFFAMEES DERECHEF , ET VOUEES D'AILLEURS , PAR DEFINITION , A LA MORT .
	3	14	167	6	IL LA CONSIDERAIT TENDREMENT , DE SES YEUX SI OURS AUX CAILLES ET AUX GELINOTTES .
2 2.6					
CONSIGNE	8	26	58	20	IL S'AGISSAIT DE GAGNER LE SUCRIER CONSIGNE DANS LE TESTAMENT DU DUC D'AUMALE* , ET DU SUCRE ETAIT PROMIS PAR LES CONCURRENTS A LEURS MONTURES .
	17	36	155	27	TOUT CELA JULIETTE* L'AVAIT BIEN CHERCHE , LE SOIR OU ELLE AVAIT VOULU REALISER LE VOEU CONSIGNE A LA PAGE 80 DU CARNET 0 - CONNAITRE UN RUSSE* , - L'EPOUSER - ETRE ABANDONNEE DE LUI , - LE REPRENDRE , - L'ABANDONNER
2 2.6					
CONSISTAIT	2	15	111	10	ELLE CONSISTAIT A REGARDER CERTAINS ETRES , CERTAINS PAYS , AU VISAGE ET D'AUTRES AU CORPS .
1 1.3					
CONSISTE	12	24	99	5	IL N'AVAIT JAMAIS COMMIS MEME CE MENSONGE DE FAIT QUI CONSISTE A APPELER GARDE FEU , LE PARE ETINCELLES DU LISERE LE PASSE POIL .
	9	65	175	12	TOUS LES DESASTRES DE LA PRECIOSITE MAL QUI CONSISTE A TRAITER LES OBJETS COMME DES HUMAINS , LES HUMAINS COMME S'ILS ETAIENT DIEUX ET VIERGES , LES DIEUX COMME DES CHATS OU DES BELETTES , MAL QUE PROVOQUE , NON PAS LA VIE DANS LES BIBLIOTHEQUE MAIS LES RELATIONS PERSONNELLES AVEC LES SAISONS , LES PETITS ANIMAUX , UN EXCESSIF PANTHEISME ET DE LA POLITESSE ENVERS LA CREATION , JULIETTE* LES ENTASSAIT SOUS SES PAS .

Tableau 1. Concordance



a : n° page ; b : n° ligne
c : place du mot dans la phrase
d : fréquence absolue ; r : fréquence relative
e : fréquences par chapitre.

JULIETTE AU PAYS DES HOMMES PAGE 60

PREFERAIT (a) (b) (c)
R. 72 2 4 (e)
ABS (d) 1 REL (r) 1.3 0 0 1 0 0 0 0 0

PREFERE
R.110 9 76 R.110 22 3 R.165 16 7 R.181 13 16
R.188 21 6
ABS 5 REL 6.5 0 0 0 0 2 0 1 2

PREFERENCE
R. 72 5 10
ABS 1 REL 1.3 0 0 1 0 0 0 0 0

PREFERENCES
R.143 5 3 R.166 2 8
ABS 2 REL 2.6 0 0 0 0 0 1 1 0

PRELIMINAIRE
R.112 6 58
ABS 1 REL 1.3 0 0 0 0 1 0 0 0

PREMIER
R. 13 4 13 R. 22 17 2 R. 26 21 25 R. 34 8 17
R. 37 1 4 R. 46 17 4 R. 55 27 1 R. 69 24 7
R.102 25 13 R.105 25 23 R.108 22 59 R.128 16 9
R.129 15 4 R.133 11 7 R.145 23 6 R.156 3 10
R.160 7 13 R.161 9 27 R.161 15 14 R.161 24 11
R.178 6 19 R.181 20 39 R.191 18 5
ABS 23 REL 30.1 2 4 2 0 3 4 5 3

Tableau 2. Index

– *lemmatisation.* Cette opération – semi-automatique – est nécessaire pour le regroupement des formes qui appartiennent au même vocable et qu'on range derrière l'infinitif, le masculin et le singulier, selon qu'il s'agit d'un verbe, d'un adjectif ou d'un substantif. Certes cette phase du traitement peut être automatisée complètement si l'on dispose d'un dictionnaire de toutes les formes flexionnelles possibles. Mais les cas d'homographie seront alors difficiles à résoudre, ce qu'on peut faire aisément ici en recourant aux concordances. On trouvera dans *C.U.M.F.I.D.*, n° 3, page 1 à 8 notre programme de lemmatisation.

– *fusion.* Lorsque le fichier d'un certain texte ainsi lemmatisé a été constitué sur disque (ou sur bande) et que l'ordre alphabétique a été contrôlé, on le compare au dictionnaire des textes déjà traités et fusionnés, afin de faire apparaître les mots nouveaux. Ceux-ci reçoivent leurs codes grammatical et sémantique (opération manuelle) et le fichier du texte est modifié en conséquence. Alors seulement, la fusion peut être assurée et les fréquences du texte considéré sont portées dans le dictionnaire de l'auteur (voir *C.U.M.F.I.D.*, n°3, programme 13). Au total chaque texte occasionne trois opérations manuelles (perforation, lemmatisation, codage) et 6 passages en ordinateur (index et concordance, création du fichier lemmatisé, tri, extraction des mots nouveaux, correction du fichier codé, fusion). En réalité un dictionnaire plus détaillé a été constitué qui tient compte des fréquences par chapitres, ce qui permet des analyses plus fines portant notamment sur la composition et ses conséquences lexicales. Quand toutes ces opérations fastidieuses ont été menées à bien – il en coûte quelques jours à la machine et cent fois plus au chercheur – alors commence la tâche noble de l'exploitation statistique.

2. La structure du vocabulaire

2.1. La richesse lexicale

Un programme de base[4] établit pour chaque texte et pour l'ensemble du corpus :

– le relevé des occurrences (N)

– celui des vocables ou mots différents (V) (tableau 3)

4. Programme n° 16, *C.U.M.F.I.D.*, n° 3, 1971, p. 27.

– la distribution des classes de fréquences, c'est-à-dire l'effectif des vocables employés 1 fois, 2 fois, *n* fois (tableau 4).

Occurrences *N*		Vocables *V*	Moyenne *N/V*
PRO	10899	2162	5.04
IND	10890	2181	4.99
51M	47389	4986	9.50
SUZ	63811	6136	10.39
S&L	63912	6590	9.69
JUL	36281	4857	7.46
BEL	49697	5820	8.53
EGL	50100	5385	9.30
SIE	21850	2957	7.38
AMP	25022	2910	8.59
BAR	49198	5293	9.29
JUD	24296	2974	8.16
INT	2242	3156	7.10
COM	14075	2545	5.53
GUE	20504	2613	7.84
ELE	28497	2967	9.60
CAN	8704	1507	5.77
ELU	16016	2599	6.16
OND	25360	2812	9.01
APO	7939	1434	5.53
SOU	22214	2571	8.64
FOL	14868	3312	1.50
LUC	27420	3025	9.06
TOT	671364	15771	42.56

Tableau 3. Relevé du vocabulaire (N, V et N/V)

		PRO	IND	SIM	SUZ	S&L	JUL	BEL	EGL	SIE	AMP	BAR	JUD	INT	COM	GUE	ELE	CAN	ELU	OND	APO	SOU	FOL	LUC	TOT
FREQ.	1	1329	1296	2357	2791	2949	2485	2828	2529	1576	1446	2910	1342	1668	1462	1360	1493	853	1440	1426	851	1311	1758	1574	4618
FREQ.	2	319	356	826	1062	1248	871	1019	899	457	468	879	511	551	434	423	448	247	455	473	214	441	586	481	2222
FREQ.	3	134	147	451	551	598	414	523	491	244	257	483	223	273	189	212	270	112	219	212	93	209	239	236	1356
FREQ.	4	87	88	240	358	361	221	327	310	148	147	292	129	157	102	122	138	66	108	128	64	114	145	139	971
FREQ.	5	51	54	194	225	276	158	207	192	96	81	184	97	84	76	61	88	38	55	77	33	82	96	92	675
FREQ.	6	28	32	121	174	179	131	147	152	65	57	136	97	66	44	58	73	27	46	60	16	55	79	66	541
FREQ.	7	30	23	97	130	135	84	101	106	49	53	108	47	34	26	40	63	15	44	46	17	52	50	52	457
FREQ.	8	25	19	85	97	92	50	85	84	24	43	94	38	25	28	32	34	17	21	37	11	28	36	60	337
FREQ.	9	12	14	69	79	75	49	70	71	27	34	62	25	20	12	30	28	14	17	35	14	21	27	32	328
FREQ.	10	19	16	59	63	52	42	40	58	13	25	45	23	20	17	24	32	7	14	24	11	17	29	29	277
FREQ.	11	10	12	33	47	49	36	43	45	19	15	44	21	22	5	17	19	10	11	25	6	14	17	10	220
FREQ.	12	8	9	37	46	61	28	26	31	20	26	28	20	15	11	10	17	7	16	22	8	12	19	15	197
FREQ.	13	9	7	33	34	41	17	29	34	14	21	38	10	11	4	17	12	4	10	14	4	15	13	13	163
FREQ.	14	6	7	18	33	36	16	33	35	8	13	38	14	15	12	9	7	3	10	20	4	10	17	12	177
FREQ.	15	3	5	27	30	18	18	18	20	15	11	14	13	10	7	11	9	2	7	10	4	10	9	7	138
FREQ.	16	6	6	22	29	22	13	27	14	10	8	21	10	10	6	11	12	5	10	6	6	9	11	10	140
FREQ.	17	3	4	14	25	21	9	16	21	7	10	12	6	8	9	6	11	1	7	15	4	7	4	3	128
FREQ.	18	0	3	15	15	21	12	14	13	7	11	17	4	12	5	6	5	4	6	6	1	9	7	8	99
FREQ.	19	5	5	12	8	20	4	17	23	7	8	12	7	5	5	5	8	2	5	6	3	9	5	12	104
FREQ.	20	4	3	12	14	17	9	8	12	6	7	10	11	6	6	6	11	1	6	6	3	4	9	7	106
FREQ.	21	4	3	14	10	12	9	16	10	7	8	9	8	6	4	6	8	2	3	5	4	2	4	7	85
FREQ.	22	3	0	6	11	17	9	5	4	3	8	16	6	4	5	6	5	2	3	8	5	7	11	6	94
FREQ.	23	1	4	18	17	12	13	3	9	7	9	10	5	4	5	4	11	5	1	5	4	3	9	3	73
FREQ.	24	3	1	6	3	9	4	7	11	6	5	7	4	8	5	4	7	1	1	7	1	7	3	9	87
FREQ.	25	4	1	10	17	11	1	10	5	5	3	6	7	3	0	6	3	1	1	6	1	2	4	6	73
FREQ.	26	4	3	9	2	13	6	8	7	3	3	11	6	2	0	1	3	2	2	7	0	2	4	4	76
FREQ.	27	0	3	7	7	11	2	5	6	3	3	5	6	2	0	3	5	5	3	3	3	2	7	1	56
FREQ.	28	2	1	4	8	11	2	8	8	4	5	5	4	2	1	3	4	3	3	1	1	5	0	2	48
FREQ.	29	0	4	4	5	6	3	2	7	1	5	9	6	2	5	5	4	2	2	2	2	4	4	4	56
FREQ.	30	2	1	4	10	7	3	6	5	4	5	8	3	1	3	0	6	2	2	3	1	3	2	4	54
FREQ.	31	1	2	3	7	0	1	5	4	1	1	7	4	2	1	1	3	2	0	7	1	1	3	4	49
FREQ.	32	0	3	2	9	7	3	1	4	2	3	2	3	1	2	1	3	1	3	2	1	1	2	2	38

Tableau 4. Distribution des fréquences

On peut alors étudier les rapports qui lient entre elles les différentes classes de fréquence. Le modèle peut être construit soit à partir de la formule de Waring-Herdan, soit d'après la loi binomiale, laquelle impose moins de contraintes. Nous avons réalisé les deux programmes[5] et nous sommes arrivé à la conclusion que la formule binomiale est plus intéressante et qu'elle souligne mieux l'exceptionnelle importance de la fréquence 1. On peut aussi établir la liste des vocables répartis par fréquences décroissantes et tester la loi de Zipf. Nous l'avons fait dans un programme[6] qui calcule les meilleurs facteurs de correction *a* et *b* de la formule développée :

$$(rang + b)^a \times fréquence = constante.$$

Et nous nous sommes convaincu que la pente *a* réelle n'est pas nécessairement le quotient *Log N/Log V*.

On peut surtout tenter de cerner la richesse lexicale des textes en présence, en mettant à l'épreuve plusieurs formules qui par des pondérations diverses mettent en rapport *N*, *V* et V_1 (V_1 représentant l'effectif des mots employés 1 fois). On a ainsi avancé les équations :

$$V / \sqrt{N} = constante$$
$$V_1^3 / V^2 = constante$$
$$V_b = \sqrt[2]{p}\, V_a$$
$$V_{1b} = \sqrt[3]{p}\, V_{1a}$$

où V_a et V_b désignent respectivement le vocabulaire des textes *a* et *b*, *p* étant le rapport du second au premier (en occurrences). Les programmes[5] réalisés pour ces différentes équations ont donné des résultats quelque peu décevants, l'influence de la longueur des textes n'étant jamais complètement éliminée. L'indice de Yule-Herdan a donné lieu à un autre programme, avec des résultats assez peu satisfaisants, le poids des fréquences élevées s'y montrant prépondérant.

La simple règle de trois se révélant insuffisante comme aussi les systèmes de pondération même sophistiqués, force est de recourir aux lois et formules de la statistique classique, et notamment à la loi binomiale. Pour l'exposé de cette formule nous renvoyons le lecteur à Charles Muller,

5. *C.U.M.F.I.D.*, n° 3, 1971, p. 55 et n° 4, p. 209.
6. *C.U.M.F.I.D.*, n° 7, 1973, p. 61.

Initiation à la statistique linguistique (Paris, Hachette, 1973, p. 38 et sqq)[7]. Nous nous bornerons ici à expliquer que la formule

$$q\,V_1 + q^2V_2 + q^3V_3 + \ldots q^kV_k$$

permet de calculer le vocabulaire théorique d'un texte ; ce dernier est comparé à l'effectif réellement observé, et l'écart est apprécié grâce au test de X^2 – ce qui autorise finalement un classement où l'on retrouve en tête tous les romans, en queue les tragédies ou pièces antiques et entre les deux les comédies ou pièces modernes (tableau 5).

Oeuvre	Vocab.	Théorique	Ecart	X2	Complém	X2 tot	Rang
PRO	2162	2367.95	-205.95	17.91	3.16	21.07	7
INO	2181	2366.63	-185.63	14.56	2.57	17.13	5
SIM	4986	5499.51	-513.51	47.94	25.67	73.61	44
SUZ	6136	6373.06	-237.06	8.81	5.97	14.78	4
S8I.	6590	6378.00	212.00	7.04	4.78	11.82	1
JUL	4857	4185.09	71.91	1.08	0.47	1.55	3
8EL	5820	5633.55	186.45	6.17	3.42	9.59	2
EGL	5385	5656.54	-271.54	13.03	7.28	20.31	6
SIE	2957	3610.27	-653.27	118.20	35.09	153.29	15
AMP	2910	3901.32	-991.32	251.89	82.79	334.68	19
8AR	5293	5604.94	-311.94	17.36	9.57	26.93	9
JUU	2974	3836.70	-862.70	193.98	62.36	256.34	17
INT	3156	3664.35	-508.35	70.52	21.34	91.86	12
CON	2545	2777.53	-232.53	19.46	4.16	23.62	8
SUE	2613	3479.49	-866.49	215.78	61.08	276.86	18
ELE	2967	4196.26	-1229.26	360.10	130.54	490.64	23
CAN	1507	2049.55	-542.55	143.62	21.45	165.07	16
ELU	2599	3004.86	-405.86	54.81	12.90	6T.71	10
OND	2812	3931.00	-1119.00	318.53	105.75	424.28	22
APO	1.434	1929.82	-495.82	127.38	17.76	145.14	14
SOD	2571	3644.92	-1073.92	316.41	95.10	411.51	21
FOL	3312	3887.64	-575.64	85.23	27.88	113.11	13
LUC	3025	4107.25	-1082.25.	285.17	100.41	385.58	20

X2=2694.98

Tableau 5. Richesse lexicale

7. On pourra consulter également l'ouvrage fondamental de G. Herdan, *The Advanced theory Language as choice and chance*, Springer, 1966 et le n° 3 de nos *C.U.M.F.I.D.* Par d'autres voies, Pierre Guiraud aborde aussi cette question dans *Problèmes et Méthodes de la Statistique lexicale*, PUF, 1960, p. 84 et sqq.

Il est improbable que le hasard puisse expliquer une distribution qui reproduit si fidèlement la distinction des genres. La conclusion s'impose : quand un texte appartient au théâtre et que le sujet en est antique et le ton tragique, le vocabulaire de Giraudoux perd en étendue et en variété – le cas limite étant atteint par *Electre* dont le vocabulaire est comme enfermé dans le cercle étroit des hantises de l'héroïne. A l'extrême inverse, c'est l'abondance luxuriante des thèmes et des termes qui vaut à *Siegfried et le Limousin* la première place. Entre ces deux extrêmes, les textes se répartissent selon les dosages divers des trois critères pertinents (genre, sujet et ton) tout en suivant une évolution chronologique où Giraudoux manifeste une tendance à la sobriété lexicale. Les derniers romans sont ainsi moins riches que les premiers, et les dernières pièces tragiques plus dépouillées que les premières. La corrélation n'est pas douteuse entre l'ordre chronologique et le classement selon la richesse lexicale ; le coefficient de Spearman atteint une valeur très élevée (R = 0,71) que le hasard ne peut produire qu'une fois sur mille. Nous retrouverons d'ailleurs tout au long de notre étude cette loi du temps.

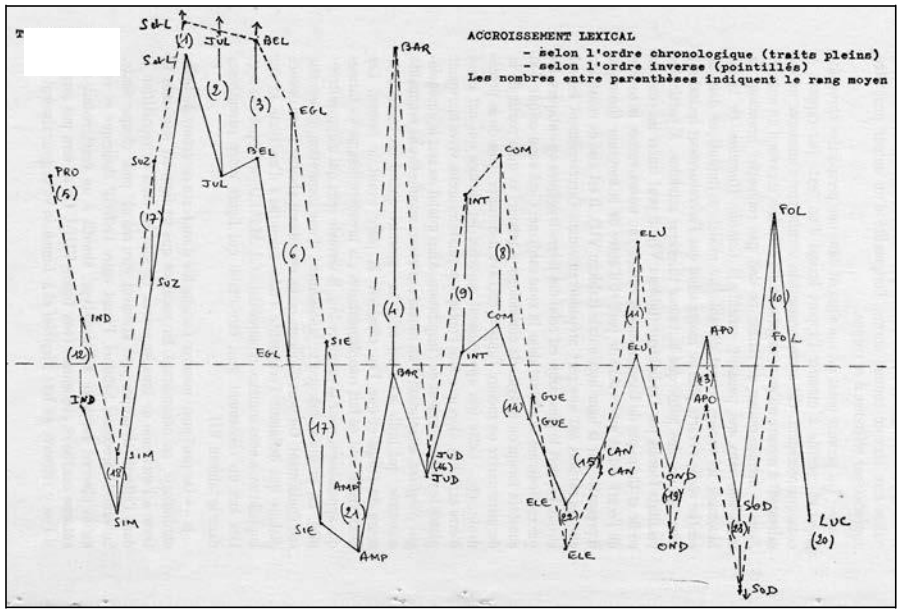
2.2. L'originalité lexicale

Alors que la richesse lexicale peut être mesurée dans l'absolu, l'originalité du vocabulaire est une notion relative. C'est par rapport à l'oeuvre antérieure (ou postérieure) ou par rapport à l'ensemble de l'oeuvre qu'un texte apparaît original. Ou si l'on dispose d'un corpus plus vaste qui puisse être accepté comme norme, l'originalité d'un auteur peut être appréciée par référence à ce corpus.

a – Un premier point de vue s'inscrit dans une perspective chronologique et consiste à mesurer l'apport lexical d'un texte par rapport à ceux qui précèdent. Cet accroissement du vocabulaire s'amenuise nécessairement à mesure qu'on se rapproche des dernières oeuvres d'un auteur. Ainsi le dernier roman de Balzac ne doit pas comporter beaucoup de mots nouveaux qui puissent enrichir la *Comédie Humaine*. Or la loi binomiale ici encore permet d'établir un modèle asymptotique dont la courbe réelle s'écarte plus ou moins selon que l'accroissement lexical est plus fort ou plus faible que la valeur théorique attendue. À partir des résultats fournis par l'ordinateur (tableau 6), on peut tracer les courbes réelle et théorique de l'accroissement cumulé, ou mieux encore la courbe de l'écart réduit qui évolue de part et d'autre de la moyenne théorique représentée par l'horizontale (graphique 1).

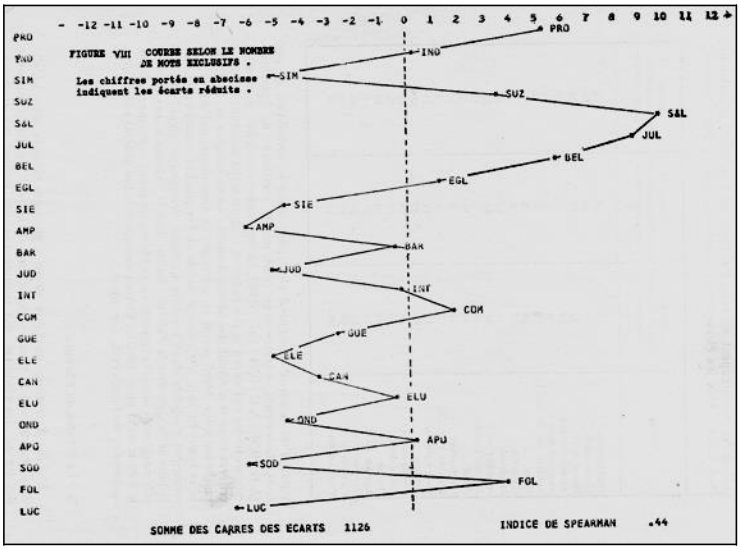
	Théo-manqu.	*Théo-cum*	*Réel-cum*	*Accroiss-théo*	*Accrois-réel*	*Ecart*	*Probab.*	*Type*	*Réduit*	*X2*
IND	12187.267	3583.733	3411	1304.693	1249	-55.693	.083	34.593	-1.610	2.377
SIM	9147.254	6623.746	6192	3040.013	2781	-259.013	.193	49.537	-5.229	22.068
SUZ	6899.172	8871.828	8516	2248.082	2384	135.918	.143	43.904	3.096	8.218
S&L	5375.995	10395.005	10505	1523.176	1929	405.824	.097	37.095	10.940	108.12,
JUL	4683.807	11087.193	11370	692.188	865	172.812	.044	25.723	6.718	43.144
BEL	3869.571	11901.429	12384	814.237	1014	199.763	.052	27.786	7.189	49.010
EGL	3164.748	12606.252	13095	704.822	711	6.178	.045	25.948	.238	.054
SIE	2886.185	12884.815	13281	278.564	186	-92.564	.018	16.541	-5.596	30.758
AMP	2585.231	13185.769	13468	300.954	187	-113.954	.019	11.180	-6.633	43.148
BAR	2041.960	13729.040	14004	543.271	536	-7.271	.034	22.901	-.318	.097
JUD	1794.309	13976.691	14191	247.651	187	-60.651	.016	15.611	-3.885	14.854
INT	1576.242	14194.758	14416	218.068	225	6.933	.014	14.661	.473	.220
COM	1444.114	14326.886	14564	132.127	148	15.873	.008	11.441	1.387	1.907
GUE	1257.840	14513.160	14720	186.274	156	-30.274	.012	13.567	-2.232	4.920
ELE	1010.220	14760.780	14892	247.621	172	-75.621	.016	15.611	-4.844	23.094
CAN	936.983	14834.017	14940	73.236	48	-25.236	.005	8.535	-2.957	8.696
ELU	805.119	14965.881	15076	131.865	136	4.135	.008	11.434	.362	.130
OND	603.287	15167.713	15225	201.831	149	-52.831	.013	14.111	-3.744	13.829
APO	541.826	15229.174	15295	61.461	70	8.539	.004	7.817	1.092	1.186
SOD	373.802	15397.198	15400	168.024	105	-63.024	.011	12.891	-4.889	23.640
FOL	192.400	15578.600	15654	181.402	254	72.598	.012	13.390	5.422	29.054
LUC	.000	15771.000	15771	192.400	117	-75.400	.012	13.781	-5.471	29.549

Tableau 6. Accroissement du vocabulaire



Graphique 1. Accroissement lexical

Il est aisé de constater que les romans ont une part prépondérante dans l'accroissement lexical et que l'apport le plus faible est celui des pièces tragiques ou antiques, les pièces modernes se situant dans la zone médiane. C'est aussi ce que l'on constate lorsqu'on imagine la chronologie à rebours et qu'à partir de la dernière oeuvre on mesure l'apport de l'avant-dernière, puis de la précédente, etc. Cette fois apparaissent en relief les thèmes qui ont cessé d'être exploités, alors que dans l'optique normale les arêtes vives marquent plutôt le renouvellement de l'inspiration. Ainsi quand on suit la chronologie, *Siegfried*, adaptation d'un roman antérieur, manifeste une originalité émoussée, qui brille d'un vif éclat au contraire quand on remonte le cours du temps et que l'on considère la pièce avant le roman. Les deux visées sont en fait complémentaires : la première désigne les thèmes qui apparaissent pour la première fois, la seconde ceux qui n'apparaîtront plus. La première marque le jaillissement d'une inspiration, la seconde son épuisement. On peut d'ailleurs en les combinant aboutir à une donnée globale qui définirait l'originalité « tous azimuts » d'un texte et sur laquelle nous nous sommes expliqué dans *C.U.M.F.I.D.*, n° 3, p. 97 et sqq. Dès lors un classement peut être établi qui figure entre parenthèses dans le graphique 1.



Graphique 2. Courbe selon le nombre de mots exclusifs

b – On peut aussi mesurer l'originalité d'un texte en dehors de toute chronologie, en dénombrant les vocables qui ne figurent que dans ce texte à l'exclusion de tous les autres. Ce sont les mots de répartition 1, dont l'effectif théorique peut aisément être calculé pour chaque texte. Suivant le processus habituel, l'écart entre l'effectif théorique et celui qu'on observe, rapporté à l'écart-type, aboutit à un écart réduit qui autorise courbe et classement (graphique 2)[8].

On ne sera pas surpris si l'on y retrouve en tête *Siegfried et le Limousin* et si généralement les mêmes oeuvres occupent un rang identique ou très voisin de celui qui était le leur dans les classements précédents. Voir tableau ci-dessous :

Texte	*Rang selon la richesse lexicale* *A*	*Rang selon l'accroissement du vocabulaire* *B*	*Rang selon le nombre de mots exclusifs* *C*
Provinciales	9	5	4
École des I.	6	12	10
Simon le P.	11	18	18
Suzanne et le P.	4	7	6
Sieg. et le L.	1	1	1
Jul. au P. des H.	2	2	2
Bella	3	3	3
Églantine	5	6	8
Siegfried	14	17	17
Amphitryon	19	21	22
Jérôme Bardini	8	4	12
Judith	17	16	19
Intermezzo	12	9	11
Comb. avec l'A.	7	8	7
Guerre de T.	18	14	14
Électre	23	22	20
Cantique des C.	16	15	15
Choix des E.	10	11	13
Ondine	22	19	16
Apollon de B.	15	13	9
Sodome et G.	21	23	21
Folle de C.	13	10	5
Pour Lucrèce	20	20	23

8. A partir d'une distribution quelconque notre programme n° 35, *C.U.M.F.I.D.*, n° 6, 1973, trace la courbe correspondante, d'après les valeurs de l'écart réduit, et en fait apparaître les caractéristiques (X2 global et coefficient de corrélation chronologique).

Point n'est besoin d'être grand clerc pour saisir la corrélation de ces trois classements, attestée avec éclat par le coefficient de Spearman (A-B : 0,91 ; A-C : 0,84 ; B-C : 0,92). La conclusion est claire : ce sont les mêmes textes qui sont à la fois riches et originaux. Et pourtant les deux critères ne se recouvrent pas nécessairement : qu'on imagine un traité de mathématiques inséré parmi les ouvrages de Giraudoux, nul doute que la richesse lexicale n'y paraisse faible et l'originalité considérable. On remarque d'ailleurs dans les faits quelques divergences significatives : la plus nette concerne la *Folle de Chaillot* dont la richesse reste moyenne (13ème rang) mais dont l'originalité est éclatante (5ème rang) – ce que confirme la lecture de la pièce dont les thèmes sociaux et le ton général ont de quoi surprendre les familiers de Giraudoux.

2.3. Le rythme du discours

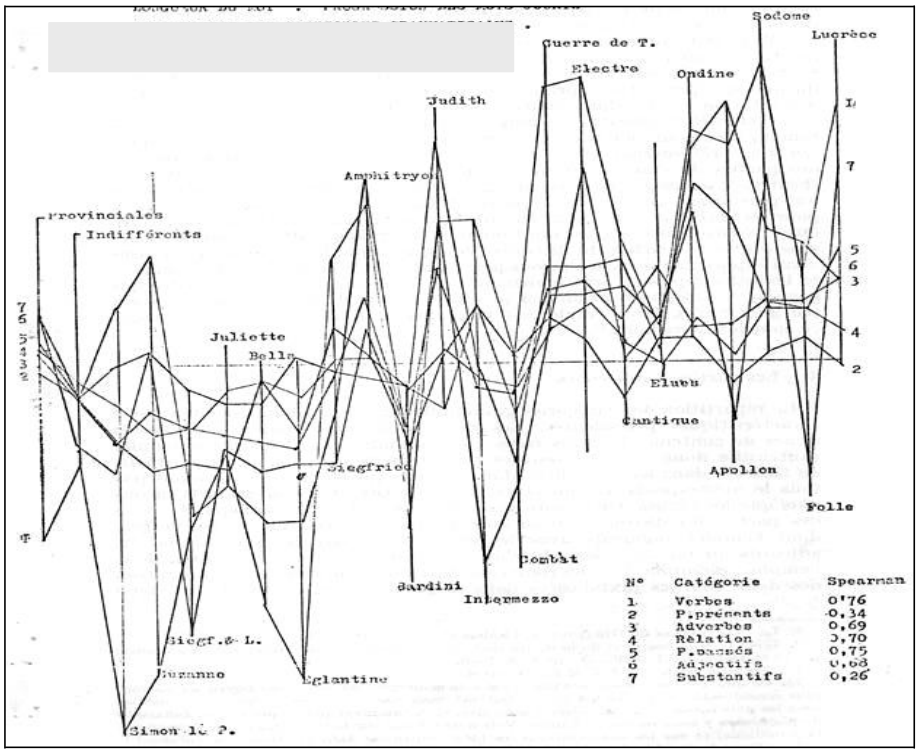
L'étude du rythme peut être située à plusieurs niveaux selon qu'elle distingue des unités plus ou moins larges. On peut étudier la longueur des textes, des chapitres, des paragraphes, des tirades, des répliques, des phrases, des mots... Sans reprendre le détail d'une publication antérieure[9], nous nous bornerons à constater le raccourcissement progressif de la phrase de Giraudoux et la rupture qui sépare en ce domaine le roman et le théâtre, et dans un même genre les pièces roses et les pièces noires, et qui oppose le rythme dur, appuyé, où le point domine, au rythme souple, enveloppé, où la virgule abonde. Quant à la longueur du mot qui peut être exprimée en nombre de syllabes ou mieux (car la délimitation en est plus rigoureuse) en nombre de lettres[10], elle s'accorde avec celle de la phrase. Le mot est plus court au théâtre que dans le roman, plus court aussi dans les pièces tragiques que dans les comédies. Certains présumeront que c'est la conséquence d'une répartition inégale des parties du discours et que les mots de relation, plus abondants au théâtre, y abaissent la moyenne générale, étant donné qu'ils sont deux fois plus courts que les mots « pleins »[11] (2,845 lettres par mot contre 5,776 pour les verbes et 5,985 pour les substantifs). L'explication doit

9. La Ponctuation de Giraudoux, *C.U.M.F.I.D.*, n° 1, 1970.

10. L'étude de la longueur du mot, par texte et par catégorie, fait l'objet des programmes 45, 46 et 48 dans *C.U.M.F.I.D.*, n°8, 1974, p. 3-46.

11. Programme n° 49, *C.U.M.F.I.D.*, n°8, 1974.

en réalité être cherchée ailleurs, dans une tendance générale qui affecte dans le même sens toutes les parties du discours (voir graphique 3) et qu'on peut aisément rapprocher des conclusions précédentes ; car la richesse et l'originalité lexicales reposent principalement sur les éléments du vocabulaire qui gagnent en compréhension (en précision) ce qu'ils perdent en extension (en généralité). Leur emploi est plus rare et l'usage ne les a pas usés : ce sont les mots longs.[12]

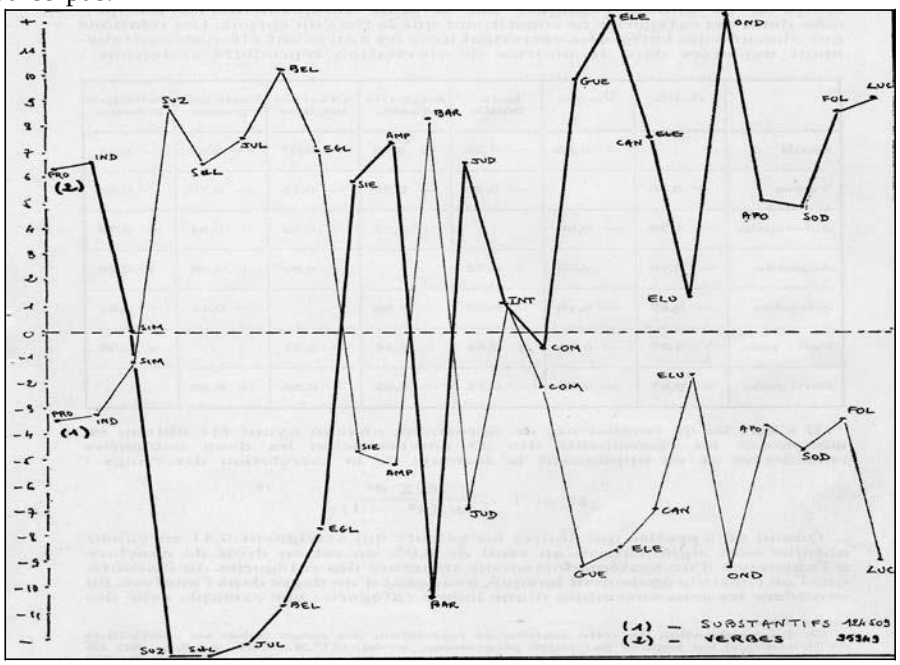


Graphique 3. Progression des mots courts dans les catégories grammaticales

12. Ce sont aussi les moins économiques, non seulement parce qu'ils exigent un nombre plus considérable de phonèmes (ou de lettres), mais parce que ces phonèmes eux-mêmes sont les plus rares, les moins aisés à articuler, et notamment parce que les combinaisons de phonèmes y sont moins courantes. Voir notre étude sur la répartition des graphèmes (2,5 millions) et sur les combinaisons les plus fréquentes dans la classe des substantifs (1,5 million de combinaisons), *C.U.M.F.I.D.*, n° 8, 1974.

2.4. Les parties du discours

La répartition des catégories grammaticales n'est pas indépendante des caractéristiques précédentes. Les mots-outils sont évidemment moins riches de contenu et moins rares que les mots dits pleins. Leur emploi s'intensifie donc dans les oeuvres relativement pauvres, principalement au théâtre, dans les tragédies. Mais les mots pleins eux-mêmes n'ont pas tous le même poids de substance, les adjectifs et les substantifs en ont plus que les verbes. On assiste ainsi chez Giraudoux à une bipolarisation des parties du discours : d'un côté les mots de relation et les verbes, dont l'emploi augmente avec le temps, de l'autre les substantifs, les adjectifs qualificatifs, les adverbes (de manière), et les participes, dont l'emploi prédomine dans les romans et baisse au fil des ans. La comparaison des deux courbes juxtaposées dans le graphique 4 manifeste de façon éclatante l'opposition qui s'affirme entre les verbes et les substantifs et qu'on ne saurait expliquer par la seule complémentarité puisqu'à elles deux ces catégories ne constituent que le tiers du corpus.



Graphique 4. Courbes des verbes et des substantifs

Les relations que chacune des catégories entretient avec les autres ont été systématiquement explorées dans la matrice de corrélation reproduite ci-dessous :

	Outils	Verbes	Substantifs	Adjectifs	Adv-manière	Part-passés	Part-présents
Outils		+ 0,55	- 0,76	- 0,76	- 0,57	- 0,27	- 0,57
Verbes	+ 0,55		- 0,90	- 0,75	- 0,15	- 0,70	- 0,65
Substantifs	- 0,76	- 0,90		+ 0,72	+ 0,32	+ 0,54	+ 0,72
Adjectifs	- 0,76	- 0,75	+ 0,72		+ 0,39	+ 0,36	+ 0,52
Adverbes	- 0,57	- 0,15	+ 0,32	+ 0,39		- 0,11	+ 0,28
Part. pass.	- 0,27	- 0,70	+ 0,54	+ 0,36	- 0,11		+ 0,36
Part. prés.	- 0,57	- 0,65	+ 0,72	+ 0,52	+ 0,28	+ 0,36	

Il s'agit ici de coefficients de Spearman, chacun ayant été obtenu en comparant les classements des 23 oeuvres selon les deux catégories considérées et en appliquant la formule de la corrélation des rangs [13] :

$$R = 1 - \frac{6 \sum d^2}{n (n^2 - 1)}$$

	PREPO	ARTIC	INDEP	RELAT	CONJO	PERSO	POSSE	DEMON	NUMER	TEMPS	LIEU	QUANT	NEGAT	INTER	AMBIG	AVOIR	ETRE	RELAT	
PRE	1.00	.95	.21	.30	.40	-1.00	-.64	-.69	.52	.09	.42	-.08	-.92	-.95	-.27	-.66	-.91	-.87	PREPOSITIONS
ART	.95	1.00	.32	.28	.40	-.98	-.78	-.67	.56	.04	.33	-.05	-.88	-.89	-.00	-.69	-.87	-.83	ARTICLES
IND	.21	.32	1.00	.41	.14	-.24	-.29	-.30	.36	.68	.35	.13	-.36	-.38	.17	-.42	-.38	-.20	INDEFINIS
REL	.30	.28	.41	1.00	.39	-.36	-.19	-.29	.40	.34	.20	-.15	-.46	-.35	.20	-.31	-.29	-.23	RELATIFS
CON	.40	.40	.14	.39	1.00	-.46	-.21	-.38	.28	-.13	.08	-.52	-.48	-.29	.25	-.40	-.33	-.23	CONJONCTIONS
PER	-1.00	-.98	-.24	-.36	-.46	1.00	.73	.62	-.91	-.03	-.43	.10	.90	.90	.10	.67	.88	.86	PERSONNELS
POS	-.64	-.78	-.29	-.19	-.21	.73	1.00	.22	-.55	.16	-.53	-.10	.51	.52	-.12	.40	.52	.51	POSSESSIFS
DEM	-.69	-.67	-.30	-.29	-.38	.62	.22	1.00	-.68	-.34	-.06	.15	.82	.78	.03	.82	.87	.82	DEMONSTRATIFS
NUM	.52	.56	.36	.40	.28	-.91	-.55	-.68	1.00	.12	.19	-.42	-.76	-.57	.21	-.56	-.65	-.62	NUMERAUX
TEM	.09	.04	.68	.34	-.13	-.03	.16	-.34	.12	1.00	.26	.26	-.20	-.31	-.13	-.33	-.34	-.19	ADVERBES TEMPS
LIE	.42	.33	.35	.20	.08	-.43	-.53	-.06	.19	.26	1.00	.09	-.35	-.33	-.27	-.31	-.41	-.37	ADVERBES LIEU
QUA	-.08	-.05	.13	-.15	-.52	.10	-.10	.15	-.42	.26	.09	1.00	.28	.08	-.33	-.01	.10	.09	ADVERBES QUANT
NEG	-.92	-.88	-.36	-.46	-.48	.90	.51	.82	-.76	-.20	-.35	.28	1.00	.90	.03	.83	.96	.91	NEGATIONS
INT	-.95	-.89	-.38	-.35	-.29	.90	.52	.78	-.57	-.31	-.33	.08	.90	1.00	.28	.63	.92	.85	INTERJECTIONS
AMB	-.27	-.00	.17	.20	.25	.10	-.12	.03	.21	-.13	-.27	-.33	.03	.28	1.00	-.14	.11	.10	AMBIGUS
AVO	-.66	-.69	-.42	-.31	-.40	.67	.40	.82	-.56	-.33	-.31	-.01	.83	.63	-.14	1.00	.87	.87	AVOIR
ETR	-.91	-.87	-.38	-.29	-.33	.88	.52	.87	-.65	-.34	-.41	.10	.96	.92	.11	.87	1.00	.98	ETRE
REL	-.87	-.83	-.20	-.23	-.23	.86	.51	.82	-.62	-.19	-.37	.09	.91	.85	.10	.87	.98	1.00	TOTAL

Tableau 7. Corrélation Bravais-Pearson des mots de relation

13. La constitution de cette matrice de corrélation des rangs (selon les coefficients de Spearman) est assurée par notre programme, n° 32, *C.U.M.F.I.D.*, n° 6, 1973, p. 81. Le programme n° 33, *C.U.M.F.I.D.*, p. 97, réalise une autre matrice de corrélation, fondée sur les coefficients de Bravais-Pearson. Enfin le programme n° 34, p. 121, reproduit la corrélation graphique.

Quand on a précisé que toutes les valeurs qui atteignent 0,41 en valeur absolue sont significatives au seuil de 0,05, on est en droit de conclure à l'existence d'un système fortement structuré des catégories du discours, que l'on constate également lorsque, avançant d'un degré dans l'analyse, on considère les sous-ensembles d'une même catégorie, par exemple celle des mots de relation (tableau 7).

On y remarquera l'opposition des déterminants faibles (articles, indéfinis, numéraux), plus fréquents dans les romans, et des déterminants forts (démonstratifs, possessifs), caractéristiques du théâtre, la prédominance dans les romans des catégories descriptives (adverbes de lieu, de temps) et de celles sur lesquelles s'articule l'enchaînement des mots ou des propositions (prépositions, relatifs, conjonctions), et l'abondance au contraire dans l'oeuvre théâtrale des catégories «affectives» (personnels, négations, interjections). Nous avons même poussé l'investigation plus loin, étudiant certains sous-ensembles dans le détail des composants et constatant par exemple que la distribution du singulier différait de celle du pluriel, le féminin du masculin, et la première personne de la seconde et de la troisième. A la limite cependant l'analyse finit par rencontrer le mot individuel et dès lors l'étude de la structure bascule dans celle du contenu.

3. Le contenu lexical

3.1. Le vocabulaire significatif

Alors que dans la perspective précédente on ignorait l'identité et la charge sémantique des mots, il s'agit ici de faire apparaître la coloration thématique d'un texte ou d'un ensemble de textes, à travers les mots-clés qui reviennent avec le plus d'insistance et qui constituent le vocabulaire significatif de l'oeuvre considérée. Là encore pour chaque mot le programme[14] calcule la fréquence qu'on devrait théoriquement rencontrer dans chaque texte, établit la comparaison avec la fréquence observée et apprécie la différence par un écart réduit. Plus cet écart est grand –

14. Publié dans *C:U.M.F.I.D.*, n° 4, 1972, p. 163.

notamment s'il dépasse 2 en valeur absolue – plus le mot en question est représentatif du texte étudié. On peut ainsi faire apparaître le squelette thématique d'une oeuvre. L'exemple de *Pour Lucrèce* est à ce titre très caractéristique : dans la liste des mots significatifs fournie par l'ordinateur (tableau 8), si l'on écarte les mots-outils et ceux de fréquence basse, on reconnaît aisément le sujet de la pièce ; la série : *mari, femme, pureté, séducteur, homme, vertu, vice, amant, vengeance, coupable, crime*, indique assez ce dont il est question : *viol*, adultère et *pureté* perdue.

JEROME BARDINI — VOCABULAIRE CARACTERISTIQUE POSITIF

MUTISME	9	5	.65	5.60	CATARACTE	6	4	.43	5.65	STERILISE PP	6	4	.43	5.65
CUIR	24	9	1.75	5.69	PROMENADE	97	22	7.10	5.80	PRINTEMPS	95	22	6.96	5.92
TOUSSER	27	10	1.97	5.94	INCONNU	211	38	15.46	5.95	MUSIQUE	69	18	5.05	5.98
JOUER	179	34	13.11	5.99	ASSISTE P	8	5	.58	6.03	FERMIER	8	5	.58	6.03
QUATUOR	8	5	.58	6.03	CONTINUEL	8	5	.58	6.03	LUI	2694	282	197.41	6.25
COMBLE PP	21	9	1.53	6.27	ORPHELIN	25	10	1.83	6.27	COMMUNAUTE	5	4	.36	6.30
CONTREBANDIER	5	4	.36	6.30	IRONIQUE	5	4	.36	6.30	FAMILIAL	17	8	1.24	6.30
CETTE	2396	256	175.57	6.30	DESPOTE	3	3	.21	6.33	IMMERGE PP	3	3	.21	6.33
LUTHERIE	3	3	.21	6.33	OURAGAN	3	3	.21	6.33	SAVOURANT	3	3	.21	6.33
SEMBLER	455	69	33.34	6.41	GESTE	267	47	19.56	6.44	COMPAGNON	47	15	3.44	6.47
ASSURE PP	37	13	2.71	6.49	VIOLON	12	7	.87	6.82	L'	11888	1065	871.15	6.82
SON	3235	341	237.06	7.01	FIANCE PP	120	29	8.79	7.08	DISPARITION	21	10	1.53	7.11
HAUTBOIS	4	4	.29	7.16	INSIGNE	4	4	.29	7.16	TRIO	11	7	.80	7.20
LAC	81	23	5.93	7.28	SA	2616	290	191.70	7.37	VERANDA	8	6	.58	7.39
FUI	31	13	2.27	7.39	LIBERTE	101	27	7.40	7.48	AH	34	14	2.49	7.57
DE	34147	28762	502.29	7.76	FUITE	25	12	1.83	7.81	SACRE PP	15	9	1.09	7.87
EVADE PP	17	10	1.24	8.17	NEIGE	129	34	9.45	8.29	DEVINE PP	50	19	3.66	8.33
CLUB	16	10	1.17	8.48	CET	339	91	39.49	8.51	AUCUN	322	64	23.59	8.64
ALTO	6	6	.43	8.82	EVASION	10	8	.73	8.84	ELLE	6199	639	454.26	9.00
S'	2998	350	219.69	9.13	VAGABOND	19	12	1.39	9.34	BASSON	9	8	.65	9.47
BESACE	8	8	.58	10.12	POLICEMAN	12	10	.87	10.16	IL	9819	993	719.53	10.59
HUI	65	29	4.76	11.54	LYNX	40	32	2.93	17.64	ENFANT	730	238	53.49	26.20

POUR LUCRECE

	FREQUENCE TOTALE	FREQ. REELLE	FREQ. THEORIQ.	ECART REDUIT										
PISTOLET	7	4	.28	7.18	VIOL	7	4	.28	7.18	CRIME	106	19	4.32	7.21
DUEL	19	7	.77	7.25	ASSOMMANT	4	3	.16	7.26	CYNISME	4	3	.16	7.26
DEGRAFE PP	4	3	.16	7.26	MUQUEUX	4	3	.16	7.26	CELLES	114	20	4.65	7.26
COUPABLE	53	13	2.16	7.93	VENGEANCE	44	12	1.79	7.79	IGNOMINIE	6	4	.24	7.84
AMANT	168	27	6.86	7.85	TU	3060	216	124.97	8.31	CONFITURE	8	5	.32	8.46
MANTELET	3	3	.12	8.49	PALISSANDRE	3	3	.12	8.49	REQUISITOIRE	3	3	.12	8.49
CE	9394	551	383.66	8.72	MA	1326	117	54.15	8.72	VICE	86	20	3.51	8.98
LEPRE	7	5	.28	9.11	VERTU	104	24	4.24	9.79	NE	10456	627	427.04	9.88
JE	11713	693	478.36	10.01	PAS	5606	378	228.96	10.05	HOMME	1474	139	60.20	10.37
SEDUCTEUR	13	8	.53	10.48	GREFFIER	10	7	.40	10.66	PROCUREUSE	9	7	.36	11.30
PURETE	54	19	2.20	11.56	PROCUREUR	13	10	.53	13.28	HIER	177	49	7.22	15.87
VOTRE	835	126	34.10	16.06	FEMME	1343	191	54.85	18.77	EST VERBE	7132	611	291.28	19.12
A VERBE	2921	333	119.29	19.97	MARI	355	106	14.49	24.54	VOUS	5162	634	210.82	29.75

VOCABULAIRE SIGNIFICATIF NEGATIF

PERE	429	1	17.52	-4.02	COMME	2857	72	116.68	-4.22	SE	3812	103	155.68	-4.31
GUERRE	444	0	18.13	-4.34	OU CONJ	2448	57	99.98	-4.38	CES	1776	35	72.53	-4.49
PAR	3411	87	139.31	-4.52	DU	4280	114	174.80	-4.69	ILS	1900	37	77.59	-4.70
DE	34147	11121	394.03	-7.72	DES	7307	164	298.43	-7.94	LES	12553	315	512.68	-8.91

Tableau 8. Le vocabulaire significatif de *Jérôme Bardini* et de *Pour Lucrèce*

De même l'inspiration des *Aventures de Jérôme Bardini*, dont le héros fuit le bonheur conjugal et rejoint un enfant vagabond au bord du Niagara, se marque fort clairement dans le champ sémantique circonscrit par les mots : *enfant, policeman, besace, vagabond, évasion, évadé, fuite, liberté, fui, disparition* (tableau 8).

On peut aussi se demander quels sont les mots dont l'absence, ou la faible représentation peut négativement caractériser une oeuvre. C'est le

vocabulaire significatif négatif. Il ne s'agit là que de mots courants, le plus souvent de mots-outils : car on ne peut s'étonner de l'absence d'un mot rare. Ainsi le verbe *avoir* et le verbe *être* appartiennent au vocabulaire significatif négatif des romans, au point même que ce seul critère suffit à désigner un texte comme roman.

On peut enfin chercher à l'extérieur du corpus la norme de référence : il ne s'agirait plus alors de comparer un texte de Giraudoux à l'ensemble des oeuvres de cet auteur, mais au corpus global des textes qui constituent la littérature de son temps et plus précisément la prose littéraire de la première moitié du XX^{e} siècle. Il est permis de contester le dosage des écrivains et des oeuvres dans le grand corpus du *Trésor de la Langue Française,* mais il faut reconnaître que l'énormité de ses dimensions (37 millions de mots pour la prose du XX^{e} siècle), et l'identité des normes de dépouillement offrent à la comparaison quelque légitimité : et dès lors c'est l'originalité de Giraudoux qui apparaîtra dans son vocabulaire significatif, positif et négatif, dans ses silences comme dans ses préférences. Ayant disposé du premier tome du dictionnaire de Nancy, nous avons pu vérifier le fait et constater par exemple combien l'emploi intensif de mots comme *abeille* ou *ablette* marque la tendresse que Giraudoux porte aux animaux et qu'il partage avec La Fontaine.

Mais si l'on s'en tient au seul corpus de Giraudoux, des regroupements de textes permettent d'isoler un genre déterminé, ou une époque particulière. Ainsi peut être constituée la liste des mots caractéristiques du roman giralducien (en ordre décroissant : *de, des, par, à, les, se, un, sa, son, ses, île, du, comme, sembler, dont, soudain, au, la, sur, chaque, ou,* etc.) et la liste caractéristique du théâtre : *être, vous, tu, avoir, je, te, ce, pas, toi, ne, votre, ta, oui, nous, moi, cela, tes, ton, vouloir, dire*, etc.[15] De la même façon le dictionnaire entier de Giraudoux peut être présenté par tranches chronologiques, ou par genres, ou par périodes à l'intérieur d'un genre. Le tableau 9 donne un extrait du dictionnaire ainsi réalisé.[16]

15. *C.U.M.F.I.D.*, n°7, 1973, p. 109.
16. *C.U.M.F.I.D.*, n^{o} 7, 1973, p. 121.

	R.1	R.2	R.3	R.4	R.T	MOD	ANT	COM	TRA	THE	TOT
CERTITUDE	0	24	13	22	16	11	11	33	0	I0	14
CERVEAU	29	36	22	44	30	35	44	66	27	38	34
CES	1193	2255	2464	1182	2074	1269	1359	1464	1172	1300	1776
CESSE	9	32	4	44	19	29	66	50	45	44	29
CESSE PP	0	8	9	0	6	23	0	25	0	10	8
CFT	310	413	788	669	550	480	595	485	534	520	539
CETTE	1504	1945	3424	3011	2484	2170	2322	2267	2245	2254	2396
CEUX	436	491	410	535	455	504	451	460	443	461	458
CHACUN	281	335	337	200	317	160	161	117	181	173	262
CHACUNE	77	163	76	22	107	59	22	25	27	36	80
CHAGRIN	97	32	40	22	45	17	22	8	22	20	36
CHAISE	29	49	58	0	45	35	16	33	22	25	38
CHAINETTE	0	12	13	0	9	5	5	8	4	5	8
CHAIR	48	53	108	200	83	154	116	100	140	126	100
CHAISE	97	53	49	66	60	47	0	50	9	23	46
CHALDEEN	0	24	0	0	9	0	0	0	0	0	6
CHALET	19	28	9	0	17	5	0	0	0	2	12
CHALEUR	77	20	72	0	47	17	61	41	36	36	43
CHAMBELLAN	0	8	0	0	3	0	0	0	72	41	18
CHAMBRE	291	208	324	379	276	183	178	292	108	171	236
CHAMBRIERE	9	8	27	0	14	5	5	8	0	5	11
CHAMEAU	0	8	0	0	3	0	11	0	13	7	5
CHAMOIS	0	28	4	0	13	11	0	8	0	5	10
CHAMP	106	110	130	89	115	53	77	58	67	62	95
CHAMPAGNE	0	28	18	0	17	11	0	16	0	5	13
CHAMPIGNON	0	20	4	0	9	5	11	8	9	7	9
CHAMPION	9	12	27	22	17	17	0	25	0	7	14
CHANCE	106	94	72	133	91	213	167	167	162	178	125
CHANDAIL	0	12		0	6	5	0	0	0	2	5
CHANGE	9	28	31	44	27	5	5	0	9	5	19
CHANGE PP	38	53	58	111	56	65	21	58	31	49.	54
CHANGEANT	19	16	18	0	16	5	16	0	13	10	14
CHANGEMENT	0	24	63	66	37	29	11	16	27	20	31
CHANGER	135	106	198	200	151	201	100	142	117	147	150
CHANSON	19	16	4	22	13	41	5	58	4	20	16
CHANT	19	102	45	22	61	29	150.	41	117	90	73
CHANTAGE	9	8	4	22	8	29	38	50	27	33	18
CHANTANT	19	8	0	0	6	41	0	50	58	51	24
CHANTE PP	0	12	18	22	13	47	0	16	9	20	16
CHANTER	164	65	63	66	81	177	94	184	117	152	109
CHANTEUR	9	28	9	22	17	77	22	100	27	51	31
CHANTIER	29	4	4	0	8	0	5	8	0	2	6
CHAOS	0	12	9	0	8	0	22	25	4	10	9
CHAPEAU	232	98	103	89	122	88	0	75	9	38	90
CHAPELLE	9	53	13	0	27	17	0	8	13	12	22
CHAPITRE	126	24	31	44	45	17	5	25	0	10	32
CHAQUE	1329	1293	806	401	1058	480	484	326	538	497	842
CHAR	38	20	0	0	14	11	11	0	9	10	13
CHARBON	9	4	4	0	4	17	11	8	22	18	10
CHARDON	9	20	0	0	9	5	11	8	9	7	9
CHARDONNERET	0	12	0	0	4	0	11	0	9	5	5
CHARGE	29	8	36	66	26	35	33	33	31	33	29

R.1 :romans 1ère période, de *Provinciales* à *Simon le P.* ; R.2 : romans 2 ème période ; R•3 : romans 3ème période . Le cycle de Fontranges ; R•4 : derniers romans ; R.T : tous romans groupés ; MOD : pièces modernes ; ANT ; pièces antiques ; COM : comédies ; TRA : pièces tragiques ; THE : ensemble des pièces ; TOT : fréquence totale .

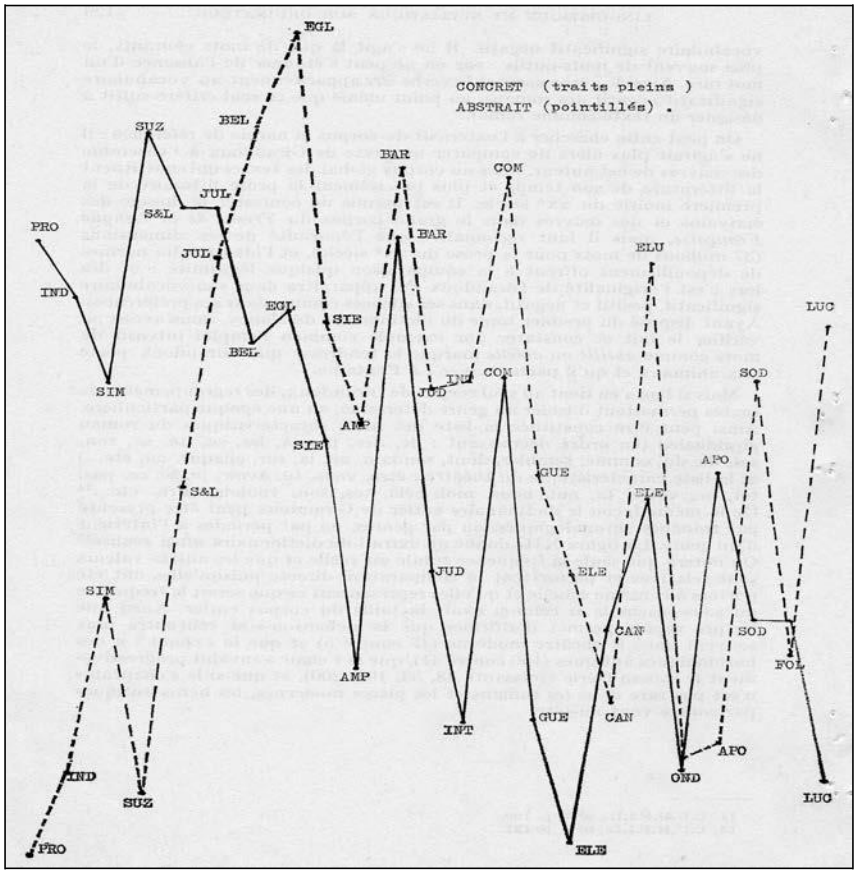
Tableau 9. Fréquences relatives par genres et par périodes

On notera que seule la fréquence totale est réelle et que les autres valeurs sont relatives et permettent la comparaison directe puisqu'elles ont été portées à la même échelle et qu'elles représentent ce que serait la fréquence du sous-ensemble si celui-ci avait la taille du corpus entier. Ainsi une lecture rapide permet d'affirmer que la *chanson* se rencontre plus souvent dans le théâtre moderne (45 contre 5) et que le *chant* a des harmoniques antiques (150 contre 41), que la *chair* envahit progressivement le roman (série croissante 48, 53, 108, 200), et que si le *chapeau* n'est pas rare dans les romans et les pièces modernes, les héros antiques par contre vont nu-tête.

3.2. Les champs sémantiques

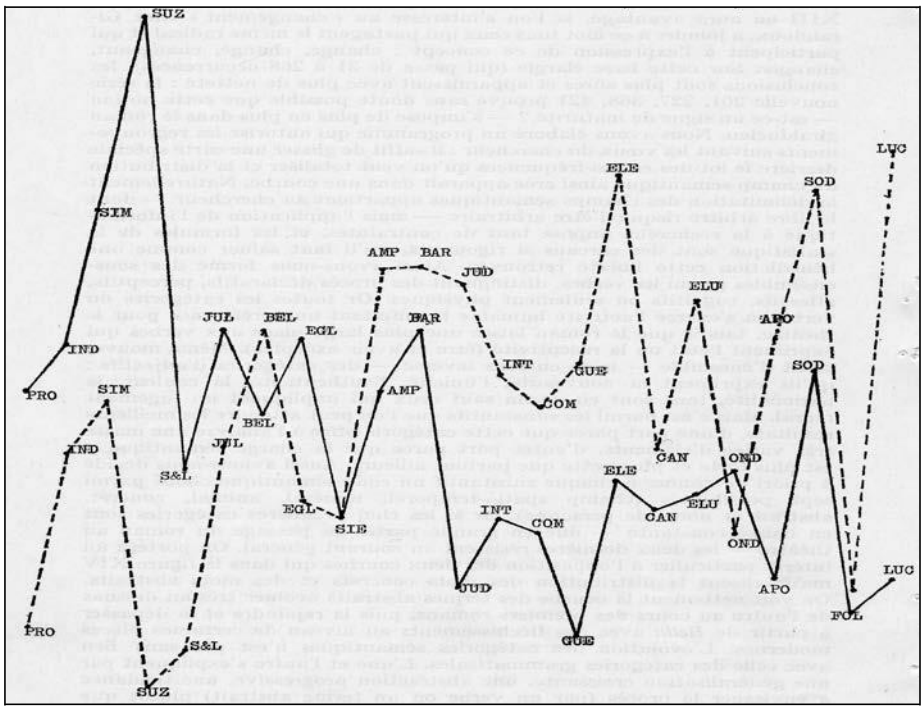
Si l'on peut regrouper les textes, on peut aussi regrouper les mots qui s'inscrivent dans le même champ sémantique. Ainsi en reprenant le tableau 9 on aura avantage, si l'on s'intéresse au *changement* chez Giraudoux, à joindre à ce mot tous ceux qui partagent le même radical et qui participent à l'expression de ce concept : *change, changé, changeant, changer*. Sur cette base élargie (qui passe de 31 à 268 occurrences), les conclusions sont plus sûres et apparaissent avec plus de netteté : la série nouvelle 201, 227, 368, 421 prouve sans doute possible que cette notion – est-ce un signe de maturité ? – s'impose de plus en plus dans le roman giralducien. Nous avons élaboré un programme qui autorise les regroupements suivant les voeux du chercheur : il suffit de glisser une carte spéciale derrière le lot des cartes-fréquences qu'on veut totaliser et la distribution du champ sémantique ainsi créé apparaît dans une courbe. Naturellement la délimitation des champs sémantiques appartient au chercheur dont le libre arbitre risque d'être arbitraire – mais l'application de l'informatique à la recherche impose tant de contraintes, et les formules de la statistique sont des carcans si rigoureux, qu'il faut saluer comme une bénédiction cette liberté retrouvée. Ainsi avons-nous formé des sous-ensembles parmi les verbes, distinguant des procès déclaratifs, perceptifs, affectifs, cognitifs ou seulement physiques. Or toutes les catégories du verbe où s'exerce l'activité humaine manifestent une préférence pour le théâtre, tandis que le roman laisse une plus large place aux verbes qui expriment l'état ou la réceptivité (*être* et *avoir* exceptés). Même mouvement d'ensemble – mais en sens inverse – des catégories d'adjectifs ; qu'ils expriment la nouveauté, l'unicité, l'authenticité, la couleur, la nationalité, tous sont en déclin sauf ceux qui impliquent un jugement moral.

Mais c'est parmi les substantifs que l'on peut attendre les meilleurs résultats, d'une part parce que cette catégorie offre à l'analyse une masse très variée d'éléments, d'autre part parce que la charge sémantique y est plus forte et plus nette que partout ailleurs. Aussi avions-nous décidé à priori de donner à chaque substantif un code sémantique choisi parmi sept possibilités (champ spatio-temporel, minéral, animal, concret, abstrait et noms de personne). Or si les cinq premières catégories sont en baisse constante – due en grande partie au passage du roman au théâtre – les deux dernières résistent au courant général.



Graphique 5. Courbes du concret et de l'abstrait

On portera un intérêt particulier à l'opposition des deux courbes qui dans le graphique 5 matérialisent la distribution des mots concrets et des mots abstraits. On voit nettement la courbe des termes abstraits évoluer très au dessous de l'autre au cours des premiers romans, puis la rejoindre et la dépasser à partir de *Bella* avec des fléchissements au niveau de certaines pièces modernes. L'évolution des catégories sémantiques n'est pas sans lien avec celle des catégories grammaticales. L'une et l'autre s'expliquent par une généralisation croissante, une abstraction progressive, une tendance à envisager le procès (par un verbe ou un terme abstrait) plutôt que l'objet, la relation plutôt que la découverte, la morale plutôt que l'esthétique, l'action plutôt que la description, et l'homme plutôt que le monde[17].



Graphique 6. La nature (traits pleins) et l'homme (pointillés)

17. Les données quantitatives confirment donc les conclusions auxquelles est parvenu dans sa thèse R. Marill Albérès, *Esthétique et morale chez Jean Giraudoux*, Nizet, 1957, 569 p.

Ainsi en distinguant parmi les substantifs deux zones qui correspondant respectivement à l'homme et à la nature, on fait apparaître deux mouvements opposés : la courbe de la nature culmine avec *Suzanne* perdue sur son île loin des humains et descend par degrés jusqu'à la fin, tandis que celle du champ humain s'élève progressivement jusqu'au sommet de la dernière pièce (graphique 6).

Ces deux tendances opposées se manifestent avec un relief éclatant dans les trois substantifs les plus employés de Giraudoux :

homme	1474	occurrences,	corrélation	chronologique	R =	+ 0,58
jour	1368	«	«	«	R =	- 0,46
femme	1343	«	«	«	R =	+ 0,56

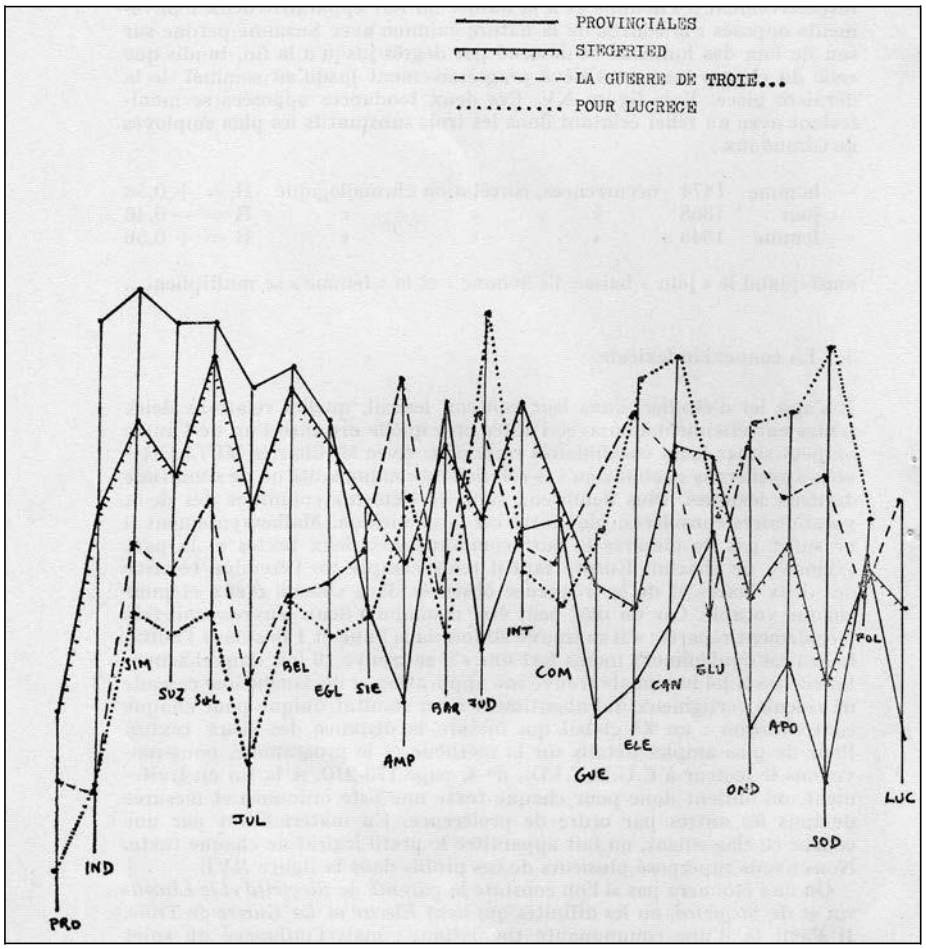
Ainsi quand le *jour* baisse, l'*homme* et la *femme* se multiplient...

3.3. La connexion lexicale

Il s'agit ici d'étudier, dans leur contenu lexical, quelles relations deux textes entretiennent l'un avec l'autre et à quelle distance l'un de l'autre on peut situer leurs vocabulaires respectifs. Avec Charles Muller[18] nous appellerons *connexion* la relation de communauté ou de similitude de deux lexiques. Plus nombreux sont les éléments communs des deux vocabulaires considérés, plus forte est la connexion. Malheureusement il ne suffit pas de mesurer la part commune aux deux textes et la part exclusive de chacun. Encore faut-il tenir compte de l'étendue relative des deux textes et de la fréquence observée dans chacun d'eux et pour chaque vocable. Car un mot peut être commun à deux oeuvres mais très inégalement réparti : s'il se trouve 20 fois dans l'une et 1 fois dans l'autre, le lien est évidemment moins fort que s'il se trouve 10 fois dans chacune. Là encore la loi binomiale trouve son application et déclanche une cascade de calculs vertigineux qui aboutissent à un résultat unique pour chaque confrontation : un X2 global qui mesure la distance des deux textes. Pour de plus amples détails sur la méthode et le programme, nous renvoyons le lecteur à *C.U.M.F.I.D.*, n°4, page 173-210. A la fin du traitement, on obtient donc pour chaque texte une liste ordonnée et mesurée de tous les autres par ordre de préférence. En

18. Charles Muller, *Initiation à la statistique linguistique*, Hachette, 1973, p. 211 et sqq.

matérialisant par une courbe ce classement, on fait apparaître le profil lexical de chaque texte. Nous avons superposé plusieurs de ces profils dans la graphique 7.



Graphique 7. Connexion lexicale (étude de contenu)

On ne s'étonnera pas si l'on constate la parenté de *Siegfried et le Limous*in et de *Siegfried*, ou les affinités qui lient *Electre* et *La Guerre de Troie*. Il s'agit là d'une communauté thématique ; mais l'influence du sujet

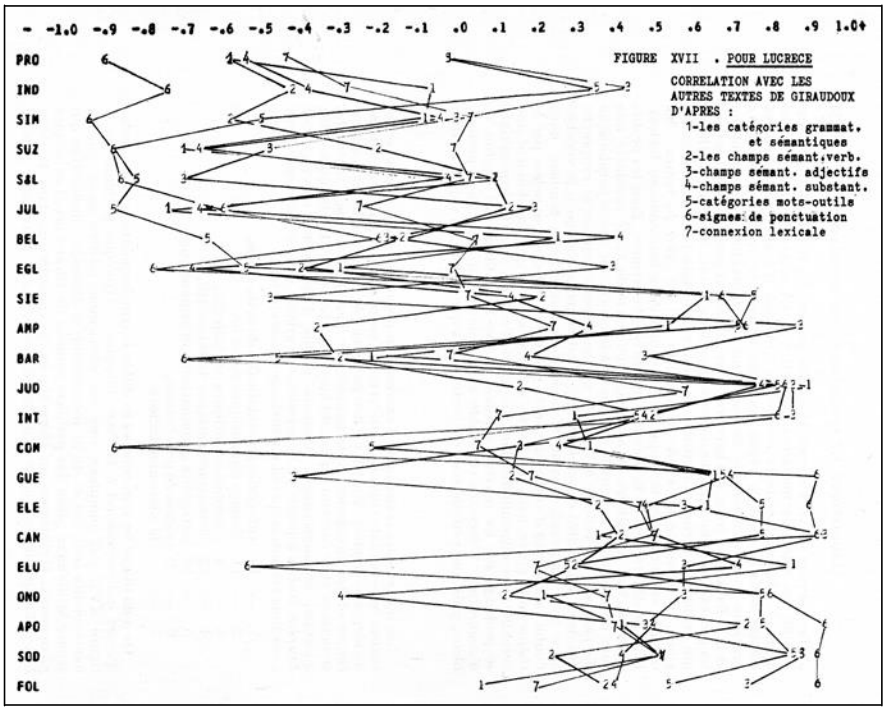
s'efface le plus souvent devant la loi des genres : les romans forment une famille, les comédies modernes une autre et les tragédies antiques une autre. La révolution de la *Préface de Cromwell* n'a donc pas aboli les classes sociales dans la république des lettres, et quoique Giraudoux pratique avec insolence le déclassement des individus, l'anachronisme et le mélange des tons, des barrières n'en subsistent pas moins qui isolent les genres.

Cependant par dessus les oppositions de ton, de genre, de thème qui expliquent certaines ruptures, certains reliefs dans la chaîne chronologique, on constate un ample mouvement qui est celui du temps : les oeuvres voisines dans le temps le sont aussi dans leur contenu lexical. Rien ne diffère plus de la première oeuvre que la dernière. Tous les profils observés font apparaître au voisinage immédiat du texte considéré une zone de connexion maximale qui donne à la courbe un dessin parabolique (la parabole est réduite à l'une de ses branches lorsqu'on se rapproche des deux extrémités chronologiques). Tout se passe comme si le vocabulaire était mouvant, vivant et périssable. Des cellules tombent dans l'oubli tandis que d'autres naissent, non point tellement appelées par le thème ou le genre, mais par la vie, les circonstances qui leur donnent une actualité, une disponibilité dans l'esprit de l'auteur, et cette disponibilité se prolonge sur les oeuvres voisines, même très différentes de ton et d'inspiration. L'écrivain ressemble à un peintre qui aurait une période bleue, puis une période rose ou verte, quels que soient les sujets traités.

Les traitements statistiques s'exercent en des domaines si variés, et les indices relevés sont si divers qu'au terme de notre étude se pose le problème de leur convergence. Y a-t-il un lien par exemple entre les champs sémantiques et les catégories grammaticales et plus généralement entre les structures et le contenu du vocabulaire ?

Le graphique 8 répond à cette question, en superposant plusieurs profils d'une même oeuvre, *Pour Lucrèce,* établis à partir de la corrélation variable que ce texte entretient avec tous les autres quand on envisage successivement :

1 – 18 catégories grammaticales de mots-outils,
2 – 11 champs sémantiques du verbe,
3 – 9 champs sémantiques de l'adjectif,
4 – 12 champs sémantiques du substantif,
5 – 14 catégories grammaticales et sémantiques,
6 – 9 signes de ponctuation,
7 – la connexion lexicale.



Graphique 8. *Pour Lucrèce*. Corrélation avec les autres textes de Giraudoux

On a distingué les sept courbes en différenciant les points par un numéro d'ordre qui correspond à la liste ci-dessus. Quoique le détail du graphique soit complexe, une vue d'ensemble montre assez que les mouvements sont liés et qu'ils manifestent la distance considérable qui sépare la dernière pièce des premiers romans de Giraudoux. Les mêmes plissements, stylistiques ou thématiques, y apparaissent dans le relief lexical lorsqu'un genre succède à

l'autre, tandis que le mouvement général de la chaîne s'oriente suivant l'axe chronologique. Ces courbes ressemblent à des courbes de niveau...

Rappelons que plusieurs fois déjà l'occasion s'est offerte dans les pages qui précèdent de souligner la cohérence des résultats et la convergence des approches. Ainsi le classement qu'on obtient à partir de la richesse lexicale s'accorde avec celui qui se fonde sur l'originalité du vocabulaire – laquelle peut elle-même être estimée de trois manières concordantes (accroissement, mots exclusifs, mots significatifs) –, il s'accorde aussi – parfois au prix d'une inversion – avec les classements qui rendent compte de la longueur du mot, de celle de la phrase, de la distribution des verbes, ou des substantifs, ou des mots-outils[19] ; la convergence est parfois si étroite qu'elle rassure et inquiète à la fois. Elle rassure sur la fiabilité des résultats – dont le contrôle manuel n'est pas toujours aisé lorsque les calculs s'enchaînent à l'infini : les résultats aberrants ont en effet la propriété de s'accorder rarement entre eux. Elle pourrait inquiéter aussi en suggérant l'hypothèse de quelque redondance : car la cohérence des mesures n'a rien de si admirable quand c'est toujours la même chose qu'on mesure. Et en effet il s'agit partout d'un même objet, la langue et le style d'un auteur, ensemble complexe mais en même temps système clos, surdéterminé, où les liaisons sont multipliées et les atomes doublement crochus. Le langage – qu'on dit chose humaine, incertaine et imprévisible – est un bloc compact qui laisse peut-être moins de place au hasard que la physique nucléaire. Le langage est comme une boule : on peut multiplier à l'infini les angles de vue, c'est toujours la même image qu'on obtient.

En ce qui concerne Giraudoux tout conduit à la conclusion que chez ce professionnel de la jeunesse que fut à sa manière Giraudoux, le temps a pourtant fait son oeuvre. Une évolution naturelle le conduit de la fantaisie à la gravité, du concret à l'abstrait, de la richesse à la sobriété, des expériences à l'expérience, du roman au théâtre. La succession des périodes et la distinction des genres s'accordent chez lui : plus que le hasard de la route, c'est l'évolution nécessaire de son cheminement intérieur qui a poussé Giraudoux sur la scène.

---

19. Ou même de la distribution des phonèmes.

Le cas de Giraudoux ouvre deux directions de recherche comparative. La première concerne la critique des genres littéraires : les faits observés chez Giraudoux résultent-ils de choix personnels, volontaires ou spontanés, ou bien sont-ils déterminés par des formes à priori, par la loi des genres ? La seconde met en cause la loi de la nature : est-ce qu'en vieillissant tous les auteurs vieillissent, comme Giraudoux[20], et cèdent – ou tendent – comme lui à l'abstraction et au dépouillement ? La réponse à ces questions ne peut venir que d'études parallèles dont le faisceau devrait éclairer un peu mieux la vérité du style, de la langue et de l'homme sous le feu croisé des comparaisons.

20. Corneille, étudié par Charles Muller dans son *Etude de statistique lexicale* (Klincksieck, 1964) manifeste des symptômes assez semblables à ceux de Giraudoux.