

A. SALEM

La typologie des segments répétés dans un corpus, fondée sur l'analyse d'un tableau croisant mots et textes

Les cahiers de l'analyse des données, tome 9, n° 4 (1984), p. 489-500

http://www.numdam.org/item?id=CAD_1984__9_4_489_0

© Les cahiers de l'analyse des données, Dunod, 1984, tous droits réservés.

L'accès aux archives de la revue « Les cahiers de l'analyse des données » implique l'accord avec les conditions générales d'utilisation (<http://www.numdam.org/conditions>). Toute utilisation commerciale ou impression systématique est constitutive d'une infraction pénale. Toute copie ou impression de ce fichier doit contenir la présente mention de copyright.

NUMDAM

Article numérisé dans le cadre du programme
Numérisation de documents anciens mathématiques

<http://www.numdam.org/>

LA TYPOLOGIE DES SEGMENTS RÉPÉTÉS DANS UN CORPUS, FONDÉE SUR L'ANALYSE D'UN TABLEAU CROISANT MOTS ET TEXTES

[SEG. REP.]

par A. Salem

0 Cooccurrence et environnement syntagmatique : On a souvent recours à l'Analyse Factorielle des Correspondances afin d'obtenir une typologie des différentes parties d'un ensemble de textes. Pour construire cette typologie on commence, le plus souvent, par effectuer le relevé dans chacune des parties du corpus, des occurrences de toutes les formes du vocabulaire qu'elle contient.

Certaines études qui se donnent pour objectif l'indexation automatique des textes (1) écartent d'emblée une série de "mots-outils", ou formes fonctionnelles, réputés fréquents et bien répartis. Cette manière de procéder permet de réduire considérablement la liste des formes soumises à l'analyse. On peut alors employer des algorithmes de Classification Ascendante Hiérarchique pour rapprocher les formes fortement cooccurentes dans les mêmes parties.

Dans les études lexicologiques que nous pratiquons au laboratoire de St-Cloud nous avons dû renoncer à ce genre de solution devant l'impossibilité, maintes fois constatée, d'établir, *a priori* et une fois pour toutes, une liste de formes fonctionnelles satisfaisante pour l'ensemble des textes que nous étudions.

D'autre part, certains de ces mots fonctionnels jouent dans l'analyse des textes politiques un rôle de premier plan (cf. par exemple le paradigme : LA LIBERTE, LES LIBERTES, etc.).

Enfin, nombre de linguistes s'intéressent tout particulièrement aux catégories grammaticales formant ce que l'on appelle à la suite d'Emile BENVENISTE : l'appareil formel de l'énonciation (pronoms personnels, pronoms et adjectifs possessifs, pronoms et adjectifs démonstratifs, etc.) en ce qu'elles sont les traces explicites de la présence du locuteur dans le texte qu'il a produit.

Partant du décompte des occurrences de chacune des formes graphiques dans les parties du corpus analysé, l'A.F.C. fournit des facteurs portant à la fois sur les parties comparées et sur le vocabulaire qu'elles utilisent.

En parcourant les listes de formes les plus fréquentes dans un sous-ensemble du corpus, on est souvent tenté, d'après les affinités qu'on devine entre certaines d'entre elles, de reconstruire des lexies figées, des locutions, dont on suppose qu'elles expliquent la présence simultanée de ces formes dans certaines parties du texte.

(1) Voir par exemple : Akila Att Hamlat. *Analyse des Répétitions et Indexation Automatique des documents*. C.A.D. Vol IX (1984) n° 2 pp 173-204.

(*) Laboratoire : *Lexicologie et Textes Politiques*. U.R.L. 3 de l'I.N.A.L.F. (C.N.R.S.) E.N.S. de St Cloud.

Mais cette opération doit être effectuée avec la plus grande prudence. Nombreux sont en effet les cas où le retour au texte montrera, en fin de compte, que ces associations n'existent que dans l'imagination du chercheur, alors même qu'il tentait, par la mise en oeuvre d'une méthode statistique réputée objective, de considérer les textes qu'il étudie avec moins de subjectivité.

Ces risques sont encore plus grands lorsqu'on a affaire à des textes produits à une époque éloignée de la nôtre, tels ceux qui forment le corpus de la presse révolutionnaire (1793) que nous étudions au laboratoire de St-Cloud.

L'existence de tournures figées couramment employées à l'époque mais tombées par la suite en désuétude et aujourd'hui perdues, la constitution, à l'inverse, dans la période qui a suivi d'autres associations lexicales qui nous sont familières mais qui n'existaient nullement alors, constituent autant de pièges qui finiront à la longue, par mettre en défaut la compétence décalée du chercheur.

Pour toutes ces raisons, l'étude de l'environnement syntagmatique des formes caractéristiques de chaque type de texte se révèle tout à fait indispensable pour la connaissance des corpus que l'on étudie. Cependant, la plupart du temps, la perte des liens syntagmatiques du texte qui résulte de son morcellement en occurrences de formes graphiques (ou de vocables, si l'on tient à la lemmatisation) devient, au moment de l'interprétation des résultats fournis par les procédures statistiques un obstacle important à la production de tout commentaire d'ordre textuel sur les grandes oppositions constatées dans l'emploi du vocabulaire.

Dans cette étude nous tenterons de convaincre le lecteur que l'utilisation conjointe des Inventaires de Segments Répétés (2), élaborés au laboratoire de Lexicologie politique de l'E.N.S. de St-Cloud d'une part, et des techniques de "mise en élément supplémentaire" couramment utilisée en Analyse Factorielle des Correspondances d'autre part, peut lui permettre, et ce sans faire d'hypothèses supplémentaires, de replacer automatiquement les formes sélectionnées par les méthodes statistiques dans leur environnement syntagmatique récurrent, de sélectionner les associations lexicales les plus caractéristiques pour chacun des grands types de texte mis en évidence.

1 *Les tableaux lexicométriques* : Qu'il s'agisse de faire des parentages entre des textes d'auteurs différents en rapprochant ceux qui ont un vocabulaire similaire, de construire une typologie des chapitres d'une même oeuvre ou encore de chercher à dégager les changements survenus dans le vocabulaire d'une organisation politique ou syndicale au cours d'une période donnée, les applications de l'A.F.C. aux corpus de données textuelles se font sensiblement de la même manière.

On commence par se fixer des règles de segmentation qui permettront de reconnaître au fil du texte les occurrences des différents mots graphiques (nous dirons : formes) qui constituent le vocabulaire du corpus étudié.

(2) *La méthode des I.S.R. est issue d'un travail en collaboration avec Pierre LAFON que je tiens à remercier tout particulièrement ici.*

J'ai aussi bénéficié des conseils amicaux et des critiques constructives de : Maurice TOURNIER, Annie GEFFROY, Pierre FIALA, Jacques GUILHAUMOU et de tous les autres membres du laboratoire de St-Cloud.

Dans un deuxième temps, on construit un tableau à double entrée qui compte autant de colonnes que le corpus a de parties, autant de lignes que le corpus de textes compte de formes différentes. A l'intersection de la colonne *j* et de la ligne *i* de ce tableau on note le nombre de fois que la forme *i* est employée dans la partie *j*.

Si l'on prend soin de classer ensuite les lignes du tableau ainsi constitué en commençant par la forme la plus fréquente dans le corpus (en général dans un texte français la préposition "DE") et en finissant par les formes qui n'apparaissent qu'une seule fois et que l'on nomme hapax, les formes de même fréquence étant rangées par ordre alphabétique, on donne alors à ce tableau une forme "canonique" que nous avons proposé d'appeler le Tableau Lexicométrique Entier (T.L.E.) du corpus ainsi partitionné.

On peut voir sur la figure 2 (§ 3) un fragment d'un tableau de ce type correspondant aux formes les plus fréquentes du corpus HR8 décrit ci dessous (§2).

La gamme des fréquences des formes qui composent un texte et par voie de conséquence la gamme des fréquences marginales des lignes du T.L.E., présente ordinairement une régularité qu'on exprime par des lois associées au nom de G.K. ZIPF.

La pratique assidue, et presque exclusive, de ce genre de tableau nous a permis de dégager certains enseignements touchant à l'A.F.C. des Tableaux Lexicométriques, qu'ils soient entiers ou amputés de la partie qui correspond aux formes les moins fréquentes (3). Ainsi, dans la plupart des cas les résultats varient peu si l'on retire les hapax du tableau analysé. Cette règle permet d'alléger considérablement les tableaux de résultats fournis par l'A.F.C. .

S'il est en général assez aisé, et toujours très instructif, de dépouiller les facteurs sur l'ensemble des parties, qui dépasse rarement la centaine d'unités dans les études que nous faisons, il en va tout autrement en ce qui concerne le dépouillement des facteurs sur l'ensemble des formes qui compte souvent plusieurs milliers d'unités. Pour que cette opération puisse être réalisée dans des délais raisonnables, il faut, la plupart du temps, recourir à des procédures automatisées de tri sur les valeurs du facteur, qui sélectionneront un sous-ensemble de formes jugées les plus intéressantes, en se basant pour cela, soit sur les coordonnées factorielles, soit sur les contributions absolues au facteur.

Notre habitude est de nous borner aux formes dont la contribution absolue dépasse un certain seuil (par exemple 3 fois la contribution moyenne) formes qui sont ensuite rangées suivant les valeurs décroissantes du facteur.

2 Le corpus HR8 : Nous baserons notre exposé sur un corpus que nous avons souvent eu l'occasion d'étudier avec des méthodes diverses : le corpus HR8.

Ce corpus est constitué de seize journaux révolutionnaires tous parus à Paris au cours de l'été 1793. Ces livraisons ont été choisies essentiellement pour leur taille à seule fin de permettre des expériences lexicométriques simples, sur des données présentant un maximum d'homogénéité tant sur le plan des conditions socio-historiques de production de ces textes que sur celui de leurs caractéristiques

(3) Voir : *Signalement et inventaire lexical : Textes politiques Français de 1793. in Pratique de l'Analyse des Données Tome 3, DUNOD, Paris 1981.*

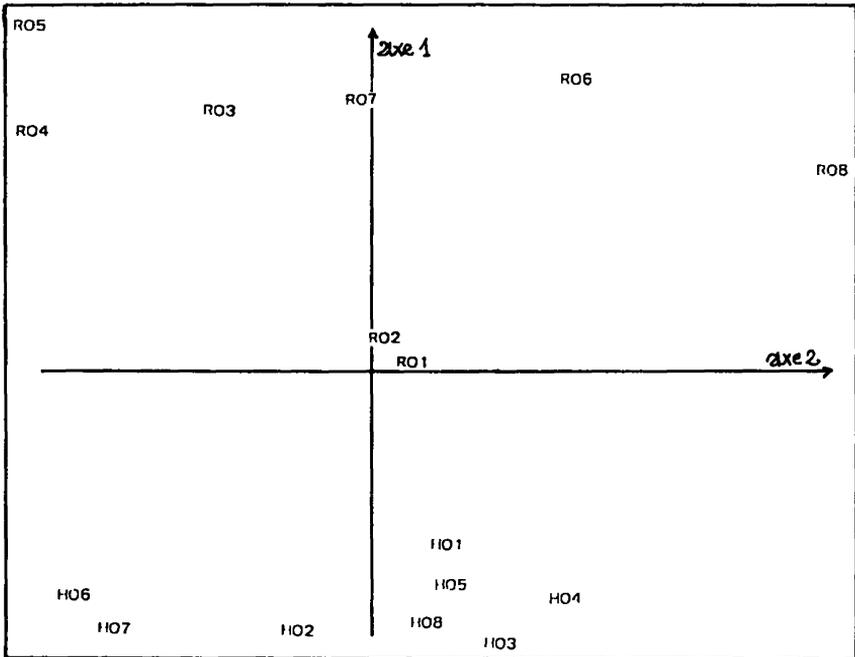
quantitatives d'ensemble (4).

Les huit premiers textes notés R1, R2, ..., R8, sont des livraisons du "Publiciste de la République Française" écrits par Jacques ROUX (1752-1794), un des chefs de file du mouvement des "enragés".

Les huit autres, des numéros du célèbre "Père Duchesne" dans lesquels s'exerçait la verve de Jacques-René HEBERT (1757-1794) membre du Club des Cordeliers, sont notés H1, H2, ..., H8.

Ce corpus compte 25023 occurrences pour 4403 formes, réparties à peu près pour moitié entre les deux auteurs. Le texte le plus long H4 compte 1675 occurrences, le plus court R8 en compte 1427.

LES FORMES REPELEES DU CORPUS rR8																
J1	Q1T	POID	INRT	1#F	COR	CTR	2#F	COR	CTR	3#F	COR	CTR	4#F	COR	CTR	
1	RO1	496	57	65	68	3	1	66	2	2	-896	433	347	40	1	1
2	RO2	554	63	64	82	4	2	45	1	1	-686	282	225	186	21	18
3	RO3	415	64	62	470	138	77	-227	32	24	-344	74	57	-203	26	22
4	RO4	500	66	65	423	110	64	-513	162	125	210	27	22	-451	125	110
5	RO5	927	63	75	638	206	138	-498	126	112	318	51	48	1034	541	545
6	RO6	353	65	67	570	192	114	305	55	43	93	5	4	-130	10	9
7	RO7	437	64	64	534	172	98	-23	0	0	118	8	7	-627	238	205
8	RO8	819	57	75	424	83	56	1072	531	472	470	102	95	134	8	8
9	HO1	393	58	60	-304	54	29	87	4	3	-446	117	87	152	14	11
10	HO2	151	59	58	-429	115	59	-73	3	2	146	13	10	-23	0	0
11	HO3	235	65	57	-484	161	87	184	23	16	-13	0	0	-37	1	1
12	HO4	303	67	60	-400	110	59	285	55	39	68	3	2	283	55	44
13	HO5	124	63	53	-348	87	41	104	8	5	180	23	15	-38	1	1
14	HO6	293	61	56	-406	109	55	-432	123	82	272	49	34	41	1	1
15	HO7	292	63	63	-454	125	70	-391	93	69	259	41	32	-131	11	9
16	HO8	160	65	55	-391	110	54	94	6	4	179	23	16	-169	21	15
		22466.00		1000			1000			1000			1000			1000



[SEG. RÉP]: § 2; fig 1 : Analyse du tableau (1848 formes) x (16 textes).

(4) Voir A. Salem article paru dans M.O.T.S. n° 4 (cf. bibliographie).

On peut voir sur le tableau 1 les résultats de l'analyse du tableau (16 textes x 1848 formes répétées dans le corpus) sur l'ensemble des journaux. On constate sur le premier facteur une opposition assez nette entre les deux groupes de textes. Les numéros du "Père Duchesne" ont tous en effet sur cet axe une coordonnée négative qui les oppose aux numéros écrits par J. ROUX, lesquels ont une coordonnée positive.

Notons au passage que l'opposition que l'on peut constater sur les deuxième et troisième facteurs est due comme c'est souvent le cas dans ce genre d'analyse à l'utilisation, tour à tour par chacun des deux auteurs, de sous-systèmes particuliers de la catégorie des substituts et référents personnels : JE, IL, NOUS, VOUS, ON... et des formes auxiliaires qui leur sont associées : AI, A, ETIONS, ETIEZ...

On examinera plus loin les principales formes responsables du clivage constaté entre les deux séries de textes sur le premier axe factoriel (figures 3 et 4). Signalons seulement ici que l'on retrouve côte à côte sur ce tableau les formes NOUS et TOUS particulièrement abondantes dans le discours du "Père Duchesne" dont on se souvient qu'il prône l'unité tout au long de cette période 1793 (5).

On peut penser tout d'abord à des locutions en "NOUS TOUS". Le retour au texte pour ces formes qui comptent respectivement 164 et 151 occurrences se révèle une opération assez fastidieuse même pour un corpus aussi modeste. Il montrera en fin de compte que les formes NOUS et TOUS n'apparaissent que très exceptionnellement dans les mêmes phrases dans le corpus HR8.

3 Les inventaires de segments répétés (I.S.R.) : Un ensemble de programmes conçus au laboratoire de St-Cloud permet de constituer l'inventaire de toutes les séquences de formes, que nous appellerons segments de longueur 2, 3,... et ce jusqu'à 25), qui se rencontrent plusieurs fois dans un corpus de textes (6).

On voit tout l'intérêt que peuvent présenter les I.S.R. pour les différents spécialistes de l'étude des textes. L'historien tentera à travers eux d'étudier la circulation de divers mots d'ordre au cours de la période étudiée en concentrant son attention sur les segments les plus longs tels par exemple : METTRE A L'OMBRE TOUS LES HOMMES SUSPECTS que l'on rencontre deux fois chez HEBERT.

A l'inverse le grammairien s'intéressera sans doute plus particulièrement aux syntagmes figés les plus fortement récurrents et à leurs variations internes : CONTRE LA MONTAGNE - CONTRE LA REPUBLIQUE - CONTRE LES CITOYENS.

Sans multiplier les exemples, nous dirons que les segments répétés peuvent représenter soit des locutions fonctionnant comme un seul mot (ex : "mettre à l'ombre" pour "emprisonner") soit des équivalents en plusieurs mots de formes fléchées ("de la ville" en français pour "urbis" en latin) soit les produits d'une rhétorique de la répétition ("faudrait-il leur rappeler que"... (bis) ; (ter)...). La distinction entre segment répété et mot unique étant souvent conventionnelle dans la mesure où la segmentation précise du discours en mots est par essence postérieure au langage.

Ceux qui se préoccupent d'études quantitatives tenteront de soumettre à des procédures d'analyse statistique le tableau qui résulte du décompte des occurrences de chacun des segments répétés dans les différentes parties du texte.

(5) Voir le travail de Jacques GUILHAUMOU.

(6) Voir : P.Lafon, A. Salem, article paru dans M.O.T.S. n° 6.

	R1	R2	R3	R4	R5	R6	R7	R8	H1	H2	H3	H4	H5	H6	H7	H8
1 DE	79	85	88	79	84	93	88	78	64	73	66	78	76	82	92	58
2 LA	41	64	63	67	53	76	82	50	35	59	39	55	61	50	56	51
3 LES	37	53	58	54	61	60	71	37	53	57	69	43	51	59	56	57
4 ET	48	52	43	59	32	43	40	31	31	42	39	39	42	35	40	30
5 DES	30	37	47	40	33	43	79	30	19	18	27	16	26	22	33	26
6 LE	28	27	43	44	26	29	37	38	31	23	33	23	27	26	28	41
7 à	20	36	23	33	49	34	17	33	18	30	33	35	40	37	30	26
8 QUE	32	31	33	23	29	25	21	26	27	29	26	24	23	33	36	29
9 L	28	24	37	38	21	42	29	22	19	15	21	17	15	22	24	19
1 DE LA LIBERTE ET DE L'EGALITE	0	0	3	0	0	0	1	0	0	0	0	0	1	0	0	0
2 A L'ARMEEE DES COTES DE BREST	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3 LA FACTION SCCELERATE DES HOMMES D'ETAT	0	0	0	0	0	0	1	2	0	0	0	0	0	0	0	0
4 NIERAS TU QUE TU AS VOULU AFFAMER	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0
5 S IL EST VRAI QUE VOUS LA	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0
6 A LA TETE DE LA PLUS BELLE	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0
7 CHEFS DE LA FACTION SCCELERATE DES HOMMES	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0
8 COUP DE COLLIER POUR SAUVER LA REPUBLIQUE	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9 DES TRAITRES ET DES SANGSUES DU PEUPLE	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0
20 DIREZ VOUS QUE LES DEPUTES MONTAGNARDS	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0
21 DONNER LA CLEF DES CHAMPS AU PETIT	0	0	0	0	0	0	1	0	0	0	0	0	2	0	0	0
2225 DE LA	19	21	23	26	17	28	36	19	4	14	7	10	15	9	24	6
2226 TOUS LES	4	9	1	1	0	3	2	0	12	12	14	3	12	9	14	12
2227 à LA	5	6	4	9	2	10	7	11	5	4	6	9	7	9	2	10
2228 DE L	7	5	13	8	10	8	9	1	7	3	3	4	2	8	3	3
2229 LA LIBERTÉ	3	5	14	10	6	8	8	5	3	4	1	1	2	2	3	1

[SEG. REP]; §3; figure 2: Corpus HR8: Tableau des 9 formes les plus fréquentes (DE; LA; LES.....) avec deux fragments du tableau des segments répétés (DE LA LIBERTÉ ET DE L'ÉGALITÉ.....)

Deux voies distinctes s'ouvrent alors dans l'exploitation de ces résultats. Nous avons déjà quelque peu exploré la première de ces voies, dans un article écrit en collaboration avec Pierre LAFON (6), qui consiste à partir des tableaux de segments répétés pour asseoir de nouvelles typologies en suivant la même démarche que lors de l'analyse des tableaux de formes simples. Nous avons alors obtenu des résultats sur ce même corpus HR8 qui nous avaient paru encourageants. En effet le critère de satisfaction que nous avions posé au départ d'une meilleure discrimination des deux groupes de journaux se trouvait nettement mieux réalisé dans les analyses faites à partir de ces tableaux de segments que dans des analyses précédentes basées sur les tableaux de formes simples.

Dans la présente étude nous tenterons de montrer l'intérêt qu'il y a, selon nous, à utiliser les tableaux de segments répétés pour éclairer les résultats obtenus à partir des tableaux de formes simples.

4 Inventaires de segments répétés et "mise en élément supplémentaire": La technique de "mise en élément supplémentaire" est bien connue. Dans le domaine de l'analyse des tableaux lexicométriques on utilise cette technique soit pour annuler l'influence, jugée perturbatrice, de quelques unes des parties sur la typologie d'ensemble, soit encore pour neutraliser les conséquences de l'emploi par certaines des parties d'un vocabulaire trop lié à un thème particulier. Dans cette dernière hypothèse on préférera souvent mettre en élément supplémentaire les quelques vocables responsables de cette hétérogénéité.

A l'issue de cette opération l'ensemble des formes (et parfois également l'ensemble des parties) se trouve partagé en deux sous-ensembles bien distincts :

1) l'ensemble des éléments dits "principaux" qui seuls serviront de base au calcul des axes factoriels, des valeurs propres relatives à ces axes et des coordonnées factorielles.

(6) Voir : P. Lafon, A. Salem, article paru dans M.O.T.S. n° 6.

2) l'ensemble des éléments dits, pour cette raison, "supplémentaires" qui ne participeront en rien aux calculs mentionnés plus haut et pour lesquels on calculera après coup et indépendamment pour chaque élément, les coordonnées factorielles qui auraient été affectées à une forme ayant la même répartition dans les parties du corpus mais participant à l'analyse avec une masse infinitésimale.

Recommençons maintenant l'analyse décrite plus haut en impliquant cette fois en qualité d'éléments supplémentaires les 2945 lignes qui correspondent à la répartition de chacun des segments répétés dans les seize parties du corpus.

Nous soumettrons donc à l'A.F.C. un tableau qui compte cette fois (1848 formes répétées + 2945 segments répétés) 4793 lignes et toujours seize colonnes. Les 1848 premières lignes de ce tableau constitueront l'ensemble des éléments principaux, les autres celui des éléments supplémentaires.

Pour les raisons que nous avons évoquées plus haut, les valeurs propres, les axes factoriels, les coordonnées factorielles sur l'ensemble des éléments principaux sont exactement les mêmes que lors de la première analyse.

Nous examinerons donc directement les résultats relatifs à l'ensemble des éléments supplémentaires. Prenons par exemple les calculs supplémentaires effectués pour le premier facteur à propos du segment:

	F1	COS2	CTR
LA CONVENTION NATIONALE	1212	745	283

qui compte huit occurrences dont on peut voir la répartition dans les seize journaux au tableau 3, (figure 5)

La coordonnée sur le premier facteur est celle qui serait échue à toute forme ayant le même profil de répartition dans les différentes parties. Dans notre cas, c'est à peu de chose près, la coordonnée qui aurait été affectée au syntagme nominal LA CONVENTION NATIONALE si, toutes les choses restant égales par ailleurs, nous avions considéré lors du découpage initial que ce syntagme constituait une unité insécable, c'est-à-dire une forme.

Que l'on s'appuie sur ce genre de raisonnement ou que l'on préfère dire que l'on étudie la répartition de celles des occurrences de la forme LA qui sont suivies de CONVENTION NATIONALE importe assez peu en définitive. Il reste que, dans les analyses faites à partir de Tableaux Lexicométriques, les coordonnées factorielles calculées sur l'ensemble des segments répétés mis en supplémentaires constituent un matériel très homogène aux résultats des calculs correspondants sur les éléments principaux.

On pourra donc raisonnablement comparer entre elles les contributions aux facteurs des éléments principaux et supplémentaires qui résultent d'opérations similaires portant sur les facteurs et sur les poids relatifs.

FACTEUR NUMERO 1 NUM FORME	F1	COS2	CTR																	
+ 2240 IL FAUT	-523	246	270	41	1	1	0	1	1	2	3	0								
					3	2	9	1	5	9	1	2								
77 ETRE	-534	413	268	39	4	3	0	3	0	0	0	1								
					2	3	4	6	7	1	2	3								
98 ME	-557	110	217	29	3	0	3	0	0	0	0	0								
					11	0	5	5	0	0	0	2								
114 VOIR	-611	450	225	25	1	1	0	1	1	0	0	1								
					0	3	1	4	2	2	3	5								
27 TOUS	-646	814	1520	151	6	10	1	3	0	3	5	1								
					15	15	18	8	15	14	22	15								
25 NOUS	-655	338	1698	34	1	12	7	2	2	3	1	2								
					11	12	6	14	5	29	46	11								
+ 2226 TOUS LES	-667	727	1161	108	4	9	1	1	0	3	2	0								
					12	12	14	3	12	9	14	12								
+ 2286 POUR LES	-691	527	219	19	1	1	0	0	0	1	0	0								
					2	3	1	2	1	2	1	4								
167 VA	-729	469	218	17	2	0	0	0	0	0	0	1								
					0	1	3	3	1	3	1	2								
67 SANS-CULOTTES	-739	740	645	49	2	2	1	0	1	0	0	0								
					5	3	5	6	9	5	4	6								
139 BRAVES	-776	410	276	19	0	1	0	1	0	0	0	0								
					0	2	0	3	2	3	2	5								
+ 926 IL FAUT QUE	-814	186	224	14	0	0	0	0	0	1	0	0								
					1	0	4	1	0	6	0	1								
123 PERE	-839	302	391	23	0	0	1	0	0	0	0	0								
					7	1	4	1	1	1	6	1								
+ 2322 A ETE	-857	170	248	14	0	0	0	0	0	0	1	0								
					1	0	7	4	0	0	0	1								
+ 2290 LA SANS-CULOTTIERE	-858	495	319	18	0	0	0	0	0	1	0	0								
					1	4	3	2	1	3	3	0								
153 SANS-CULOTTIERE	-858	495	319	18	0	0	0	0	0	1	0	0								
					1	4	3	2	1	3	3	0								
159 DUCHESNE	-889	254	324	17	0	0	0	0	0	0	0	0								
					7	1	4	1	1	1	1	1								
+ 2298 PERE DUCHESNE	-889	254	324	17	0	0	0	0	0	0	0	0								
					7	1	4	1	1	1	1	1								
156 BONS	-903	479	334	17	1	0	0	0	0	0	0	0								
					0	2	1	2	3	2	5	1								
249 BOUGRES	-923	492	226	11	0	0	0	0	0	0	0	0								
					1	0	2	1	3	1	1	2								
+ 2307 LES SANS-CULOTTES	-925	585	330	16	0	0	0	0	0	0	0	0								
					1	2	2	2	4	1	1	3								
168 ALORS	-926	264	331	16	1	0	0	0	0	0	0	0								
					0	0	7	1	4	1	1	1								
+ 2377 TOUS CES	-928	428	229	11	0	0	0	0	0	0	0	0								
					7	0	1	2	1	0	3	2								
+ 2374 QUE TOUS	-939	296	234	11	0	0	0	0	0	0	0	0								
					3	3	1	0	0	0	3	1								
271 BON	-941	476	214	10	0	0	0	0	0	0	0	0								
					0	2	1	2	2	2	0	1								
41 FOUTRE	-942	852	1991	93	0	0	0	0	0	0	0	0								
					10	14	12	12	13	8	13	11								
274 CHIEN	-965	437	225	10	0	0	0	0	0	0	0	0								
					1	1	3	2	1	2	0	0								
193 BOUGRES	-982	517	326	14	0	0	0	0	0	0	0	0								
					0	1	2	3	2	1	4	1								

Figure 4 : Les formes les plus caractéristiques de HEBERT sur le premier facteur.

FACTEUR NUMERO 1 NUM FORME	F1	COS2	CTR	ET- REP- R01 H01	R01 à R08 H08	R08 à R08 H08	
+ 2790 CONTRE PARIS	1484	147	212	4	0 0 0 0	4 0 0 0	} R } H
570 FARINES	1484	147	265	5	0 0 0 0	5 0 0 0	} R } H
710 GRAINS	1484	147	212	4	0 0 0 0	4 0 0 0	
732 NIERAS	1484	147	212	4	0 0 0 0	4 0 0 0	
436 NIERAS TU QUE TU	1484	147	212	4	0 0 0 0	4 0 0 0	
+ 2954 PLAN DE	1484	147	212	4	0 0 0 0	4 0 0 0	
+ 966 QUAND TU AS	1484	147	372	7	0 0 0 0	7 0 0 0	
+ 1033 VOULU AFFAMER PARIS	1484	147	265	5	0 0 0 0	5 0 0 0	
561 DOUTER	1385	193	231	5	0 0 0 0	4 0 0 0	
+ 2662 DES SUBSISTANCES	1384	195	231	5	0 0 0 1	4 0 0 0	
390 PLAN	1373	222	363	8	0 0 1 1	6 0 0 0	
+ 991 QUE TU AS	1366	363	270	6	0 0 1 0	3 2 0 0	
515 PETION	1337	348	259	6	0 0 0 0	2 1 3 0	
+ 2719 ON TE	1327	122	212	5	0 0 0 0	0 5 0 0	
319 FACTION	1281	390	356	9	0 0 1 0	1 3 4 0	
+ 2589 L'ESCLAVAGE	1263	312	231	6	0 0 3 0	2 1 0 0	
423 ESCLAVAGE	1260	416	268	7	0 0 3 0	2 1 1 0	
+ 2514 LA FACTION	1257	314	267	7	0 0 1 0	0 3 3 0	
+ 924 DE LA PATRIE	1252	414	605	16	0 0 6 0	4 1 5 0	
+ 934 DANS LES DEPARTEMENTS	1251	455	378	10	0 0 3 1	4 1 0 1	
+ 2605 LES DEPUTES	1240	225	222	6	0 0 0 1	1 0 4 0	
469 AH	1225	601	217	6	0 0 0 1	1 2 1 1	
296 ROIS	1221	546	360	10	0 0 3 1	1 3 2 0	
+ 2503 DES ROIS	1220	398	251	7	0 0 1 1	0 3 2 0	
+ 971 C'EST PAR	1214	189	213	6	0 0 0 0	0 4 0 2	
525 TERRIBLE	1213	216	213	6	0 0 0 1	0 1 4 0	
+ 952 LA CONVENTION NATIONALE	1212	745	283	8	0 0 2 1	2 1 1 1	
+ 399 ENNEMIS DE LA PATRIE	1205	175	245	7	0 0 5 0	2 0 0 0	

Figure 3 : Les formes les plus caractéristiques de Jacques ROUX sur le premier facteur.

5 De l'inventaire statistique des formes, à la recherche des associations lexicales dans le discours : Pour toutes ces raisons, nous avons choisi lors de l'exploitation automatique des résultats de sélectionner les formes et les segments répétés à partir d'une même valeur de contribution au facteur.

Pour cet exemple nous avons choisi de fixer ce seuil à dix fois la contribution moyenne. Sur les figures 3 et 4 on voit, triés selon les valeurs décroissantes du premier facteur, les résultats correspondants aux valeurs extrêmes du premier facteur. Comme toujours, les valeurs positives de ce facteur sont associées à un emploi plus massif dans les journaux dus à Jacques ROUX.

Sur ce tableau les segments répétés sont précédés du signe "+" qui rappelle qu'ils n'ont participé à l'analyse qu'en tant qu'éléments supplémentaires. Le numéro qui suit ce signe permet de retrouver segments et formes dans les tableaux de départ.

On trouve ensuite dans l'ordre traditionnel :

- F1 coordonnée sur le premier facteur,
- COS2 contribution du facteur à l'élément,
- CTR contribution de l'élément au facteur,

puis :

- FT fréquence totale du segment ou de la forme dans le corpus.

La partie gauche de la figure restitue les fréquences dans chacune des 16 parties qui composent le tableau de départ.

Ces listes constituent indéniablement un progrès par rapport aux dépouillements précédents. Cependant, comme on l'a dit au § 3 l'examen des associations récurrentes réalisées à partir des formes graphiques renvoie à des niveaux d'analyse textuelle très différents : morphologie lexicale, syntaxe, voire rhétorique.

Dans la description qui suit, on s'est borné à classer ces associations d'après leurs caractères formels sur le seul plan des fréquences.

On retiendra quelques catégories formelles qui permettent d'affiner la description dans trois directions principales :

- a) une documentation plus précise sur les expansions invariantes.

Considérons par exemple sur la figure 3 les lignes :

```

732   NIERAS   ...
et   + 436   NIERAS TU QUE TU ...

```

Comme il est facile de le voir en se reportant aux données de départ, la forme NIERAS apparaît quatre fois dans le texte R5 de même que le segment NIERAS TU QUE TU qui contient cette forme. On en conclut aisément que la forme NIERAS n'apparaît dans le discours qu'insérée dans le contexte court ci-dessus qui constitue pour elle une expansion invariante.

On a éliminé de ces listes les segments intermédiaires NIERAS TU et NIERAS TU QUE qui n'apportent qu'une information redondante. Dans l'avenir on aura sans doute intérêt à présenter les informations que l'on décidera de conserver sous une forme plus concise :

NIERAS (dans NIERAS TU QUE TU) 4 fois indiquant à la fois la répartition de la forme et le contexte qui l'accompagne invariablement.

b) Des expansions "quasi-invariantes".

On trouve aussi des associations syntagmatiques qui, sans être aussi figées que celle décrite ci-dessus, sont à considérer comme des associations très privilégiées, expansions "quasi-invariantes" pour une forme.

323 ESCLAVAGE est à une exception près figé
dans + 924 L ESCLAVAGE qui a six occurrences.

Sept des seize occurrences du segment :

+ 924 DE LA PATRIE sont prises dans le segment
+ 399 ENNEMIS DE LA PATRIE.

On voit sur ces deux exemples l'intérêt qu'il peut y avoir à rassembler ces associations "presque figées" par des méthodes comme celle décrite dans cette étude.

c) Mise en évidence d'associations lexicales propres à l'un des deux auteurs, réalisées à partir des formes employées indifféremment par chacun d'eux.

Le segment :

+ 934 DANS LES DEPARTEMENTS employé dix fois par Jacques ROUX n'est jamais employé par HEBERT. Il est à noter qu'aucune des trois formes qui composent ce segment ne présente ce genre de distorsion dans l'utilisation qui en est faite par les deux auteurs. On voit sur la figure 6 les fréquences respectives de chacune de ces formes chez les deux auteurs.

Nous avons vu que TOUS était sur-employé par le "Père Duchesne". La méthode employée nous désigne TOUS LES et surtout TOUS CES comme encore plus caractéristiques de ce discours. Ces indications nous amèneront à remarquer en nous appuyant sur les I.S.R. (7) que, dans ces journaux, la forme TOUS sert plus souvent à amalgamer les ennemis du "Père Duchesne" qu'à rassembler ses amis autour de lui comme nous aurions pu le penser au premier abord. Chez HEBERT la forme TOUS renvoie presque toujours aux ennemis du Père DUCHESNE, même lorsqu'il met en scène ceux qu'il combat comme c'est le cas dans le contexte numéro trois de la figure 5.

Toutes ces hypothèses faites à partir d'un petit corpus devront, bien sûr, faire l'objet de vérifications sur un échantillon plus grand.

⚡ Ainsi, la méthode présentée permet d'obtenir des résultats relatifs aux segments, contextes courts des formes dont la répartition est mise en lumière par l'Analyse Factorielle des Correspondances d'un tableau lexicométrique.

Ce surcroît d'information obtenu sans hypothèses supplémentaires, peut faciliter grandement le commentaire des listes de formes sélectionnées par la méthode.

Enfin la démarche attire l'attention du chercheur sur des associations syntagmatiques, récurrentes dans certaines parties du corpus, réalisées à partir de formes dont la répartition individuelles n'appelle, quant à elle, aucun commentaire sur le plan quantitatif.

Nous espérons que ce travail intéressera les chercheurs qui tentent de reconstruire, par des méthodes automatisées, une partie des liens syntagmatiques ignorés lors des études quantitatives basées sur des vocables isolés de leur contexte.

(7) Parmi les expansions récurrentes de cette forme on trouve : TOUS LES TRAITRES - TOUS LES BRIGANDS - TOUS LES ACCAPAREURS - TOUS LES HOMMES SUSPECTS - TOUS LES COUPS DE CHIENS etc. .

REF. 32657026 ... Il ne joue pas des jambes du côté de la <Vendée> ou de l'<Angleterre> . Nous lui verrons essayer à son tour le collier de <Charlotte Corday>."Ainsi. Foutre. Quand nous voudrons donner le moindre signe de vie. Tous nos ennemis seront -à quia". Qu'on mette le grain sur

tous

les contre-révolutionnaires, que tous les feuillants, royalistes, aristocrates, accapareurs soient mis à l'ombre, qu'ils soient enfermés dans des églises, et que l'on braque vis-à-vis des canons chargés à mitraille, jusqu'à ce que la paix soit faite et la constitution établie. Voilà, foutre, les moyens...

REF. 32714033 ... serruriers : ces arrangements ne conviendront pas à nos petites maîtresses et à ces godelureaux à culottes étroites et aux habits carrés ; mais une république ne peut exister avec une pareille canaille, et j'espère qu'un de ces matins les <sans-culottes> s'armeront de fouets pour foutre la chasse à

tous

ces courtauds de boutique, à tous ces clériaux, à tous ces bandits du <Palais Royal>, qui tirent la langue sur les patriotes, et qui ne cessent de faire de l'esprit à perte de vue contre la République. Nous ne voulons souffrir parmi nous que des hommes utiles, des bougres...

REF. 32705021 ... ; de faux patriotes, qui s'entendent avec nous, brouilleront les cartes à la Convention ; la République sera déchirée de toutes parts ; elle ressemblera à tour de <Babel> ; personne ne s'entendra que nous ; les patriotes s'armeront les uns contre les autres, se mangeront le blanc des yeux ;

tous

les mécontents se joindront à nous ; enfin nous deviendront les plus forts ; c'est alors que nous ferons nos prouesses, et que les <Jacobins> apprendront de quel bois nous nous chauffons". Voilà pauvres sans-culottes, qui ne voyez pas plus loin que votre nez toutes les manigances de vos ennemis. Ils...

Figure 5 : Trois contextes de la forme TOUS.

	ROUX	HEBERT	TOTAL
DANS	80	84	164
LES	445	431	876
DEPARTEMENTS	16	19	35

Figure 6 : Répartition des trois formes qui composent le segment
DANS LES DEPARTEMENTS.

BIBLIOGRAPHIE

- Pierre LAFON, André SALEM, "L'inventaire des segments répétés d'un texte" in M.O.T.S. 6 mars 1983. pp 161-177.
- Jacques GUILHAUMOU, "Nous, vous, tous... La Fête de l'Union du 10 août 1793" à paraître dans M.O.T.S. n° 10.
- Maurice TOURNIER, "Un vocabulaire ouvrier en 1848. Essai de lexicométrie" St-Cloud 1975.
- Annie GEFFROY, "Personnes du discours et figures du pouvoir dans l'Ami du peuple de Leclerc" in Actes du 2-ème colloque de lexicologie politique. St-Cloud, octobre 1980.
- André Salem, "Analyse factorielle et lexicométrie synthèse de quelques expériences" M.O.T.S. n° 4 mars 1982.