QUATRIEME PARTIE

LA STRUCTURE LEXICALE

La structure lexicale est définie comme l'ensemble des rapports qui lient entre elles les fréquences indépendamment des mots eux-mêmes.

Pour ne pas déroger à la règle de la terminologie devenue classique en lexicométrie, nous appellerons **N** le nombre des occurrences du corpus (sa longueur), **V** le nombre de ses vocables, c'est-à-dire de ses mots différents, et $\mathbf{V_1}$ l'effectif des *hapax* (mots employés une seule fois). **N** est aussi appelé l'**étendue du corpus**, et **V** l'**étendue de son vocabulaire**.

Ces données lexicométriques sont très utiles pour présenter la structure quantitative du corpus et en étudier par la suite, la *richesse lexicale*, l'*accroissement du vocabulaire*, la *connexion lexicale*, les catégories lexicales, etc. Ils serviront de base à tous les calculs ultérieurs.

Chapitre 9

Les caractéristiques lexicométriques

1. Les principales caractéristiques lexicométriques du corpus

Les principales caractéristiques lexicométriques que nous livre le dépouillement de notre corpus présentent les données quantitatives suivantes :

L'étendue globale du corpus **N** atteint 62 240 occurrences. Cette étendue globale est composée de l'étendue concernant les noms communs qui sont au nombre de 61 177 occurrences (ou items) et de l'étendue concernant les noms propres qui atteignent, eux, 1 063 occurrences. Étant donné que les analyses statistiques des noms propres ont été effectuées à part, l'étendue de notre corpus **N** est confondue à celle des noms communs, c'est-à-dire que **N** est égal à 61 177.

L'étendue globale du vocabulaire **V**, quant à elle, atteint 7 079 vocables (c'est-à-dire des mots différents) se répartissant en 6 652 vocables correspondant à des noms communs, et 427 vocables correspondant à des noms propres différents. Comme pour l'étendue du corpus, l'étendue du vocabulaire **V** que nous considérons est celle du vocabulaire commun, c'est-à-dire que **V** est égal à 6 652.

Le nombre total des *hapax* $\mathbf{V_1}$ (mots employés une seule fois) de notre corpus est de 3 749 *hapax* se répartissant en 3 443 noms communs et 306 noms propres ayant la fréquence 1. $\mathbf{V_1}$ correspond, bien entendu, pour nous au nombre des *hapax* des noms communs, c'est-à-dire que $\mathbf{V_1}$ est égale à 3 443. La fréquence maximale dans notre corpus est de 7 825 qui correspond à la fréquence de l'article défini ال « *al* ».

Les principales caractéristiques lexicométriques que nous venons d'exposer peuvent être résumées dans le tableau suivant :

Les principales caractéristiques lexicométriques d'*al-ÞimtâÝ wa l-muÞânasa*

	Vocabulaire commun	Noms propres	Total
Étendue du corpus :	**61 177**	1 063	62 240
Étendue du vocabulaire :	**6 652**	427	7 079
Fréquence maximale :	**7 825**	198	7 825
Nombre des *hapax* :	**3 443**	306	3 749

Tableau 55

Il est important de noter que ce qui est pris en compte dans les principales caractéristiques lexicométriques que nous venons de présenter c'est le lemme et non la forme (le mot-forme). En ce qui concerne le nombre des occurrences (l'étendue du corpus), il n'y a aucune différence entre le nombre des occurrences des formes et celui des occurrences des lemmes puisqu'en effet, à chaque occurrence de forme est attribuée une occurrence de lemme. Ce n'est pas le cas pour les vocables où l'on ne compte plus le nombre d'occurrences des formes ou des lemmes mais le nombre des formes différentes ou des lemmes différents. Et ce faisant, il y a une inadéquation entre le nombre des "vocables-formes" et celui des "vocables-lemmes" étant donné que plusieurs formes peuvent avoir un même lemme.

Tout comme le nombre des vocables, du fait du regroupement de plusieurs formes différentes sous un même lemme, le nombre des *hapax* se voit aussi diminuer en passant des formes aux lemmes.

Dans nos différents calculs et analyses, ce que nous avons retenu comme base des caractéristiques lexicométriques ce sont les occurrences des lemmes. Néanmoins, nous donnons à titre comparatif, dans le tableau suivant, les principales caractéristiques lexicométriques à la fois pour les formes et les lemmes :

Comparatif des principales caractéristiques lexicométriques d'*al-ÞimtâÝ wa l-muÞânasa* entre les formes et les lemmes

	Vocabulaire commun		Noms propres		Total	
	Formes	Lemmes	Formes	Lemmes	Formes	Lemmes
occurrences :	**61 177**	**61 177**	1 063	1 063	6 2240	62 240
vocables :	**10 174**	**6 652**	452	427	10 626	7 079
Fréquence maximale :	**7 825**	**7 825**	198	198	7 825	7 825
Nombre des *hapax* :	**6 763**	**3 443**	325	306	7 088	3 749

Tableau 56

2. Les principales caractéristiques lexicométriques des Nuits

L'étendue du corpus, nous l'avons vu, mesure 61 177 occurrences. Cette longueur est celle des quinze nuits réunies ; mais ces occurrences ne sont évidemment pas réparties identiquement entre les nuits. L'étendue de celles-ci est variable d'une nuit à une autre passant de 906 à 10 788 occurrences. Nous observons donc un écart entre la plus grande et la plus petite étendue de 10 788 – 906 = 9 882 occurrences.

Les nuits les plus courtes, c'est-à-dire celles qui ont les plus petites étendues, sont la cinquième nuit avec 906 occurrences (c'est l'étendue minimum) puis les nuits 16, 15 et 3 avec respectivement 1 286, 1 773 et 2 004 occurrences.

Les nuits les plus longues, c'est-à-dire celles qui ont les plus grandes étendues, sont la huitième nuit avec 10 788 occurrences (c'est l'étendue maximum) puis les nuits 10, 6 et le préambule avec respectivement 9 564, 7 079 et 5 062 occurrences. La figure suivante montre bien la variation de l'étendue de chacune des quinze nuits :

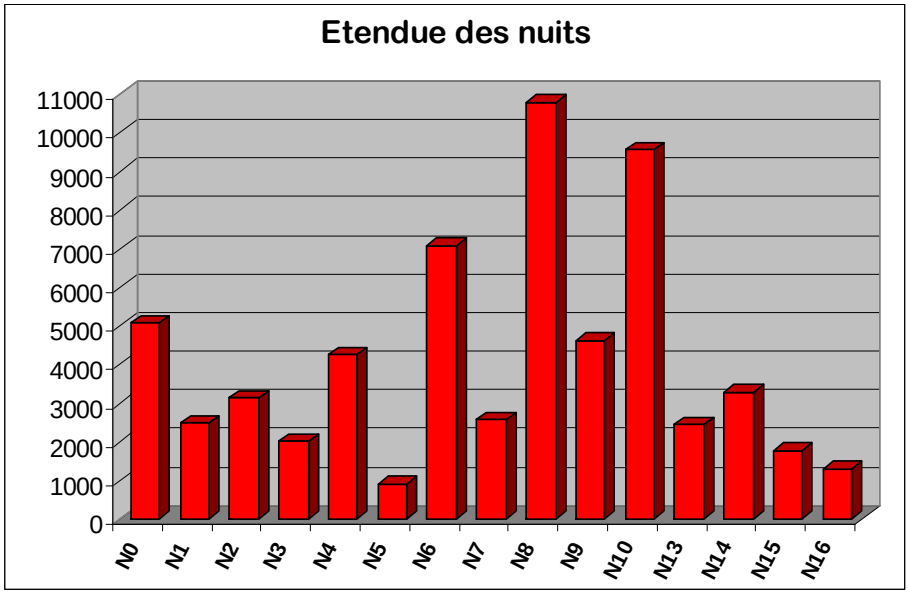


Figure 96

L'étendue moyenne est de $\frac{61177}{15} = 4\,078$ occurrences. Autour de cette moyenne, la distribution est asymétrique. En effet, seulement six nuits ont des étendues supérieures à l'étendue moyenne dont une, la quatrième nuit, est légèrement au-dessus de celle-ci alors que neuf nuits s'en écartent négativement ayant des étendues bien inférieures à cette même étendue moyenne. La figure suivante est beaucoup plus parlante quant à cette distribution autour de l'étendue moyenne :

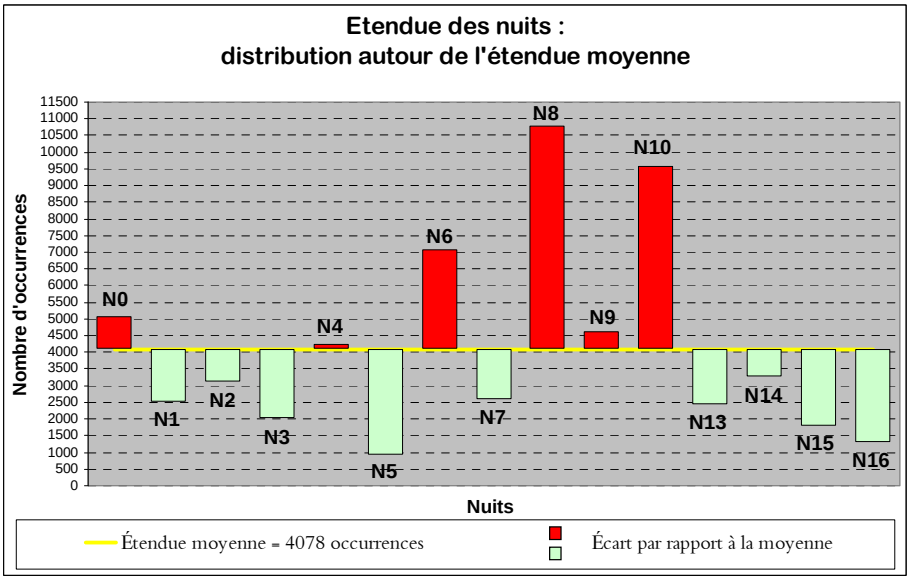


Figure 97

Il est à noter que les deux nuits qui ont, au niveau des occurrences, les valeurs extrêmes, la nuit 5 avec l'étendue minimale et la nuit 8 avec l'étendue maximale, occupent aussi au niveau du vocabulaire les mêmes places extrêmes. La cinquième nuit présente la plus petite étendue du vocabulaire qui est de 340 vocables alors que la huitième nuit a la plus grande étendue du vocabulaire avec 1 967 vocables. Il n'est pas sans intérêt de noter ici que la variation de l'étendue des parties d'un corpus n'est pas toujours conforme à la variation de l'étendue du vocabulaire de ces mêmes parties. En revanche, il pourrait y avoir entre les deux une certaine corrélation ; c'est ce que nous tenterons de démontrer un peu plus loin.

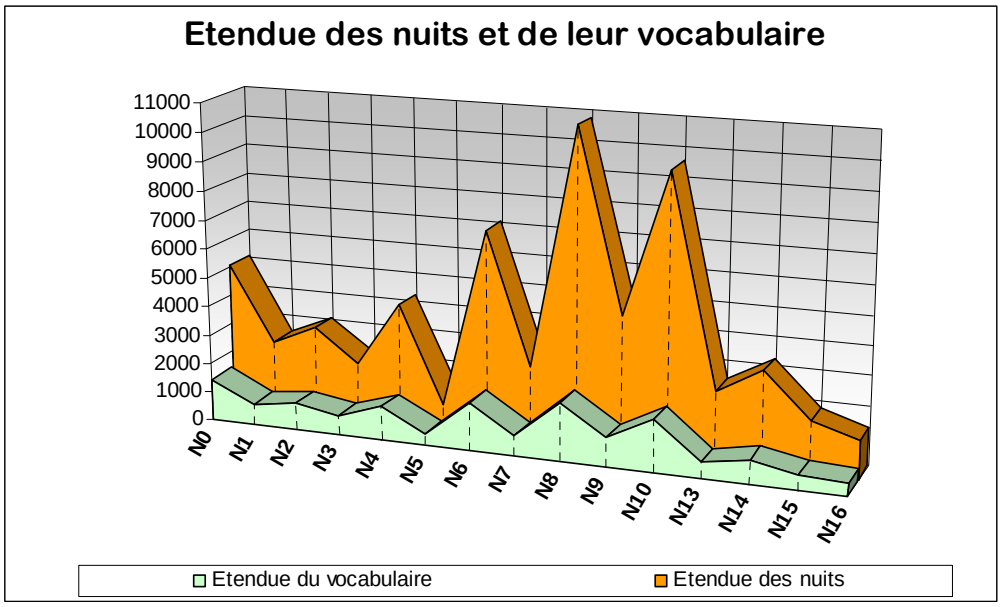


Figure 98

Au niveau des *hapax* aussi, la cinquième nuit présente le plus petit nombre de *hapax* qui est de 265 formes de fréquence 1 et la huitième nuit présente, elle, le plus grand nombre de *hapax* avec 1 249 formes de fréquence 1. Pareillement, au niveau de la fréquence maximale, c'est bien la cinquième nuit qui a la plus petite fréquence maximale de toutes les nuits qui est de 107 (correspondant au coordonnant و « *wa* ») alors que la huitième nuit a la plus grande fréquence maximale qui est de 1 317 (correspondant à l'article défini ال « *al* »).

Les principales caractéristiques lexicométriques de toutes les nuits sont présentées dans le tableau suivant :

Les principales caractéristiques lexicométriques des nuits dans *al-ÞimtâÝ wa l-muÞânasa*

	Nbr occurrences	Nbr vocables	Nbr hapax	Fréq. Max	Vocable
N00	5 062	1 412	1 048	691	و
N01	2 478	703	498	286	ال
N02	3 115	896	649	388	ال
N03	2 004	627	470	218	و
N04	4 248	1 133	834	498	ال
N05	906	340	265	107	و
N06	7 079	1 644	1 140	945	و
N07	2 569	688	464	372	ال
N08	10 788	1 967	1 249	1 317	ال
N09	4 607	1 003	637	777	ال
N10	9 564	1 772	1 033	1 229	ال
N13	2 427	535	342	343	ال
N14	3 271	819	546	477	ال
N15	1 773	510	344	248	ال
N16	1 286	420	313	187	ال

Tableau 57

Étant donné que l'étendue du vocabulaire d'un corpus est fonction de l'étendue de ce dernier, les valeurs de V ne sont pas directement comparables entre elles. Les rapports entre étendue du vocabulaire et étendue du corpus seront étudiés en détail plus loin en étudiant la richesse lexicale. En revanche, l'observation des données quantitatives peut nous permettre de dégager, d'ores et déjà, certaines tendances et considérations générales.

En effet, nous pouvons, par exemple, voir comment sont classées les nuits selon les valeurs de V et de N. Il faut savoir que les classements des parties d'un corpus ne sont pas semblables selon qu'ils sont basés sur N ou sur V.

Il existe cependant, un coefficient qui permet de mesurer globalement l'ampleur des différences entre ces types de classements : il s'agit du coefficient de corrélation des rangs de Spearman.

Le coefficient de corrélation des rangs de Spearman est une mesure de corrélation non paramétrique qui sert à déterminer la relation qui existe entre deux séries de données.

Ce coefficient est souvent utilisé comme test statistique afin de déterminer s'il existe une relation entre deux variables aléatoires. Il est donc utilisé comme un test d'hypothèse en vue d'étudier la dépendance entre deux variables aléatoires. Généralement désigné par ρ, le coefficient de corrélation des rangs est défini par :

$$\rho = 1 - \frac{6\sum_{i=1}^{n} d_i^{\,2}}{n\left(n^2 - 1\right)}$$

où

$d_i = R_{x_i} - R_{y_i}$
R_{x_i} = le rang que prend la valeur X_i
R_{y_i} = le rang que prend la valeur Y_i
n = le nombre de valeurs des variables

Calcul du coefficient de corrélation pour les 15 nuits d'*al-ÞImtâÝ wa l-muÞânasa* :

Après avoir classé les 15 nuits par ordre croissant de leur étendue, nous obtenons un classement selon N que nous présentons dans la colonne « Rang (N) » du tableau 4. Pareillement, après avoir classé les 15 nuits par ordre croissant de l'étendue de leur vocabulaire, nous obtenons un autre classement, selon V cette fois-ci, que nous portons dans la colonne « Rang (V) ». Nous calculons ensuite le carré de la distance entre chaque couple de valeurs pour calculer enfin la somme des carrés des distances.

Nuits	Rang (N)	Rang (V)	d^2
N00	12	12	0
N01	6	7	1
N02	8	9	1
N03	4	5	1
N04	10	11	1
N05	1	1	0
N06	13	13	0
N07	7	6	1
N08	15	15	0
N09	11	10	1
N10	14	14	0
N13	5	4	1
N14	9	8	1
N15	3	3	0
N16	2	2	0
			$\Sigma d^2 = 8$

Tableau 58

Nous calculons ensuite le coefficient de corrélation des rangs de Spearman :

$$\rho = 1 - \frac{6x8}{15 \times (225 - 1)} = 1 - \frac{48}{3360} = 1 - 0{,}0143 = \mathbf{0{,}9857}$$

Au seuil de signification α = 0,05, on peut rejeter l'hypothèse nulle d'absence de corrélation, autrement dit, il y a une corrélation positive et très significative, presque parfaite, entre le rang de l'étendue des nuits et de celui de l'étendue de leur vocabulaire, ceci est représenté dans le graphique suivant :

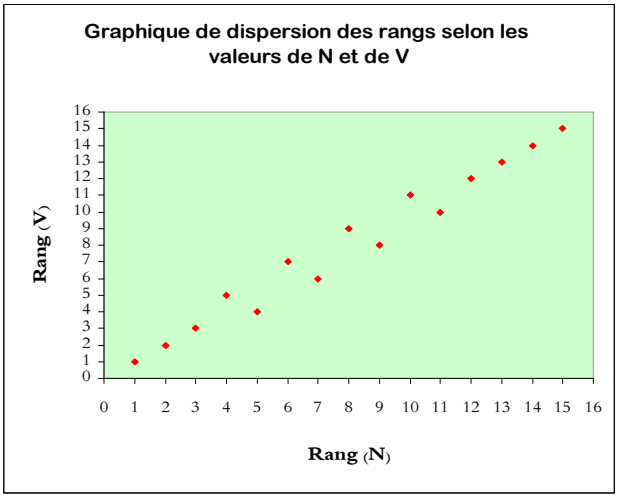


Figure 99

Une autre tendance que l'on peut mettre en évidence c'est la corrélation qui peut exister entre la longueur des nuits et l'étendue de leur vocabulaire.

En effet, la représentation graphique suivante (figure 5) permet de révéler l'existence d'une tendance générale de corrélation entre l'étendue des nuits (leur longueur) et l'étendue de leur vocabulaire. En abscisse sont portés les valeurs de N, en ordonnée celles de V. Il est clair que les points dont l'abscisse et l'ordonnée correspondent aux valeurs de chaque nuit forment un nuage de points orienté du coin inférieur gauche vers le coin supérieur droit du graphique : ce qui se traduit par l'existence d'une corrélation entre les valeurs de N et celles de V.

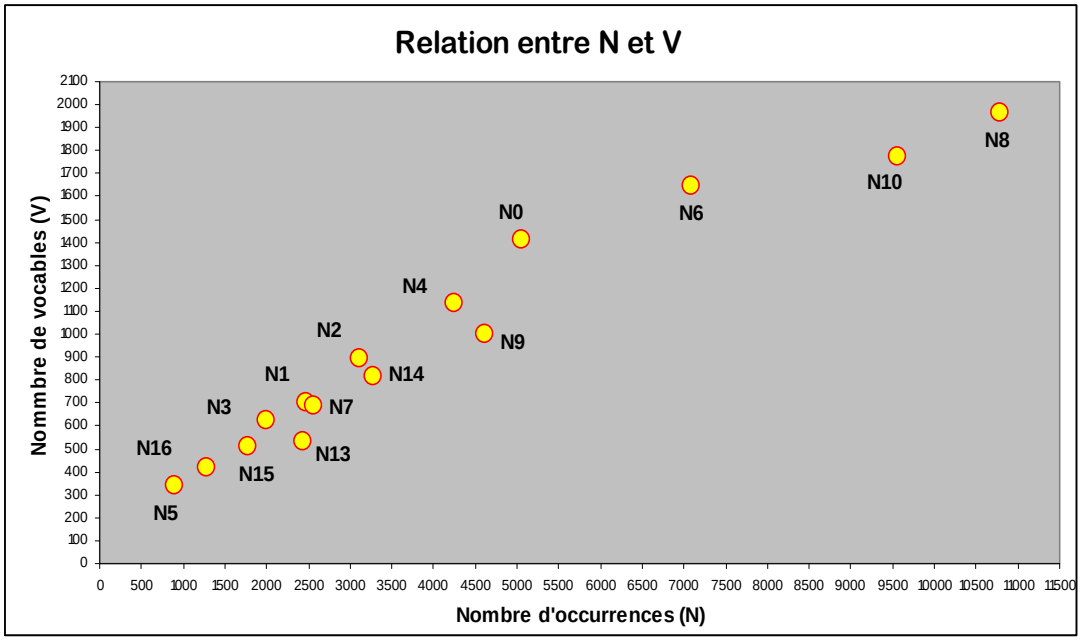


Figure 100

Nous présentons enfin dans le tableau de la page suivante (tableau 5) pour chaque nuit : son étendue, son rang dans l'ordre décroissant du nombre d'occurrences, les occurrences cumulées, sa moyenne théorique, son écart par rapport à la moyenne arithmétique, sa fréquence relative, les fréquences cumulées et son coefficient de variation qui est égal à l'écart-type divisé par la moyenne.

Aussi, avons-nous résumé dans ce même tableau des mesures de trois types :

- Des mesures de tendance centrale à savoir :
  - la moyenne arithmétique (= 4 078,47)
  - la médiane (= 3 115)
  - le mode qui est ici l'étendue maximale (=10 788)
- Des mesures de dispersion à savoir :
  - la variance (= 8 715 438,552)
  - l'écart-type (= 2 952,19)
  - le coefficient de variation (=0,72)
- Des mesures de position qui sont les quartiles (le 1er quartile = 2 004, le 3ème quartile = 5 062, le 2ème quartile étant la médiane = 3 115).

La description statistique des données quantitatives relatives aux nuits d'*al-ÞimtâÝ wa l-muÞânasa*

	Occ.	Rang	Occ. Cum.	Moyenne	Écart / moy.	freq. rel.	Freq. cum.	Coef. de var.
N00	5 062	4	5 062	5 062,00	983,53	8,27 %	8,27 %	0,58
N01	2 478	10	7 540	3 770,00	-1 600,47	4,05 %	12,32 %	0,78
N02	3 115	8	10 655	3 551,67	-963,47	5,09 %	17,42 %	0,83
N03	2 004	12	12 659	3 164,75	-2 074,47	3,28 %	20,69 %	0,93
N04	4 248	6	16 907	3 381,40	169,53	6,94 %	27,64 %	0,87
N05	906	15	17 813	2 968,83	-3 172,47	1,48 %	29,12 %	0,99
N06	7 079	3	24 892	3 556,00	3 000,53	11,57 %	40,69 %	0,83
N07	2 569	9	27 461	3 432,63	-1 509,47	4,20 %	44,89 %	0,86
N08	10 788	1	38 249	4 249,89	6 709,53	17,63 %	62,52 %	0,69
N09	4 607	5	42 856	4 285,60	528,53	7,53 %	70,05 %	0,69
N10	9 564	2	52 420	4 765,45	5 485,53	15,63 %	85,69 %	0,62
N13	2 427	11	54 847	4 570,58	-1 651,47	3,97 %	89,65 %	0,65
N14	3 271	7	58 118	4 470,62	-807,47	5,35 %	95,00 %	0,66
N15	1 773	13	59 891	4 277,93	-2 305,47	2,90 %	97,90 %	0,69
N16	1 286	14	61 177	4 078,47	-2 792,47	2,10 %	100,00 %	0,72
	61 177							

Moyenne :	**4 078,47**
Minimum :	**906**
Maximum :	**10 788**
1er quartile :	**2 004**
Médiane :	**3 115**
3ème quartile :	**5 062**
Variance :	**8 715 438,552**
Ecart-type :	**2 952,19**
Coef. de variation :	**0,72**

Tableau 59

3. Intertexte[284] et répartition générique

Parce que directement repérables et facilement quantifiables, nous ne considérons ici des textes convoqués par *TawÎîdî* que les citations, qu'elles soient poétiques, coraniques ou prophétiques.

Comme chez la majorité des écrivains classiques, les sources majeures auxquelles font référence les citations de *TawÎîdî* sont le Coran, la Tradition prophétique (*Íadî*×) et la poésie :

3.1. Les principales caractéristiques lexicométriques de la partition "Intertexte"

Les principales caractéristiques lexicométriques qui résultent du dépouillement de la partition "Intertexte" nous livrent les données quantitatives suivantes :

L'étendue de la partie "Poésie" est de N = 1 096 occurrences, l'étendue de son vocabulaire est de V = 467 vocables. Les *hapax* sont au nombre de 358. L'article défini ال « *al* » est le lemme ayant la fréquence maximale avec 111 occurrences.

La partie "Coran" présente une étendue de N = 78 occurrences, un vocabulaire de V = 45 vocables. Les *hapax* y sont au nombre de 32. Comme pour la partie "Poésie", le lemme ayant la fréquence maximale est l'article défini ال « *al* » avec seulement 8 occurrences.

[284] Nous utilisons "intertexte" dans le sens d'un ensemble de textes ou de fragments textuels convoqués liés par des relations intertextuelles. Voir dans ce sens : D. Maigueneau, *Genèses du discours*, Mardaga, Liège, 1984. Voir aussi : P. Charaudeau et D. Maigueneau (sous la dir.), *Dictionnaire d'analyse du discours*, Seuil, Paris, 2002, Article « Intertexte », p. 327.

Quant à la partie "*Íadî*×", son étendue n'a que 11 occurrences, son vocabulaire atteint 10 vocables. Neuf de ces 10 vocables n'ont qu'une seule occurrence, ce sont les *hapax* de cette partie, et un seul vocable, la préposition ب « *bi* », a deux occurrences ayant, par là même, la fréquence maximale.

Les principales caractéristiques lexicométriques de la partition « Intertexte » dans *al-ÞimtâÝ wa l-muÞânasa*

	Nbr occurrences	Nbr vocables	Nbr hapax	Fréq. Max	Lemme
Poésie	1 096	467	358	111	ال
Coran	78	45	32	8	ال
***Íadî*×**	11	10	9	2	ب

Tableau 60

L'ensemble des trois parties de la partition "Intertexte" présentent une étendue de 1 185 occurrences ; rapportée à l'ensemble des occurrences du corpus, cette étendue représente 0,02 % de l'étendue totale d'*al-ÞImtâÝ wa l-MuÞânasa*. Le vocabulaire des trois parties est de 522 vocables représentant 0,08 % du vocabulaire total d'*al-ÞImtâÝ wa l-MuÞânasa*.

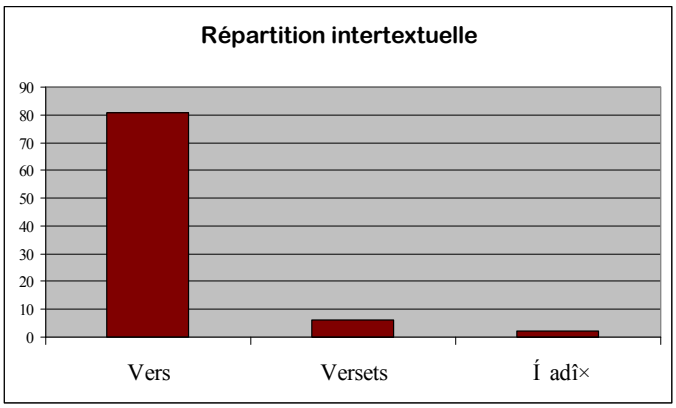


Figure 101

À l'intérieur de la partition "Intertexte", en termes de pourcentage, la poésie représente 92 % de l'ensemble des occurrences, le Coran en représente 7 %, alors que le

Íadî×, ne représente qu'un petit 1 % de l'ensemble des occurrences de la partition "Intertexte".

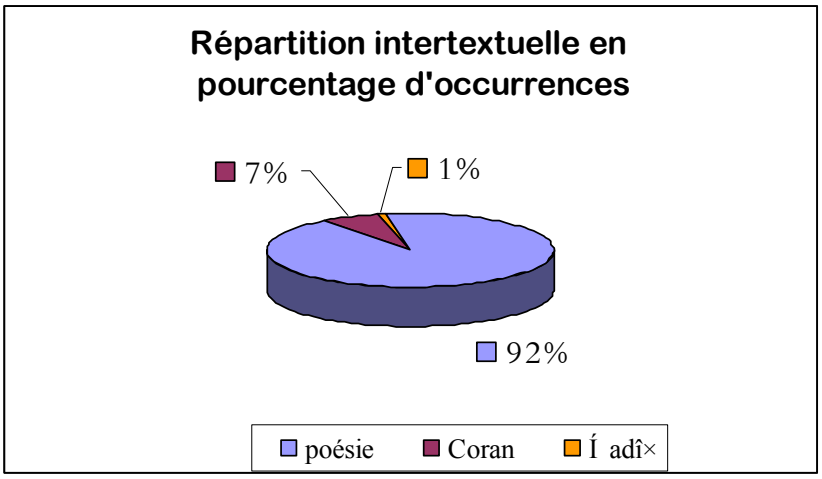


Figure 102

Après avoir présenté les principales caractéristiques lexicométriques de la partition "Intertexte" pour lesquelles seules les mots (occurrences, vocables, lemmes) sont considérés, voyons maintenant comment se répartissent les citations à un niveau supérieur au mot, les versets coraniques, les vers cités, les *Íadî×*.

3.1.1. Le Coran

Dans le préambule et les quatorze nuits composant le premier tome de notre corpus *al-ÞImtâÝ wa l-MuÞânasa*, Abû Íayyân at-TawÎîdî ne cite que 6 versets coraniques représentant, comme nous l'avons vu, 78 occurrences ou 45 vocables. Ce sont, dans l'ordre d'apparition dans le corpus, le verset 26 de la sourate 3 et le verset 55 de la sourate 43, dans la nuit 6, le verset 63 de la sourate 24, dans la nuit 7 et enfin le verset 103 de la sourate 35, le verset 46 de la sourate 3 et le verset 54 de la sourate 5, dans la nuit 8.

Nous résumons dans le tableau 7 de la page suivante toutes les citations coraniques d'*al-ÞImtâÝ wa l-MuÞânasa* avec indication de la nuit et de la page dans lesquelles chaque citation est apparue.

Les citations coraniques dans le corpus *al-ÞImtâÝ wa l-MuÞânasa*

Nuit	Page	Sourate / Verset	Verset (en gras souligné c'est la citation exacte)
N6	82	آل عمران الآية 26 **3** / **26**	**قُلِ اللَّهُمَّ مَالِكَ الْمُلْكِ تُؤْتِي الْمُلْكَ مَنْ تَشَاءُ وَتَنْزِعُ الْمُلْكَ مِمَّنْ تَشَاءُ وَتُعِزُّ مَنْ تَشَاءُ وَتُذِلُّ مَنْ تَشَاءُ بِيَدِكَ الْخَيْرُ إِنَّكَ عَلَى كُلِّ شَيْءٍ قَدِيرٌ**
N6	83	الزخرف الآية 55 **43** / **55**	**فَلَمَّا آسَفُونَا انْتَقَمْنَا** مِنْهُمْ فَأَغْرَقْنَاهُمْ أَجْمَعِينَ
N7	103	الفرقان الآية 63 **24** / **63**	وَعِبَادُ الرَّحْمَنِ الَّذِينَ يَمْشُونَ عَلَى الْأَرْضِ هَوْنًا **وَإِذَا خَاطَبَهُمُ الْجَاهِلُونَ قَالُوا سَلَامًا**
N8	118	الصافات الآية 103 **35** / **103**	**فَلَمَّا أَسْلَمَا وَتَلَّهُ لِلْجَبِينِ وَنَادَيْنَاهُ** أَنْ يَا إِبْرَاهِيمُ
N8	118	آل عمران الآية 46 **3** / **46**	**وَيُكَلِّمُ النَّاسَ فِي الْمَهْدِ وَكَهْلًا** وَمِنَ الصَّالِحِينَ
N8	131	المائدة الآية 54 **5** / **54**	يَا أَيُّهَا الَّذِينَ آمَنُوا مَنْ يَرْتَدَّ مِنْكُمْ عَنْ دِينِهِ فَسَوْفَ يَأْتِي اللَّهُ بِقَوْمٍ يُحِبُّهُمْ وَيُحِبُّونَهُ أَذِلَّةٍ عَلَى الْمُؤْمِنِينَ أَعِزَّةٍ عَلَى الْكَافِرِينَ يُجَاهِدُونَ فِي سَبِيلِ اللَّهِ وَلَا يَخَافُونَ لَوْمَةَ لَائِمٍ **ذَلِكَ فَضْلُ اللَّهِ يُؤْتِيهِ مَنْ يَشَاءُ** وَاللَّهُ وَاسِعٌ عَلِيمٌ

Tableau 61

Dans la figure 8, c'est la représentation graphique de la répartition des versets coraniques cités selon les nuits que nous présentons. On y voit les seules trois nuits qui ont fait apparaître les versets coraniques convoqués par Abû Íayyân at-TawÎîdî.

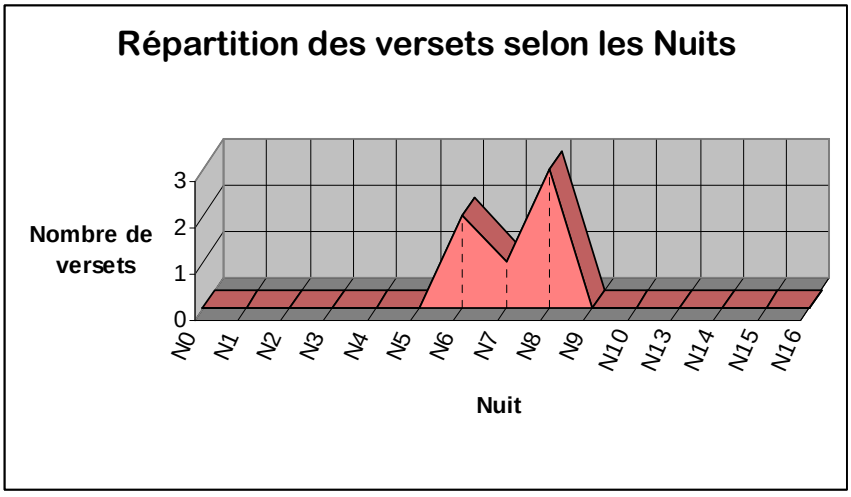


Figure 103

3.1.2. Le *Íadî*×

Dans tout notre corpus, il n'est question que de deux citations puisées dans la Tradition prophétique, *Íadî*×. Les deux citations de la Tradition convoquées par Abû Íayyân at-TawÎîdî regroupent, nous l'avons vu, 11 occurrences ou 10 vocables.

3.1.3. La poésie

Les 1 096 occurrences ou 467 vocables "poétiques" sont répartis sur 81 citations poétiques convoquées par Abû Íayyân at-TawÎîdî, nous trouvons dans ces 81 citations poétiques, 78 vers complets et 3 fragments composés d'un seul hémistiche. Ces citations apparaissent dans 43 endroits différents du corpus. La figure suivante représente graphiquement la répartition des vers cités selon les nuits.

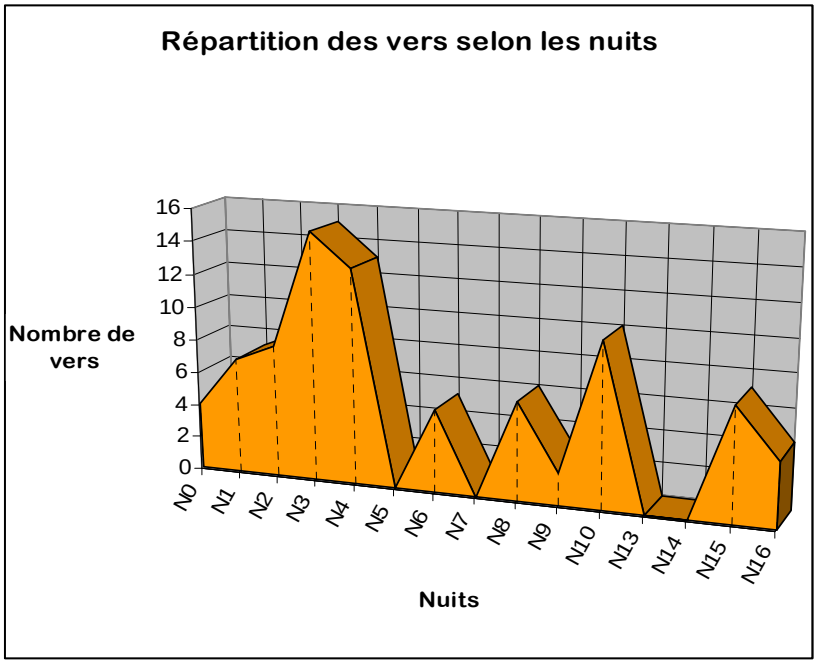


Figure 104

Il est clair que les vers ne sont pas répartis d'une façon uniforme entre les nuits : la nuit la plus riche en vers est la troisième nuit avec un effectif de 15 vers, la nuit la plus pauvre est la neuvième nuit avec 2 vers alors que quatre nuits sont à effectif nul c'est-à-dire qu'on n'y trouve aucun ver de poésie.

La moyenne des vers est de 5,40 vers soit 6 vers par Nuit. Dans le tableau suivant, nous présentons le nombre de vers dans chaque nuit, la moyenne des vers par nuit et l'écart de chaque nuit par rapport à cette moyenne.

Nuit	**N00**	**N01**	**N02**	**N03**	**N04**	**N05**	**N06**	**N07**	**N08**	**N09**	**N10**	**N13**	**N14**	**N15**	**N16**
Nbr de vers	4	7	8	15	13	0	5	0	6	2	10	0	0	7	4
Moyenne	**5,4** vers par Nuit														
Ecart / Moy.	- 1,4	1,6	2,6	9,6	7,6	- 5,4	- 0,4	- 5,4	0,6	- 3,4	4,6	- 5,4	- 5,4	1,6	- 1,4

Tableau 62

L'écart des vers de chaque nuit par rapport à la moyenne est représenté dans la figure suivante où l'on voit les deux nuits présentant un écart très important par rapport au nombre de vers moyen par nuit, qui sont les nuits 3 et 4. Il est clair aussi dans cette figure que les quatre nuits 5, 7, 13 et 14 accusent un écart négatif de la valeur de la moyenne (en valeur absolue), ce qui signifie l'absence totale de vers dans ces nuits.

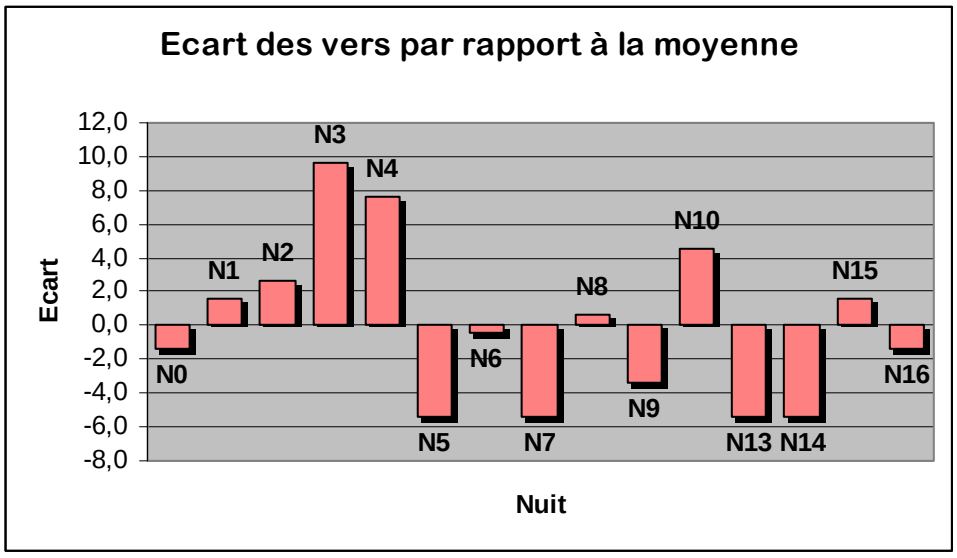
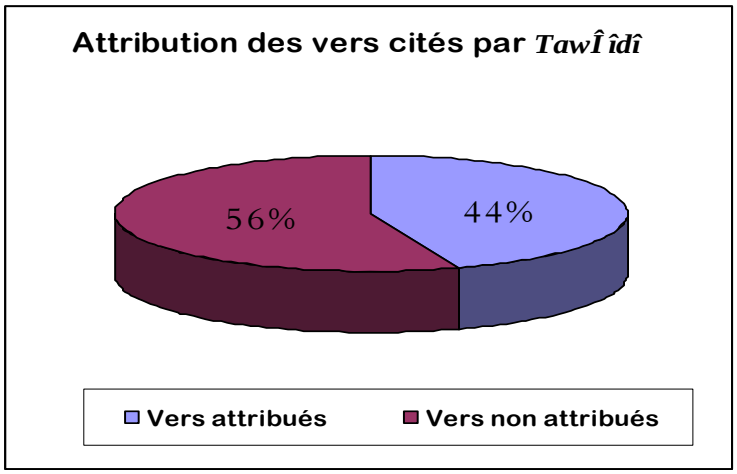


Figure 105

Les 81 vers cités par Abû Íayyân at-TawÎîdî ne sont pas tous attribués à des auteurs bien déterminés ; nous pouvons donc les répartir en deux catégories :

- Les vers attribués : sont au nombre de 36 soit 44% de l'ensemble des vers cités par *TawÎîdî*. Ces 36 vers sont attribuée à 16 poètes différents.
- Les vers non attribués : sont au nombre de 45 représentant 56%



Nous présentons dans le tableau suivant, les poètes auxquels Abû Íayyân at-TawÎîdî a attribué les 36 vers qu'il a cités, avec le nombre de vers pour chaque poète et le pourcentage des citations pour chacun d'entre eux.

Dans le cas de non attribution des vers, Abû Íayyân at-TawÎîdî utilise plusieurs formules avant de citer les vers en question : il peut utiliser le mot الأوّل « le premier », le mot الآخر « l'autre », le mot الشاعر « le poète », le mot أعرابيّ « un paysan » ou le mot قائل « On dit », comme il peut ne rien indiquer en citant directement les vers voulus.

Les poètes auxquels Abû Íayyân at-Tawîîdî a attribué des vers		
Nom du poète	**Nb de vers cités**	**% des citations**
عُروة الصعاليك	5	14%
عبْد الله بن مصعب	4	11%
ابن دارة	4	11%
عُمارة بن عَقيل	4	11%
البديهيّ	3	8%
ابن الروميّ	2	6%
ابن عبّاد	2	6%
ابن هندو	2	6%
الأسديّ	2	6%
خِراش بن زُهير	2	6%
ذو الرُمّة	1	3%
القطاميّ	1	3%
أبو شريح أوس بن حَجَر	1	3%
أبو نواس	1	3%
ابن الحجّاج	1	3%
عَمْرو بن كلثوم	1	3%

Tableau 63

Les vers non attribués par Abû Íayyân at-Tawîîdî et la formule de substitution	
Mot de substitution	**Nombre de vers**
الأوّل	6
الآخر	2
الشاعر	14
أعرابيّ	15
قائل	2
Rien	6

Tableau 64

3.2. Les principales caractéristiques lexicométriques de la partition "Genre"

Les principales caractéristiques lexicométriques que nous livre le dépouillement de la partition "Genre" se résument dans les données quantitatives suivantes :

La part du lion revient, bien entendu, à la prose avec une étendue de N = 60 081 occurrences, et une étendue de vocabulaire de V = 6 494 vocables. Les *hapax* sont au nombre de 3 343. L'article défini ال « *al* » est le lemme ayant la fréquence maximale avec 7 714 occurrences.

La poésie, nous l'avons vu plus haut, est caractérisée par une étendue de N = 1 096 occurrences, une étendue de vocabulaire de V = 467 vocables. Un nombre de *hapax* de 358. L'article défini ال « *al* » étant le lemme ayant la fréquence maximale avec 111 occurrences. La prose constitue 98 % de l'ensemble des occurrences du corpus d'*al-ÞImtâÝ wa-l-MuÞânasa*, alors que la prose n'en représente que 2 %.

Les principales caractéristiques lexicométriques de la partition « Genre » dans *al-ÞimtâÝ wa-l-muÞânasa*

	Nbr occurrences	Nbr vocables	Nbr hapax	Fréq. Max	Lemme
Prose	60 081	6 494	3 343	7 714	ال
Poésie	1 096	467	358	111	ال

Tableau 65

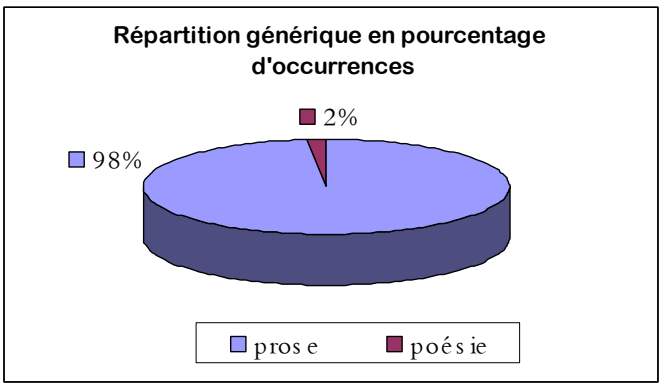


Figure 106

4. Les spécificités du vocabulaire

Le calcul de fréquences pour chacune des unités d'un texte, et la comparaison de ces fréquences d'emploi font partie d'une pratique descriptive certes importante. Mais confronter entre eux, des résultats obtenus à partir de plusieurs textes au sein d'un même corpus ou de différentes parties d'un même texte, est une technique statistique beaucoup plus judicieuse si l'on veut rendre compte des variations observées et avoir une appréciation plus fine de la variabilité de la fréquence des unités d'un corpus divisé en diverses parties. C'est dans cette perspective que l'on fait appel, dans les études lexicométriques à la notion des *spécificités du vocabulaire*.

En considérant les spécificités du vocabulaire dans le discours, on donne le pas à la particularité sur la généralité. Il est alors question d'opposer le *vocabulaire spécifique* à ce qui peut-être appelé *vocabulaire invariant*, *vocabulaire de base* (Bergounioux et al. 1982), *vocabulaire neutre* (Tournier 1975) ou *vocabulaire à spécificités nulles* (Bouterolle-Caporal 1982).

Pour comparer donc plusieurs parties d'un texte sur la base des fréquences de leur vocabulaire, il suffit d'évaluer les variations de la fréquence de chaque unité dans chacune des parties. Pour cela, on a nécessairement besoin de quatre paramètres : la fréquence du vocable dans la partie, la taille de la partie, la fréquence du vocable dans le corpus, et la taille du corpus. La figure de la page suivante explicite l'articulation de ces quatre éléments en relation avec les unités lexicales du corpus et rendant possible l'appréciation de leur variabilité dans les différentes parties composant le corpus.

Le point de départ de l'étude des spécificités est donc un tableau à double entrée obtenu par le croisement des parties du corpus et les fréquences (totales et par partie) correspondant aux différentes unités lexicales contenues dans le corpus ; il s'agit du TLE (Tableau Lexical Entier)[285].

[285] Pour le TLE de notre corpus, voir un extrait p. 127

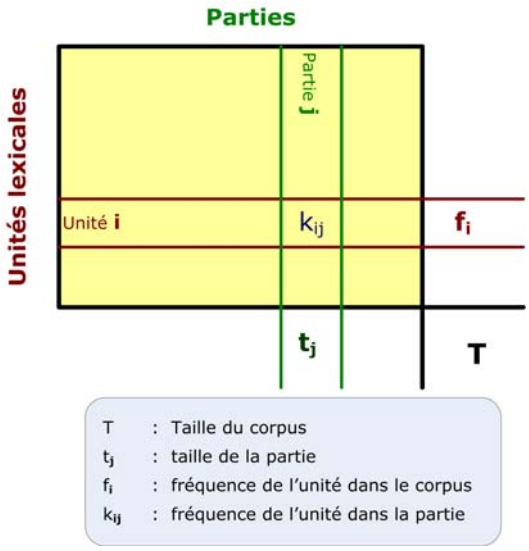


Figure 107 : informations contenues dans le TLE et servant au calcul des spécificités

À l'intersection de la ligne correspondant à l'unité ***i*** et de la colonne correspondant à la partie ***j***, le nombre k_{ij} correspond à la fréquence de l'unité ***i*** dans la partie ***j*** du corpus. f_i représente la fréquence de l'unité ***i*** dans le corpus. La taille (l'étendue) de la partie ***j*** est représentée par t_j. Et, enfin, ***T*** correspond à la taille ou l'étendue du corpus.

Pour ce qui est de l'assise théorique et méthodologique ainsi que des détails de la technique de calcul des *spécificités du vocabulaire*, nous renvoyons le lecteur à l'excellent article de Pierre Lafon (Lafon 1980), l'initiateur de la méthode des *spécificités*, ainsi qu'à l'étude de Benoît Habert (Habert 1985) où l'on trouve un panorama critique des applications de la méthode, et des propositions de son utilisation.

Nous nous contentons de présenter ici quelques uns des résultats concernant le calcul des *spécificités du vocabulaire* d'*al-ÞImtâÝ wa-l-MuÞânasa*, et renvoyons le lecteur à l'annexe G[bis] où nous présentons l'ensemble du vocabulaire spécifique (positif et négatif) du corpus. Ces résultats ont été récupérés à partir du logiciel *Lexico3*[286] qui est doté d'un module de calcul des spécificités.

L'indice de spécificité qui, normalement, se présente sous une notation exponentielle difficile à interpréter pour les non-spécialistes des calculs de probabilité,

[286] Voir la présentation de ce logiciel au chapitre 3, pp. 119-131.

est remplacé dans *Lexico3* par un coefficient simplifié résumant l'ordre de grandeur des suremplois (vocabulaire spécifique positif) et des sous-emplois (vocabulaire spécifique négatif). Ce coefficient est d'autant plus élevé que la sous-fréquence (fréquence du vocable dans la partie) s'écarte d'une répartition *neutre* (vocabulaire de base).

Nous présentons dans le tableau suivant seulement les trois lemmes les plus sur-représentés et les trois lemmes les plus sous-représentés de quelques *Nuits* d'*al-ÞImtâÝ wa-l-MuÞânasa* (*Nuit 01*, *Nuit 02*, *Nuit 07*, *Nuit 10*, *Nuit 14* et *Nuit 16*). Pour la totalité des *spécificités lexicales*, voir l'annexe G[bis].

Spécificités pour la *Nuit 01*

	Forme	Frq. Tot.	Frq.	Coeff.
Suremplois	عتيق	10	8	**10**
	قَالَ	714	58	**7**
	قديم	14	6	**6**
Sous-emplois	إنْ	179	2	**-2**
	ليس	148	1	**-2**
	ها	1080	23	**-4**

Spécificités pour la Nuit 02

	Forme	Frq. Tot.	Frq.	Coeff.
Suremplois	طبّ	10	6	**6**
	عالِم	28	8	**5**
	ي	354	35	**4**
Sous-emplois	شيء	168	2	**-3**
	ها	1080	29	**-5**
	إذا	321	3	**-5**

Spécificités pour la Nuit 07

	Forme	Frq. Tot.	Frq.	Coeff.
Suremplois	حساب	19	17	**22**
	ديوان	18	14	**16**
	بلاغة	27	14	**13**
Sous-emplois	شيء	168	0	**-4**
	ه	2936	88	**-4**
	ي	354	2	**-5**

Spécificités pour la Nuit 10

	Forme	Frq. Tot.	Frq.	Coeff.
Suremplois	ها	1080	398	**66**
	أنثى	45	44	**34**
	إذا	321	137	**31**
Sous-emplois	هم	580	12	**-28**
	كَ	495	1	**-35**
	و	6718	697	**-40**

Spécificités pour la Nuit 14

	Forme	Frq. Tot.	Frq.	Coeff.
Suremplois	سكينة	19	18	**22**
	إضافة	12	9	**10**
	شخْص	17	10	**9**
Sous-emplois	ي	354	4	**-5**
	كَ	495	5	**-7**
	ه	2936	99	**-7**

Spécificités pour la Nuit 16

	Forme	Frq. Tot.	Frq.	Coeff.
Suremplois	نا	156	18	**9**
	جهْل	20	5	**5**
	علْم	64	7	**4**
Sous-emplois	كَ	495	3	**-3**
	إذا	321	1	**-3**
	ها	1080	4	**-6**

Chapitre 10

La richesse lexicale

La richesse lexicale est un élément fondamental dans une perspective lexicométrique de l'étude des textes.

Il est pourtant très difficile de définir la notion de richesse lexicale. Il n'existe, en effet, pas de définition satisfaisante de la richesse lexicale surtout celle prenant en compte la distinction entre composante stylistique et composante thématique du vocabulaire d'un auteur, d'un genre ou d'une époque. Cependant, tous les spécialistes de la lexicométrie s'accordent à accepter un dénominateur commun qui voit dans la richesse lexicale un « lieu de comparaison entre deux ou plusieurs textes en fonction de leur étendue respective et du nombre de vocables relevés dans chacun d'eux »[287].

La notion de richesse lexicale, avec toute sa complexité, surgit dès lors qu'on se pose la question, parfois naïve et subjective : se trouve-t-on devant un vocabulaire riche ou pauvre ? C'est une opération purement descriptive dans ce sens où elle doit être considérée comme un élément lexicométrique de la description linguistique, au sens large, du discours.

Parmi d'autres critères lexicométriques, la richesse lexicale permet de décrire la structure lexicale d'une œuvre donnée ou de la comparer à d'autres corpus. La richesse lexicale est donc une notion indépendante du contenu, c'est un fait de structure. Elle ne doit pas être perçue comme une propriété qualitative ; c'est une donnée quantitative qui a le mérite d'éviter la subjectivité et les jugements "a priori" qui ont toujours entaché, dans le sens commun, les opinions sur la "richesse" ou la "pauvreté" de tel ou tel vocabulaire. Il n'est tout de même pas négligeable que, depuis des années déjà, le vocabulaire est devenu une réalité chiffrable.

D'autre part, peut-on (doit-on) qualifier, même sur des bases quantitatives, tel texte de "riche" en soi ? Ou qu'il l'est plus ou moins que tel autre ? Autrement dit, la richesse lexicale est-elle une notion absolue ou relative ? Il est vrai que la question de savoir si la richesse lexicale a un caractère absolu ou relatif a été posée et longuement débattue dans les débuts de la statistique lexicale. Mais l'on est arrivé depuis quelques années, après plusieurs contradictions, à une quasi unanimité sur le caractère relatif de

[287] Nathan Ménard, *Mesure de la richesse lexicale. Théorie et vérifications expérimentales : Etudes stylométriques et sociolinguistiques*, 1983, p. 16.

la richesse lexicale. En effet, un texte ne peut, en réalité, être qualifié de "riche" ou de "pauvre" que par rapport à d'autres textes. La relativité de la notion de richesse lexicale est une chose acquise aujourd'hui dans le domaine des études quantitatives des textes, et précisément en lexicométrie. Il y a même ceux qui appelle à la "neutralité" du terme "richesse" et pour qui « mieux vaut donc comprendre le mot "richesse" comme un terme neutre, un peu comme font les physiciens pour qui "vitesse" ne se confond pas avec "rapidité" »[288].

Il est donc très fréquent, dans les études lexicométriques des textes, que des comparaisons et des classements des différentes parties ou fragments d'un corpus soient envisagés. Mais l'établissement de la richesse lexicale « n'a pas pour but simplement de satisfaire une manie classificatrice »[289].

Mais comment peut-on, en fin de compte, quantifier une notion qui est dans le sens commun, une propriété qualitative par excellence ? Quels sont les outils pour le faire ? Quelles sont les différentes approches méthodologiques et pratiques qui nous permettent de mesurer la richesse lexicale ?

Nous allons essayer, dans les pages qui suivent, d'exposer un certain nombre des méthodes de mesure de la richesse lexicale qui existent, d'analyser leur fondement théorique et méthodologique et de les appliquer, en fin de compte, à notre corpus .

Tout d'abord, une mise en garde terminologique s'impose : par richesse lexicale, il ne faut pas entendre richesse du lexique de la langue sous-jacente au corpus étudié, du lexique de la communauté linguistique, mais plutôt la richesse de ce que Charles Muller appelle « lexique de situation »[290] c'est-à-dire du vocabulaire. Richesse lexicale équivaut donc à richesse du vocabulaire.

---

[288] Thoiron, Ph., Richesse lexicale et classement des textes, dans *Études sur la richesse et la structure lexicales*, 1988, pp. 141-163, p.142.

[289] Bernet Ch., Faits lexicaux. Richesse du vocabulaire. Résultats, dans *Études sur la richesse et la structure lexicales*, *op. cit.*, pp. 1-11, p.9.

[290] Charles Muller, *Principes et méthodes de statistique lexicale*, p.44-46.

1. Les méthodes de mesure de la richesse lexicale

La mesure de la richesse lexicale est considérée comme l'un des problèmes les plus fréquentés de la lexicométrie.

Les méthodes d'analyse de la richesse lexicale cherchent en quelque sorte à apporter une solution objective, mathématique, à un problème auquel les réponses n'ont été, pendant longtemps, que subjectives, approximatives et impressionnistes.

Tel corpus (telle partie de corpus) est-il (est-elle) plus ou moins riche que tel (telle) autre ? Comment peut-on évaluer le nombre de vocables que comporte un corpus en fonction de son étendue ? L'étendue d'un corpus influe-t-elle sur la mesure de sa richesse lexicale ? Doit-on avoir un indice, une échelle semblable au thermomètre sur laquelle on pourrait situer n'importe quel corpus ; ou plutôt utiliser une méthode de comparaison dont l'objectif serait de classer les corpus (ou les parties de corpus) les uns (les unes) par rapport aux autres et de les comparer en fonction de leur richesse lexicale ?

Les solutions aux différents problèmes soulevés par la mesure de la richesse lexicale ne manquent pas.

On peut utiliser les quotients $\frac{V}{N}$, $\frac{V_1}{V}$ et $\frac{V_1}{N}$ pour évaluer les relations qui existent entre la longueur d'un corpus, l'étendue de son vocabulaire et l'effectif de ses *hapax* et estimer par là-même certaines variations de la richesse du vocabulaire. Néanmoins ces quotients ne sont pas imperméables à l'influence de l'étendue du corpus. Il faudra donc veiller à utiliser des indices ou des méthodes qui permettraient d'isoler le facteur dont on veut étudier le comportement tout en neutralisant ou du moins en minimisant les influences parasites.

Il n'est pas sans intérêt de préciser, en outre, que quel que soit le nombre de corpus ou de parties de corpus à comparer en fonction de leur richesse lexicale, l'on est ramené à une série de comparaisons binaires entre un couple de textes ou de parties de corpus. En effet, pour n textes à comparer entre eux, le nombre de comparaisons binaires sera de :

$$C_n^2 = \frac{n(n-1)}{2}$$

Pour comparer, par exemple, nos 15 nuits deux par deux, le nombre de comparaisons binaires à faire sera de :

$$C_{15}^2 = \frac{15(15-1)}{2} = \frac{15 \times 14}{2} = \frac{210}{2} = 105 \text{ comparaisons binaires.}$$

Vu que le nombre de méthodes et de formules destinées à la mesure de la richesse lexicale est quand même assez important, il serait oiseux de tester et d'appliquer toutes ces formules. Il va donc falloir choisir parmi les formules la ou les méthodes les plus appropriées pour mieux mesurer la richesse lexicale de notre corpus mais aussi, plus généralement, pour instaurer une méthodologie de la mesure de la richesse lexicale en lexicométrie arabe, chose qui fait défaut jusqu'à l'heure actuelle. Parmi toutes les méthodes mises en œuvre, nous avons retenu quelques unes que nous allons exposer dans ce qui suit et appliquer par la suite à notre corpus.

La première méthode, la méthode de comparaison des indices, que nous avons retenue fait partie des toutes premières méthodes élaborées pour la mesure de la richesse lexicale, elle est considérée comme la plus élémentaire de toutes les mesures dans ce sens où elle manipule des indices de premier ordre, des indices bruts, non "dénaturés" par des calculs sophistiqués. C'est une méthode primaire à laquelle on fait appel pour évaluer de prime abord la richesse lexicale. La méthode de comparaison des indices, qui requiert peu de données et un minimum d'opérations, est utilisée comme un outil de première exploration de la richesse lexicale. Cependant, même si cette méthode offre l'avantage d'être simple et d'application facile, il est préférable de parfaire l'analyse de la richesse lexicale par d'autres formules.

En présence de textes dont l'étendue ne varie pas beaucoup, cette méthode donne des résultats satisfaisants et permet un classement pratiquement exhaustif de ce type de textes sur la base de la richesse lexicale. En effet, après avoir effectué 2 340 comparaisons binaires entres tranches d'étendues rapprochées et d'auteurs différents (*Gracq, Butor, Aragon, Camus, Giono et Ramuz*), Nathan Ménard conclut au fait que « pas une fois sur ces 2 340 cas, nous n'avons relevé une réponse « fausse », c'est-à-dire qui contredit le classement de base »[291]

En revanche, quand la condition des étendues rapprochées n'est pas remplie, le recours à d'autres méthodes s'avère nécessaire. Dans ce cas, la ou les méthodes utilisées viennent, soit se substituer à la méthode de comparaison des indices si celle-ci n'est pas concluante, soit la compléter pour dissiper d'éventuelles zones d'ombre au niveau de la mesure de la richesse lexicale.

Quant aux autres méthodes, notre choix a été guidé par deux considérations : d'abord parce que certaines d'entre elles ont déjà fait leurs preuves avec des spécialistes reconnus de la lexicométrie à savoir, par exemple, Charles Muller, Etienne Brunet et Charles Bernet pour ne citer que ces trois, et parce qu'elles ont marqué un moment important de la recherche concernant la mesure de la richesse lexicale et qui continuent à faire partie des principaux outils de cette mesure. Ensuite, parce ce que nous nous sommes fiés et non sans conviction, après bien entendu celle de Nathan Ménard, à l'étude comparative, plus récente et donc plus exhaustive, d'André Cossette[292] concernant un grand nombre de méthodes de mesure de la richesse lexicale. Étude dans laquelle Cossette a « entrepris de mettre au point une méthode de vérification qui soit explicite, rigoureuse et la même pour toutes les formules, permettant ainsi une comparaison de leurs valeurs respectives »[293].

Le propos de l'étude de Cossette n'était pas d'étudier la richesse lexicale en elle-même, mais plutôt d'évaluer les formules qui se présentaient comme des mesures de cette richesse surtout quant à leur capacité de minimiser l'influence de la longueur des textes sur la valeur de la richesse lexicale. C'est pourquoi il a pu établir, et non sans

[291] Nathan Ménard, *op. cit.*, p. 97.

[292] André Cossette, *La richesse lexicale et sa mesure*, 1994, 190 pages.

[293] *Ibidem*, p. 180.

réussite, une méthode d'évaluation robuste des méthodes proposées pour la mesure de la richesse lexicale.

Suite donc aux conclusions auxquelles est arrivé A. Cossette, nous avons choisi d'appliquer à notre corpus les méthodes les plus concluantes, pour son étude, vis-à-vis de la mesure de la richesse lexicale et voir si elles le sont aussi pour notre corpus en particulier et les textes arabes en général. C'est ce à quoi nous allons nous attacher dans les sections suivantes.

1.1. La méthode de comparaison des indices

Dans cette méthode, pour comparer des textes de longueur différente, il suffit de les opposer deux à deux en se basant sur les valeurs de l'étendue du corpus (**N**), de celui du vocabulaire (**V**), du nombre des vocables de fréquence 1 (**V_1**), la fréquence moyenne ($\overline{f} = \frac{N}{V}$) et du taux de répétition (**q_1** = $\frac{V - V_1}{V}$).

Cette méthode peut donner des résultats satisfaisants quand elle est appliquée à des textes de longueur comparable ; mais du moment où l'on a des textes de longueur très variée, la méthode de comparaison des indices montre rapidement ses limites comme nous allons le voir par la suite en l'appliquant à nos quinze nuits qui accusent quand même un écart entre la nuit la plus longue (10 788 occurrences) et la nuit la plus courte (906 occurrences) de pas moins de 9 882 occurrences. Autrement dit, l'ordre de grandeur des nuits varie de 1 à près de 12.

Après avoir classé les textes par ordre décroissant d'étendue, on compare un texte A à un autre texte B situé plus loin dans la liste, donc forcément plus court ; ce qui permet de partir a priori de la condition : **$N_A > N_B$**.

Les comparaisons binaires des textes vont donc se faire en confrontant les quatre indices dans l'ordre V, V_1, $\overline{f}$ et q_1. Et on écrira, par convention, **1** si $V_A > V_B$ et **0** si

$V_A < V_B$; ainsi que pour V_{1A} et V_{1B}, $\overline{f}_A$ et $\overline{f}_B$ et q_{1A} et q_{1B}. On obtient ainsi une combinaison de 4 chiffres pour chaque comparaison binaire du type (1 - 1 - 0 - 0) comme dans les exemples suivants :

		N	***V***	***V1***	$\overline{f}$	***q1***
Texte A	***Nuit 8***	10788	1967	1249	5,484	0,3650
Texte B	***Nuit 5***	906	340	265	2,665	0,2206
			1	**1**	**1**	**1**

		N	***V***	***V1***	$\overline{f}$	***q1***
Texte A	***Nuit 9***	4607	1003	637	4,593	0,3649
Texte B	***Nuit 4***	4248	1133	834	3,749	0,2639
			0	**0**	**1**	**1**

1.1.1. Combinaisons possibles et combinaisons impossibles

Théoriquement, il y a 16 combinaisons de ce type ; mais, pour des raisons algébriques, 7 d'entre elles sont impossibles étant donné que $\overline{f}$ et q_1 sont étroitement liés aux autres valeurs :

- Étant donné que $N_A > N_B$, si $V_A < V_B$, on a obligatoirement $\overline{f}_A > \overline{f}_B$[294].
  Ce qui se traduit par le fait que les combinaisons ayant un 0 en première position ($V_A < V_B$) ne peuvent avoir 0 en troisième position ($\overline{f}_A < \overline{f}_B$), donc les combinaisons suivantes sont à exclure :

  0 - 0 - **0** - 0
  0 - 0 - **0** - 1
  0 - 1 - **0** - 0
  0 - 1 - **0** - 1

- D'un autre côté, si $V_A < V_B$ et $V_{1A} > V_{1B}$, cela entraine que $q_{1A} < q_{1B}$.
  Ce qui veut dire que les combinaisons ayant 0 - 1 pour les deux premières positions ne peuvent avoir 1 en quatrième position. Les deux combinaisons suivantes sont donc exclues :

[294] Voir la démonstration mathématique dans, Charles Muller, « Sur la mesure de la richesse lexicale. Théorie et méthodes » dans *Langue française et linguistique quantitative. Recueil d'articles*, 1979, p. 281-307, p. 285

0 - **1** - 0 - **1** (déjà exclue ci-dessus)
0 - **1** - 1 - **1**

- Si, à l'inverse on a $V_A > V_B$ et $V_{1A} < V_{1B}$, cela entraine que $q_{1A} < q_{1B}$.
  Il en résulte que les combinaisons commençant par 1 - 0 ne peuvent se terminer par un 0, ce qui exclut :

1 - **0** - 0 - **0**
1 - **0** - 1 - **0**

Ces 7 combinaisons impossibles exclues, il ne reste plus que 9 combinaisons réalisables parmi les 16 théoriques :

1 - 1 - 1 - 1
1 - 1 - 0 - 0
0 - 0 - 1 - 1
0 - 0 - 1 - 0
0 - 1 - 1 - 0
1 - 0 - 0 - 1
1 - 1 - 0 - 1
1 - 0 - 1 - 1
1 - 1 - 1 - 0

1.1.2. Effets de l'étendue du texte sur les indices

Les indices ne sont pas insensibles à la variation de la longueur d'un texte. En effet, en faisant croître l'étendue, on remarque qu'en général les quatre indices croissent en même temps. V peut croître moins vite que N ou rester stable pendant une certaine étendue de texte ; V_1 quant à lui, peut croître, décroître ou rester stable pendant une certaine étendue de texte, mais dans son allure générale, il croît mais moins vite que V ; $\overline{f}$ peut croître ou décroître ou rester stable pendant une certaine étendue de texte, mais de manière générale il croît de 1 à l'infini ; alors que q_1 peut croître ou décroître ou rester stable pendant une certaine étendue de texte, mais de manière générale il croît de 0 à 1.

Autrement dit, si l'on compare deux fragments A et B d'un même texte où le lexique est stable et l'apport lexical constant, pour $N_A > N_B$ on aura :

$$V_A > V_B$$
$$V_{1A} > V_{1B}$$
$$\overline{f}_A > \overline{f}_B$$
$$q_{1A} > q_{1B}$$

Il est aussi important de noter, à la suite de Charles Muller, que « V_1 et q_1 sont les indices qui, dans un même texte, se montrent les plus sensibles à un ralentissement de l'apport lexical, et que de ce fait leur témoignage ne devra être enregistré qu'avec prudence ; qu'en revanche un écart observé sur $\overline{f}$ sera hautement significatif »[295]

1.1.3. Effets de la richesse lexicale sur les indices

D'un autre côté, il est important de noter que les deux indices V et V_1 qui sont des indices absolus ne varient pas de la même manière que les deux autres indices $\overline{f}$ et q_1 qui sont des indices relatifs. En effet, les premiers croissent d'autant **plus** vite que le lexique est plus riche, alors les seconds croissent d'autant **moins** vite que le lexique est plus riche.

Donc, à richesse lexicale égale ($\mathcal{RL}_A \approx \mathcal{RL}_B$), le texte le plus long aura ses quatre indices supérieurs à ceux du texte plus court. À longueur égale ($N_A \approx N_B$) cette fois-ci, le texte le plus riche lexicalement aura V et V_1 supérieurs, et $\overline{f}$ et q_1 inférieurs à ceux du texte plus court.

$\mathcal{RL}_A \approx \mathcal{RL}_B$ et $N_A > N_B$
$V_A > V_B$
$V_{1A} > V_{1B}$
$\overline{f}_A > \overline{f}_B$
$q_{1A} > q_{1B}$

$\mathcal{RL}_A > \mathcal{RL}_B$ et $N_A \approx N_B$
$V_A > V_B$
$V_{1A} > V_{1B}$
$\overline{f}_A < \overline{f}_B$
$q_{1A} < q_{1B}$

[295] Ch. Muller, op. cit., p. 288

1.1.4. Interprétation des indices

Pour toute opération d'interprétation, il ne faut pas oublier notre condition de départ qui postule que $N_A > N_B$.

Dans le tableau 1, les interprétations sûres figurent sans parenthèses alors que les interprétations acceptables sont encadrées par des parenthèses.

Interprétation des combinaisons d'indices

	Combinaisons	Indicateurs partiels de la richesse lexicale				Interprétation globale	Symbole
		V	V_1	$\overline{f}$	q_1		
❶	**1 - 1 - 1 - 1**	?	?	?	?	**?**	**?**
❷	**1 - 1 - 0 - 0**	?	?	A	A	**A plus** riche que **B**	**+**
❸	**0 - 0 - 1 - 1**	B	B	?	?	**A moins** riche que **B**	**–**
❹	**0 - 0 - 1 - 0**	B	B	?	A	**(A moins** riche que **B)**	**(–)**
❺	**0 - 1 - 1 - 0**	B	?	?	A	**(A moins** riche que **B)**	**(–)**
❻	**1 - 0 - 0 - 1**	?	B	A	?	**(A plus** riche que **B)**	**(+)**
❼	**1 - 1 - 0 - 1**	?	?	A	?	**(A plus** riche que **B)**	**(+)**
❽	**1 - 0 - 1 - 1**	?	B	?	?	**(A moins** riche que **B)***	**(–)**
❾	**1 - 1 - 1 - 0**	?	?	?	A	**?**	**?**

Tableau 66

Quand les textes comparés sont de longueur très différente ou quand l'écart de richesse lexicale n'est pas assez net pour compenser l'écart d'étendue, c'est la combinaison ❶ que nous obtenons. Dans ce cas, aucune conclusion ne peut être tirée et la comparaison binaire reste non résolue.

Les combinaisons ❷ et ❸ se rapportent aux situations les plus claires et produisent des résultats sûrs. Ces combinaisons traduisent le fait que l'écart de longueur est neutralisé par l'écart de richesse lexicale.

* Cette interprétation n'est acceptable que si $(V_{1B} - V_{1A}) > (V_A - V_B)$.

Quant aux combinaisons ❹, ❺ et ❻, on y trouve, combinés, des indices sûrs et d'autres douteux ; mais l'avantage est donné aux indices les plus sûrs par rapport aux plus douteux. En effet, dans les combinaisons ❹ et ❺, par exemple, l'information donnée par $V_A < V_B$ est favorisée à celle apportée par $q_{1A} < q_{1B}$; alors que , dans la combinaison ❻, la constatation que $\overline{f}_A < \overline{f}_B$ l'emporte sur celle que $V_{1A} < V_{1B}$.

Vu qu' elle ne s'appuie que sur un seul indice, l'interprétation des combinaisons ❼, ❽ et ❾ est une interprétation un peu délicate. Dans la combinaison ❼, le fait qu'on ait $\overline{f}_A < \overline{f}_B$ peut aspirer confiance et permettre une conclusion acceptable. Pour la combinaison ❽, $V_{1A} < V_{1B}$ n'est acceptable que sous la condition $(V_{1B} - V_{1A}) > (V_A - V_B)$. Alors que dans la combinaison ❾, seul l'indice q_1 est en jeu, ce qui est très insuffisant pour permettre une conclusion. Il est préférable dans ce cas de considérer comme non résolue la comparaison binaire.

1.1.5. Application au corpus *al-ÞImtâÝ wa l-muÞânasa*

Valeurs des indices pour chacune des nuits d'*al-ÞImtâÝ wa l-muÞânasa*

Nuits	*N*	*V*	V_1	$\overline{f}$	q_1
Nuit 08	10788	1967	1249	5,484	0,3650
Nuit 10	9564	1772	1033	5,397	0,4170
Nuit 06	7079	1644	1140	4,306	0,3066
Nuit 00	5062	1412	1048	3,585	0,2578
Nuit 09	4607	1003	637	4,593	0,3649
Nuit 04	4248	1133	834	3,749	0,2639
Nuit 14	3271	819	546	3,994	0,3333
Nuit 02	3115	896	649	3,477	0,2757
Nuit 07	2569	688	464	3,734	0,3256
Nuit 01	2478	703	498	3,525	0,2916
Nuit 13	2427	535	342	4,536	0,3607
Nuit 03	2004	627	470	3,196	0,2504
Nuit 15	1773	510	344	3,476	0,3255
Nuit 16	1286	420	313	3,062	0,2548
Nuit 05	906	340	265	2,665	0,2206

Tableau 67

Nous avons présenté dans le tableau 2, dans la première colonne, les nuits classées par ordre décroissant d'étendue, puis dans les colonnes suivantes respectivement l'étendue de chaque nuit ***N***, l'étendue du vocabulaire ***V***, le nombre de vocables de fréquence 1 V_1, la fréquence moyenne $\overline{f}$ et le taux de répétition q_1.

À partir de ces données nous avons donc confronté la valeur de ces indices pour toutes les nuits prises deux à deux. Comme indiqué plus haut, la valeur 1 est donnée au résultat de la confrontation pour l'indice considéré si celui de la nuit A est supérieur à celui de la nuit B, la valeur 0 dans le cas contraire. À la sortie de toutes ces confrontations, nous avons obtenu 105 combinaisons de 4 chiffres.

Le tableau 3 de la page suivante, représente les 105 comparaisons binaires résultant de la confrontation, deux à deux, des 15 nuits d'*al-ÞImtâÝ wa l-muÞânasa*. À l'intersection de chaque ligne et de chaque colonne, nous avons inscrit la combinaison de 4 chiffres représentant le résultat de la confrontation un à un des quatre indices qui sont, dans l'ordre, V, V_1, $\overline{f}$ et q_1. Pour ce qui concerne, par exemple, la comparaison entre la neuvième et la deuxième nuit, nous obtenons la combinaison (1 - 0 - 1 - 1) , et ce parce que $V_{nuit\ 9} > V_{nuit\ 2} \rightarrow (\mathbf{1})$, $V_{1nuit\ 9} < V_{1nuit\ 2} \rightarrow (\mathbf{0})$, $\overline{f}_{nuit\ 9} > \overline{f}_{nui\ t2} \rightarrow (\mathbf{1})$ et $q_{1\ nuit\ 9} > q_{1\ nuit\ 2} \rightarrow (\mathbf{1})$:

	N	***V***	V_1	$\overline{f}$	q_1
Nuit 9 **:**	4607	1003	637	4,593	0,3649
Nuit 2 **:**	3115	896	649	3,477	0,2757
Combinaison obtenue →		**(1 - 0 - 1 - 1)**			

Pour une meilleure lecture de ce tableau, nous avons utilisé une couleur différente, comme trame de fond, pour chaque combinaison différente. Sur les 9 combinaisons possibles, nous n'avons enregistré dans la comparaison des parties de notre corpus que cinq combinaisons. Les combinaisons ❹ (0 - 0 - 1 - 0), ❺ (0 - 1 - 1 - 0), ❻ (1 - 0 - 0 - 1) et ❼ (1 - 1 - 0 - 1), et qui sont, au demeurant, toutes des combinaisons acceptables, ne sont pas réalisées dans notre corpus.

Les combinaisons d'indices correspondant aux 105 comparaisons binaires des 15 nuits d'*al-ÞImtâÝ wa l-muÞânasa*

Nuit	N05	N16	N15	N03	N13	N01	N07	N02	N14	N04	N09	N00	N06	N10
N08	1 - 1 - 1 - 1	1 - 1 - 1 - 1	1 - 1 - 1 - 1	1 - 1 - 1 - 1	1 - 1 - 1 - 1	1 - 1 - 1 - 1	1 - 1 - 1 - 1	1 - 1 - 1 - 1	1 - 1 - 1 - 1	1 - 1 - 1 - 1	1 - 1 - 1 - 1	1 - 1 - 1 - 1	1 - 1 - 1 - 1	1 - 1 - 1 - 0
N10	1 - 1 - 1 - 1	1 - 1 - 1 - 1	1 - 1 - 1 - 1	1 - 1 - 1 - 1	1 - 1 - 1 - 1	1 - 1 - 1 - 1	1 - 1 - 1 - 1	1 - 1 - 1 - 1	1 - 1 - 1 - 1	1 - 1 - 1 - 1	1 - 1 - 1 - 1	1 - 0 - 1 - 1	1 - 0 - 1 - 1	
N06	1 - 1 - 1 - 1	1 - 1 - 1 - 1	1 - 1 - 1 - 0	1 - 1 - 1 - 1	1 - 1 - 0 - 0	1 - 1 - 1 - 1	1 - 1 - 1 - 0	1 - 1 - 1 - 1	1 - 1 - 1 - 0	1 - 1 - 1 - 1	1 - 1 - 0 - 0	1 - 1 - 1 - 1		
N00	1 - 1 - 1 - 1	1 - 1 - 1 - 1	1 - 1 - 1 - 0	1 - 1 - 1 - 1	1 - 1 - 0 - 0	1 - 1 - 1 - 0	1 - 1 - 0 - 0	1 - 1 - 1 - 0	1 - 1 - 0 - 0	1 - 1 - 0 - 0	1 - 1 - 0 - 0			
N09	1 - 1 - 1 - 1	1 - 1 - 1 - 1	1 - 1 - 1 - 1	1 - 1 - 1 - 1	1 - 1 - 1 - 1	1 - 1 - 1 - 1	1 - 1 - 1 - 1	1 - 0 - 1 - 1	1 - 1 - 1 - 1	0 - 0 - 1 - 1				
N04	1 - 1 - 1 - 1	1 - 1 - 1 - 1	1 - 1 - 1 - 0	1 - 1 - 1 - 1	1 - 1 - 0 - 0	1 - 1 - 1 - 0	1 - 1 - 1 - 0	1 - 1 - 1 - 0	1 - 1 - 0 - 0					
N14	1 - 1 - 1 - 1	1 - 1 - 1 - 1	1 - 1 - 1 - 1	1 - 1 - 1 - 1	1 - 1 - 0 - 0	1 - 1 - 1 - 1	1 - 1 - 1 - 1	0 - 0 - 1 - 1						
N02	1 - 1 - 1 - 1	1 - 1 - 1 - 1	1 - 1 - 1 - 0	1 - 1 - 1 - 1	1 - 1 - 0 - 0	1 - 1 - 0 - 0	1 - 1 - 0 - 0							
N07	1 - 1 - 1 - 1	1 - 1 - 1 - 1	1 - 1 - 1 - 1	1 - 0 - 1 - 1	1 - 1 - 0 - 0	0 - 0 - 1 - 1								
N01	1 - 1 - 1 - 1	1 - 1 - 1 - 1	1 - 1 - 1 - 0	1 - 1 - 1 - 1	1 - 1 - 0 - 0									
N13	1 - 1 - 1 - 1	1 - 1 - 1 - 1	1 - 0 - 1 - 1	0 - 0 - 1 - 1										
N03	1 - 1 - 1 - 1	1 - 1 - 1 - 0	1 - 1 - 0 - 0											
N15	1 - 1 - 1 - 1	1 - 1 - 1 - 1												
N16	1 - 1 - 1 - 1													

Sens de lecture du tableau

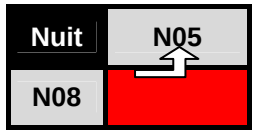


Tableau 68

Comme nous l'indiquions plus haut, seulement 5 combinaisons sur les 9 possibles sont actualisées dans notre corpus : deux combinaisons dont les réponses restent non résolues, se sont les combinaisons ❶ (1 - 1 - 1 - 1) et ❾ (1 - 1 - 1 - 0) ; elles seront donc remplacées, dans le tableau 4, par le symbole « **?** ». Deux combinaisons sûres, il s'agit de la combinaison ❷ (1 - 1 - 0 - 0) pour laquelle le texte A est plus riche que le texte B et la combinaison ❸ (0 - 0 - 1 - 1) pour laquelle le texte B est plus riche que le texte A ; elles seront remplacées, dans le tableau 4, respectivement par le symbole « + » et « - ». Quant à la cinquième combinaison ❽ (1 - 0 - 1 - 1), la condition nécessaire pour qu'elle soit considérée comme acceptable, c'est-à-dire $(V_{1B} - V_{1A}) > (V_A - V_B)$ n'est remplie pour aucune des cinq comparaisons binaires rencontrées *(Nuit13, Nuit15)*, *(Nuit07, Nuit03)*, *(Nuit09, Nuit02)*, *(Nuit10, Nuit00)*, *(Nuit10, Nuit06)*. En effet, pour, par exemple, la comparaison entre la nuit 13 et la nuit 15 :

$$V_{1B} - V_{1A} = 344 - 342 = 2$$
$$V_A - V_B = 535 - 510 = 25$$

et comme 2 n'est pas supérieur à 25, alors la condition n'est pas remplie. C'est pareil pour les quatre autres comparaisons binaires. Étant donné qu'elle ne peut être considérée comme acceptable, la combinaison ❽ doit donc être déclarée comme non résolue. Le nombre des comparaisons binaires présentant la combinaison ❽ viendra donc s'ajouter à celui des comparaisons binaires ayant les combinaisons ❶ et ❾ qui est déjà très élevé, pour représenter ensemble 81 % de la totalité des comparaisons binaires.

Le tableau suivant (tableau 4) doit être lu horizontalement, c'est-à-dire que la nuit inscrite au début de la ligne est comparée à la nuit inscrite en haut de chaque colonne. Le signe « + » à l'intersection d'une ligne et d'une colonne signifie que la nuit inscrite au début de la ligne est « plus riche » que celle qui est inscrite en haut de la colonne, le signe « - » signifie qu'elle est « plus pauvre », alors que le point d'interrogation « **?** » signifie qu'on ne peut rien déduire de cette comparaison. Nous n'avons dans notre tableau que des réponses sûres ou des réponses non résolues, l'absence de parenthèses traduit donc l'absence de réponses acceptables.

La richesse lexicale des 15 nuits d'*al-ÞImtâÝ wa l-muÞânasa* par la méthode de comparaison des indices

Nuit	N05	N16	N15	N03	N13	N01	N07	N02	N14	N04	N09	N00	N06	N10
N08	?	?	?	?	?	?	?	?	?	?	?	?	?	?
N10	?	?	?	?	?	?	?	?	?	?	?	?	?	
N06	?	?	?	?	+	?	?	?	?	?	+	?		
N00	?	?	?	?	+	?	+	?	+	+	+			
N09	?	?	?	?	?	?	?	?	?	–				
N04	?	?	?	?	+	?	?	?	+					
N14	?	?	?	?	+	?	?	–						
N02	?	?	?	?	+	+	+							
N07	?	?	?	?	+	–								
N01	?	?	?	?	+									
N13	?	?	?	–										
N03	?	?	+											
N15	?	?												
N16	?													

Réponses sûres : 19,05 %

« **+** » = 16 / 105 → 15,24 %

« **-** » = 4 / 105 → 3,81 %

Réponses non résolues : « **?** » = 85 / 105 → 80,95 %

Tableau 69

Il est inutile de rappeler que la relation « plus riche que » est une relation associative dans ce sens où lorsqu'on a **A** plus riche que **B** et **B** plus riche que **C,** alors on a forcément **A** plus riche que **C**. Dans cette logique, il n'est donc pas nécessaire de tracer, dans la figure 1, une flèche allant, par exemple, de la nuit 02 vers la nuit 13 (la nuit 02 étant plus riche que la nuit 13, ce qui traduit, dans le tableau 4 le signe « + » à l'intersection de la ligne nuit 02 et de la colonne nuit 13) et ce parce que la nuit 02 est plus riche que la nuit 01 qui est plus riche que la nuit 07 elle-même plus riche que la nuit 13. Par associativité, nous avons donc la nuit 01 plus riche que la nuit 13 et aussi la nuit 02 plus riche que la nuit 13.

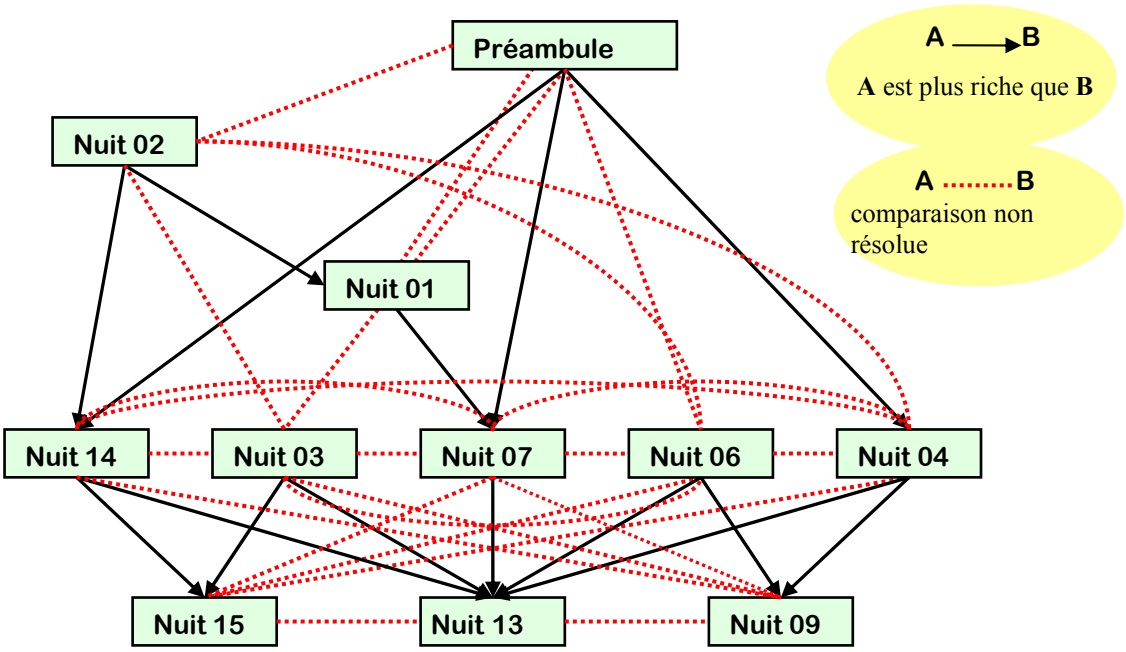


Figure 108
Classement de quelques nuits en fonction de la richesse lexicale
selon la méthode des indices

Quelques ébauches de classement se dégagent de cette méthode qui nous permettent déjà de faire les classements rudimentaires suivants :

Nuit 2 > Nuit 1 > Nuit 7 > Nuit 13
Nuit 2 > Nuit 14 > Nuit 15
Préambule > Nuit 14 > Nuit 15

Préambule > Nuit 7 > Nuit 13
Préambule > Nuit 4 > Nuit 9
Nuit 3 > Nuit 13
Nuit 3 > Nuit 15
Nuit 6 > Nuit 9
Nuit 6 > Nuit 13

Il est clair, de la figure 1 et du tableau 4, que cette méthode de comparaison des indices ne nous permet pas un classement complet de toutes les nuits que comporte notre corpus. Ce que nous avons présenté dans la figure 1, ce n'est qu'un classement rudimentaire ne représentant que 22 comparaisons binaires sur les 105 (15 flèches apparentes + 7 par associativité).

Dans le tableau 4, sur les 105 symboles au croisement des lignes et des colonnes, il n'y a que 16 signes « + » représentant 15,24 % et 4 signes « - » représentant 3,81 % ; cela ne représente que 19 % de comparaisons binaires sûres et concluantes sur l'ensemble des 105. Si l'on ajoute les comparaisons résolues par associativité, ce taux arrive à 21 % ; 79 % des comparaisons ne sont donc pas distinguées par les indices. En plus des cas résolus, nous avons représenté dans la figure 1 quelques unes des comparaisons non résolues par des liaisons en pointillé rouge. Pour ne pas trop charger la figure, nous n'avons pas voulu tracer toutes les liaisons représentant les comparaisons non résolues entre les nuits de la figure mais aussi entre elles et les nuits non représentées sur cette même figure.

Nous nous trouvons donc dans l'obligation de conclure, en dépit des quelques cas résolus et de quelques classements rudimentaires, à l'insuffisance de cette « méthode de comparaison des indices » pour des corpus tels que le notre à cause du grand écart qui existe entre les valeurs de l'étendue des nuits prises deux par deux. Mais insuffisance ne veut pas dire refus catégorique de l'utilisation de cette méthode ; la méthode des indices peut servir, dans le pire des cas, comme une sonde, comme un outil d'exploration qui peut rendre au chercheur en lexicométrie d'énormes services ne serait-ce que pour choisir quelle autre méthode utiliser en tandem avec elle.

1.2. La formule de Guiraud ($V/\sqrt{N}$)

Cette méthode n'est pas non plus indépendante de l'étendue des textes, mais elle l'est apparemment moins que la méthode de comparaison des indices puisqu'elle ne met pas en quotient V et N mais plutôt V et la racine carrée de N ($\sqrt{N}$) dans le but de minimiser l'influence de l'étendue sur la mesure de la richesse lexicale.

Dans ses *Problèmes et méthodes de la statistique linguistique*[296], Pierre Guiraud trouve que le vocabulaire d'un texte est proportionnel à la racine carrée de sa longueur :

$$\frac{V}{\sqrt{N}} = \mathrm{C}$$

C est une constante qui tend, selon P. Guiraud, vers 22. Cependant d'après nos calculs sur les nuits d'*al-ÞImtâÝ wa l-muÞânasa*, elle varie de 10,860, pour sa valeur minimale, à 19,846 se rapprochant ainsi de 20 sans jamais l'atteindre. Les variations de la valeur de **C** sont telles qu'il est très difficile de pouvoir parler de constante. La grande inégalité dans l'étendue des nuits d'*al-ÞImtâÝ wa l-muÞânasa* pourrait être à l'origine des oscillations qu'enregistre la valeur de **C** et qui ont quand même une importante amplitude.

Nous présentons dans le tableau 5 les résultats de la mesure de la richesse lexicale concernant les nuits d'*al-ÞImtâÝ wa l-muÞânasa* et dans lequel nous avons inscrit, dans la partie de gauche, dans l'ordre des colonnes, les nuits classées par ordre chronologique, l'étendue de chaque nuit N, l'étendue de son vocabulaire V, et enfin le quotient $V/\sqrt{N}$ qui est donc l'indice de la richesse lexicale selon la formule de Guiraud. Dans la partie droite du tableau, nous avons classé les nuits par ordre décroissant de richesse lexicale.

[296] Pierre Guiraud, *Problèmes et méthodes de la statistique linguistique*, 1960, p. 89.

Richesse lexicale des nuits d'*al-ÞImtâÝ wa l-muÞânasa* selon la formule de Guiraud

Nuits	*N*	*V*	$V/\sqrt{N}$
Nuit 00	5062	1412	19,846
Nuit 01	2478	703	14,122
Nuit 02	3115	896	16,054
Nuit 03	2004	627	14,006
Nuit 04	4248	1133	17,384
Nuit 05	906	340	11,296
Nuit 06	7079	1644	19,540
Nuit 07	2569	688	13,574
Nuit 08	10788	1967	18,938
Nuit 09	4607	1003	14,777
Nuit 10	9564	1772	18,119
Nuit 13	2427	535	10,860
Nuit 14	3271	819	14,320
Nuit 15	1773	510	12,112
Nuit 16	1286	420	11,712

Classement		
Nuit 00	5062	19,846
Nuit 06	7079	19,540
Nuit 08	10788	18,938
Nuit 10	9564	18,119
Nuit 04	4248	17,384
Nuit 02	3115	16,054
Nuit 09	4607	14,777
Nuit 14	3271	14,320
Nuit 01	2478	14,122
Nuit 03	2004	14,006
Nuit 07	2569	13,574
Nuit 15	1773	12,112
Nuit 16	1286	11,712
Nuit 05	906	11,296
Nuit 13	2427	10,860

Tableau 70

Ce qui est important à remarquer dans le résultat de cette méthode c'est qu'elle confirme certaines tendances que nous avons pu voir avec la méthode de comparaison des indices utilisée plus haut et qu'elle ajoute des classements que nous n'avons pu avoir précédemment. En effet, le tableau 5 nous confirme bien, par exemple, que le préambule (la nuit 00) est plus riche que les nuits 14, 15, 07, 13, 04 et 09 (il est même plus riche, selon ce classement, que toutes les nuits) et que la nuit 02 est plus riche que les nuits 14, 15, 01, 07 et 13. Aussi, avons-nous la confirmation que la nuit 13 est la moins riche de toutes les autres (sauf que, selon la méthode des indices, l'on ne peut rien dire de la comparaison de la nuit 13 avec les nuits 15 et 09).

Les informations sur les 81 % des comparaisons binaires sur la base de la richesse lexicale que nous n'avons pas pu obtenir par la méthode de comparaison des indices, nous ont bien été révélées par cette méthode, la formule $V/\sqrt{N}$ de Pierre Guiraud.

Le tableau 5 et surtout, d'une façon visuelle, la figure 2 nous résument le classement des nuits d'*al-ÞImtâÝ wa l-muÞânasa* sur la base de la richesse lexicale obtenu par la méthode Guiraud.

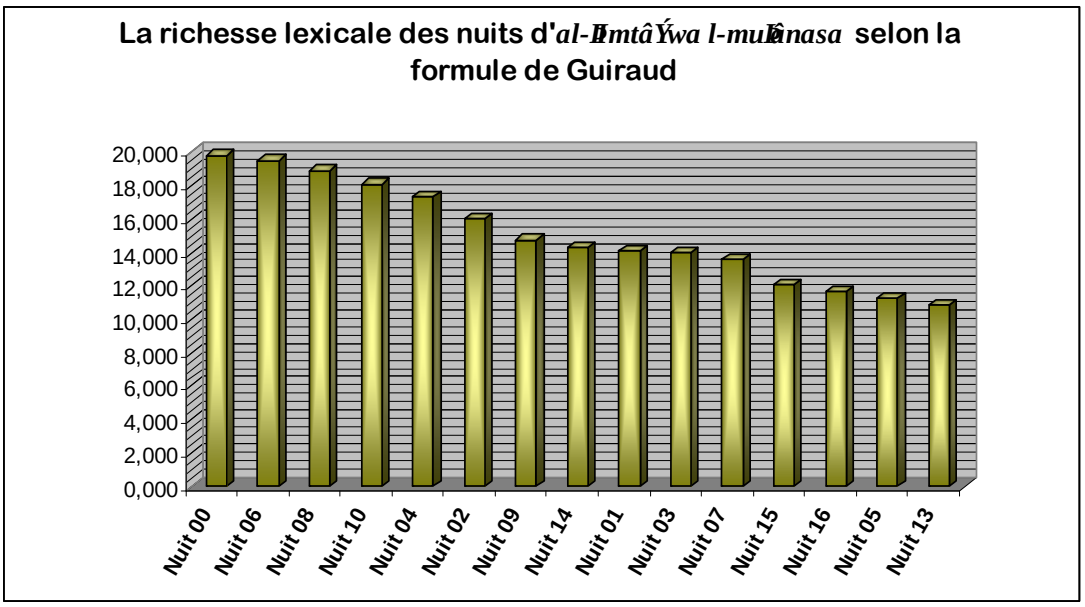


Figure 109

Une lecture plus attentive de ce graphique et du tableau 5, nous montre une sensibilité de la richesse lexicale selon cette formule aux valeurs de l'étendue des nuits. Ce constat nous a bien été confirmé par le test de corrélation des rangs de Spearman qui nous livre un coefficient de corrélation de 0,900. Au seuil de signification $\alpha = 0,05$, on peut rejeter l'hypothèse nulle d'absence de corrélation. Autrement dit, la corrélation est significative, et la dispersion des rangs sur le graphique de la (figure 3) le montre très clairement.

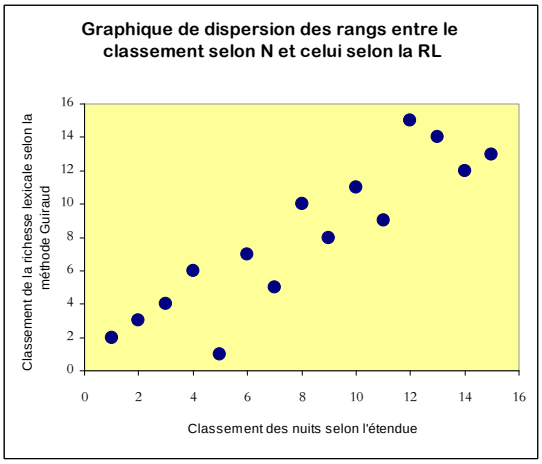


Figure 110

Au niveau de la répartition de la richesse lexicale des nuits autour de la richesse lexicale moyenne, qui est ici de 15,111, l'on voit bien dans le graphique de la figure 4, la variation en dents de scie des valeurs de richesse lexicale selon la formule Guiraud. Les nuits ayant une richesse lexicale supérieure à la moyenne (6 nuits) sont minoritaires par rapport à celles à richesse lexicale au-dessous de la moyenne (9 nuits).

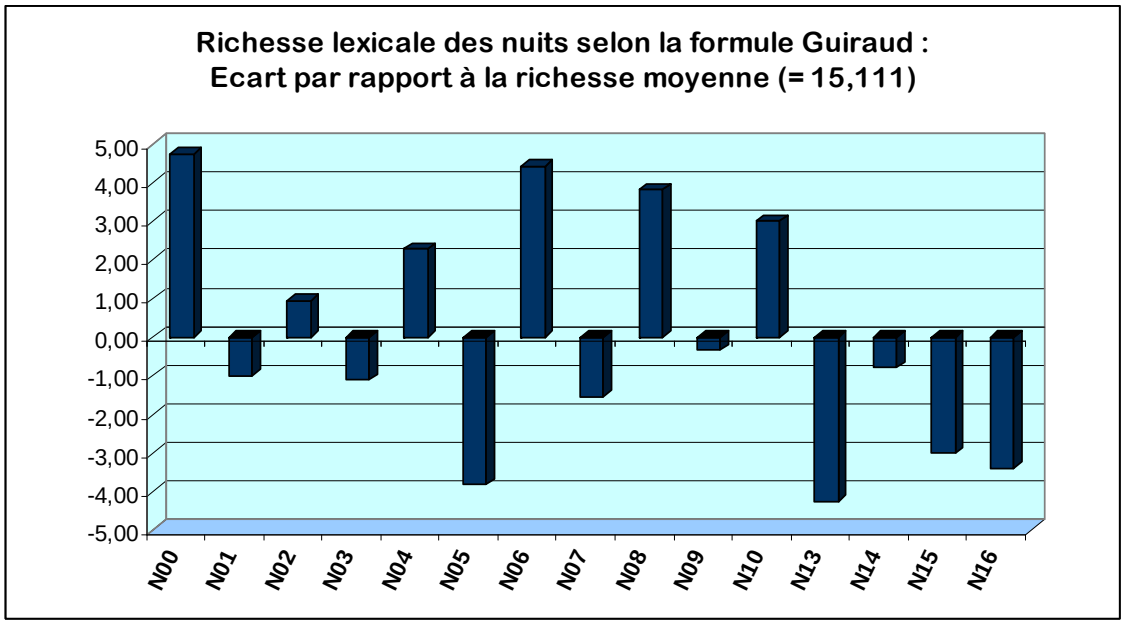


Figure 111

Les nuits les plus riches lexicalement sont , dans l'ordre, la nuit 0, la nuit 6, la nuit 8 et la nuit 10. Alors que les nuits les plus pauvres lexicalement sont les nuits 13, 5 16 et 15.

Autour de la richesse moyenne, nous trouvons la nuit 9 dont la richesse lexicale est très proche, négativement, de la moyenne. Ensuite, s'en écartent un peu plus, la nuit 14, négativement, et la nuit 2, positivement.

1.3. L'indice W de Brunet

Pierre Guiraud, nous l'avons vu plus haut, avait comme objectif, dans l'élaboration de sa formule ($V/\sqrt{N}$), de réduire par le biais de la racine carrée, la progression de N pour qu'elle accompagne plus fidèlement celle de V.

S'inscrivant dans la continuité de P. Guiraud et constatant que la racine carrée est un facteur de réduction « trop constant et trop rigide pour annuler les distorsions quand les œuvres sont d'étendue trop inégale. Et ce qui est pire, cette distorsion agit dans un sens puis dans l'autre, ce qui ruine toute possibilité de correction »[297], Etienne Brunet a mis au point un indice de mesure de la richesse lexicale se fondant également sur les deux données de base N et V, mais impliquant un facteur de réduction plus souple et plus fidèle que la simple racine carrée de N. Dans un souci donc de minimiser au maximum l'influence de l'étendue des textes sur la valeur de la richesse lexicale, E. Brunet fait jouer le rôle du facteur de réduction de N à la valeur de V, ou plus exactement à la réciproque de cette valeur (1/V), une fois qu'elle aura été elle-même convenablement réduite par un exposant fractionnaire (α). La formule de l'indice W de Brunet s'écrit donc ainsi :

$$W = N^{V^{-\alpha}} \quad (\text{ ce qui équivalent à } W = N^{1/V^{\alpha}} \text{ ou encore } W = \sqrt[V^{\alpha}]{N}\ )^{298}$$

où α est une constante dont la valeur, calculée à la suite de nombreuses opérations sur ordinateur, est de 0,172.

---

[297] Etienne Brunet, *Le vocabulaire de Jean Giraudoux. Structure et évolution. Statistique et informatique appliquées à l'étude des textes à partir des données du Trésor de la Langue Française*, Slatkine, Genève, 1978, p. 49.

[298] Nous ne prenons pas en considération le léger correctif à cette formule proposé par Brunet en 1978 en soustrayant de V un facteur « b » qui a la valeur 20 et ce pour parier au fait que l'indice W cesse d'être indifférent à l'étendue en ce qui concerne le niveau des phrases qui est le niveau le plus bas de la longueur des textes. La formule avec le correctif devient alors : $N^{1/(V-b)^{\alpha}}$. Ce même facteur de correction a été remplacé en 1981, toujours par E. Brunet-même, par ($\_\ \frac{0{,}01}{\log N}$).

Les valeurs de W se situent, empiriquement, entre 10 et 25. De plus, la valeur de l'indice W évolue en raison inverse de la richesse lexicale, c'est-à-dire que W est d'autant plus grand que le vocabulaire est plus pauvre et *vice versa*.

Pour avoir des valeurs de W qui évoluent dans le même sens que la richesse lexicale, on peut ramener chaque valeur de W à une valeur R selon la formule suivante :

$$\mathbf{R} = \frac{25 - W}{15}$$

On aura ainsi un indice de richesse lexicale variant de 0 à 1 et qui est d'autant plus grand que la richesse lexicale est grande.

Nous présentons dans le tableau suivant pour chaque nuit, dans la partie gauche et dans l'ordre, sa longueur (N), l'étendue de son vocabulaire (V), cette dernière valeur portée à la puissance α (V^{α}), la réciproque de celle-ci ($V^{-\alpha}$) et en fin la valeur de W. Dans la partie droite du tableau, nous présentons la richesse lexicale des nuits classées par ordre croissant selon l'indice W (variation inverse oblige) et par ordre décroissant selon les valeurs de cet indice ramenées à R.

Richesse lexicale des nuits d'*al-ÞImtâÝ wa l-muÞânasa* selon l'indice W de Brunet

	N	***V***	V^{α}	$V^{-\alpha}$	$W = N^{V^{-\alpha}}$
Nuit 00	5062	1412	3,4815	0,2872	11,5875
Nuit 01	2478	703	3,0880	0,3238	12,5640
Nuit 02	3115	896	3,2196	0,3106	12,1639
Nuit 03	2004	627	3,0278	0,3303	12,3174
Nuit 04	4248	1133	3,3522	0,2983	12,0875
Nuit 05	906	340	2,7253	0,3669	12,1637
Nuit 06	7079	1644	3,5738	0,2798	11,9472
Nuit 07	2569	688	3,0766	0,3250	12,8323
Nuit 08	10788	1967	3,6858	0,2713	12,4216
Nuit 09	4607	1003	3,2826	0,3046	13,0616
Nuit 10	9564	1772	3,6202	0,2762	12,5764
Nuit 13	2427	535	2,9463	0,3394	14,0904
Nuit 14	3271	819	3,1702	0,3154	12,8430
Nuit 15	1773	510	2,9221	0,3422	12,9346
Nuit 16	1286	420	2,8262	0,3538	12,5939

Classement		
	W	***R***
Nuit 00	11,5875	0,8942
Nuit 06	11,9472	0,8702
Nuit 04	12,0875	0,8608
Nuit 05	12,1637	0,8558
Nuit 02	12,1639	0,8557
Nuit 03	12,3174	0,8455
Nuit 08	12,4216	0,8386
Nuit 01	12,5640	0,8291
Nuit 10	12,5764	0,8282
Nuit 16	12,5939	0,8271
Nuit 07	12,8323	0,8112
Nuit 14	12,8430	0,8105
Nuit 15	12,9346	0,8044
Nuit 09	13,0616	0,7959
Nuit 13	14,0904	0,7273

Tableau 71

À la tête du classement de la richesse lexicale des nuits selon l'indice W de Brunet, nous trouvons, à l'image des deux premières méthodes, le préambule (la nuit 0) comme étant la partie la plus riche lexicalement de notre corpus. Pareillement, avec cette méthode, la nuit 6 confirme la place qu'elle a occupée avec la méthode Guiraud. Quant à la troisième place, elle revient à la nuit 4 qui occupait la cinquième place avec la méthode Guiraud.

En bas du classement, c'est toujours notre treizième nuit qui est reléguée à la place de la plus pauvre lexicalement de toutes les nuits. Elle est précédée par les Nuits 9 et 15, respectivement aux pénultième et antépénultième rangs.

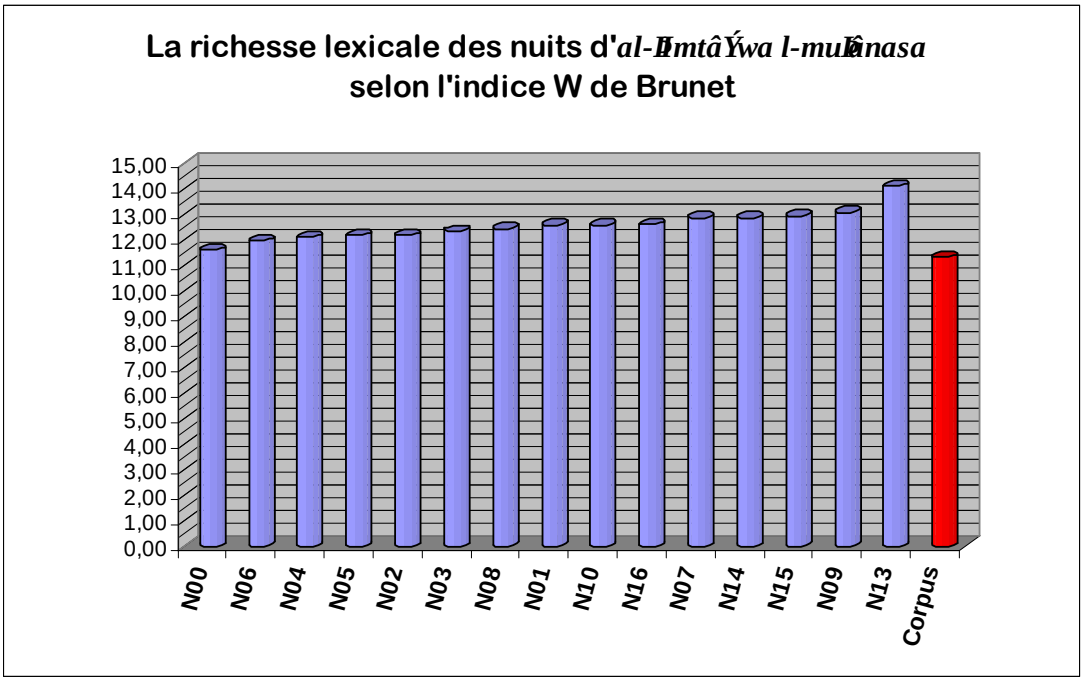


Figure 112

Quant aux nuits occupant les rangs du milieu, ce sont dans l'ordre, la nuit 8, la première et la dixième nuits occupant les septième, huitième et neuvième rangs ; elles occupaient respectivement, la troisième, la neuvième et la quatrième places avec la méthode Guiraud.

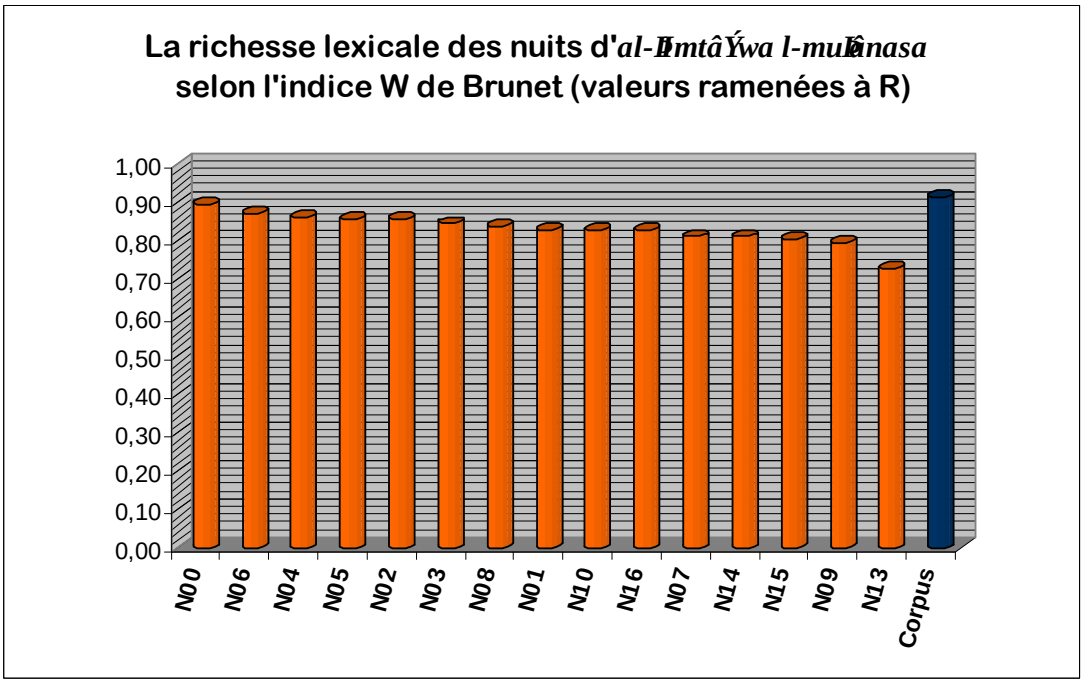


Figure 113

Dans la perspective de la réduction de l'influence de l'étendue du texte sur la richesse lexicale, nous pouvons affirmer sans réserve que l'indice W de Brunet a bien atteint son objectif avec un corpus comme le nôtre[299] qui a une étendue de plus de 61 000 occurrences et des parties dont l'étendue varie de 1 à 10. En effet, il suffit de remarquer, par exemple, que les nuits 8 et 10 qui ont, et de loin, les deux plus grandes étendues de toutes les nuits (10 788 et 9 564 occurrences) se sont vues reléguées, au niveau de la richesse lexicale, aux septième et neuvième rangs. Cette affirmation est facilement vérifiable en ayant recours au coefficient de corrélation des rangs de Spearman. Ce coefficient est de 0,257, ce qui est très peu pour pouvoir parler de quelconque corrélation entre les valeurs de l'indice W et celles de l'étendue N des nuits.

Au seuil de signification $\alpha = 0,05$, on ne peut donc pas rejeter l'hypothèse nulle d'absence de corrélation. Autrement dit, la corrélation n'est pas significative, et la dispersion des points sur le graphique de dispersion des rangs (figure 7) le montre très clairement.

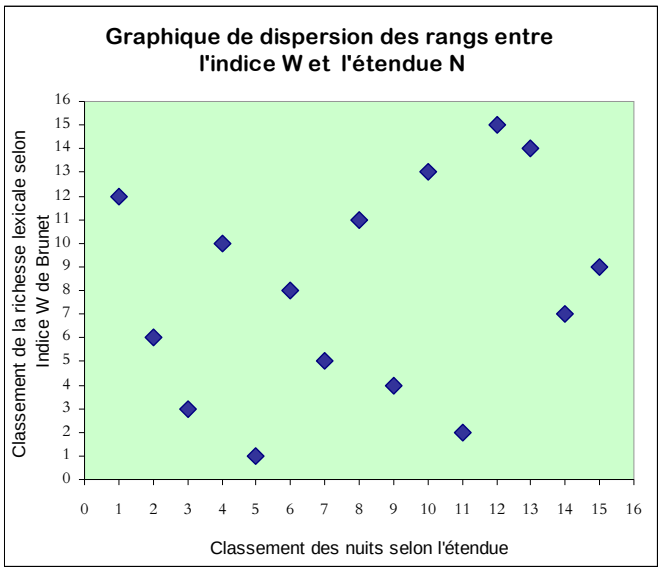


Figure 114

[299] André Cossette a appliqué l'indice W à des corpus tels que le TLF, Giraudoux, Proust, Chateaubriand, Hugo et Zola dont l'étendue varie de 671 364 à 37 652 402 occurrences et a remarqué une stabilité presque parfaite de cet indice. Voir : A. Cossette, *op. cit.*, p. 148.

Au niveau de la répartition des nuits de part et d'autre de la richesse lexicale moyenne, qui est ici de 0,8303, elles sont équitablement réparties en un groupe ayant une richesse lexicale supérieure à la moyenne composé des nuits 0, 2, 3, 4, 5, 6 et 8 ; et un groupe ayant une richesse lexicale inférieure à la moyenne composé des nuits 1, 7, 9, 10, 13, 14, 15 et 16.

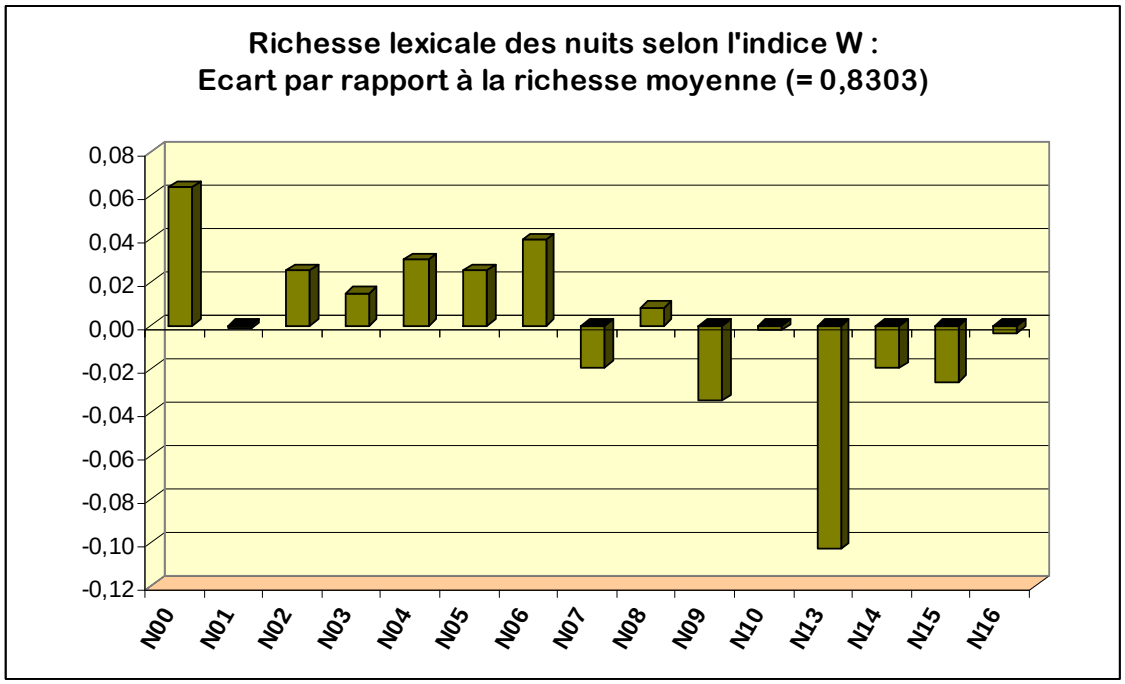


Figure 115

Les nuits ayant la richesse lexicale la plus éloignée positivement de la richesse moyenne sont la nuit 0, la nuit 6 et la nuit 4. Alors que celles ayant la richesse lexicale la plus éloignée négativement sont la nuit 13, la nuit 9 et la nuit 15.

Dans les nuits à richesse lexicale moyenne, la nuit 1, même si elle s'écarte négativement de la moyenne, reste quand même collée à celle-ci. Les nuits 10 et 16 s'en écartent aussi légèrement et négativement. La nuit 8, quant à elle, s'en écarte positivement et, en valeur absolue, plus que les trois précédentes.

1.4. La méthode binomiale de Muller

L'approche que propose Charles Muller pour mesurer et comparer la richesse lexicale de deux textes repose sur le raisonnement suivant :
Devant deux textes A et B d'étendues différentes ($N_A > N_B$), quel serait le vocabulaire de A si on l'avait arrêté à la longueur de B ?

Autrement dit, cette méthode, qui repose sur le calcul des probabilités, cherche à évaluer l'étendue du vocabulaire théorique (V') qu'aurait le texte le plus long s'il était réduit à l'étendue du plus court. Dans un deuxième temps, ce vocabulaire attendu est comparé au vocabulaire réel pour apprécier, en fin de compte, l'écart entre les deux.

Sans chercher à supprimer arbitrairement des morceaux du texte à réduire, ni à effectuer des échantillonnages aléatoires laborieux, le calcul des probabilités permet d'opérer de manière rationnelle la réduction du texte recherchée.

Dans ce sens, en effet, chaque occurrence du texte A a la probabilité $p = \frac{N_B}{N_A}$ de figurer dans le texte réduit et la probabilité complémentaire $q = \frac{N_A - N_B}{N_A} = 1 - \frac{N_B}{N_A}$ de ne pas y figurer, c'est-à-dire de figurer dans la partie éliminée.

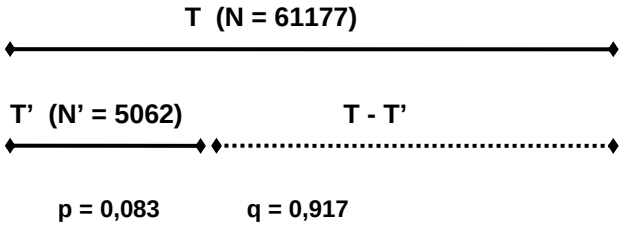
Il faut rappeler ici que l'étendue du vocabulaire ne s'exprime pas seulement par la valeur V qui est le nombre des vocables d'un texte, mais peut aussi s'exprimer par la somme des fréquences de chacun des éléments du vocabulaire :

$V_1 + V_2 + V_3 \cdots + V_n = V$ ou encore : $V = \sum_{i=1}^{n} V_i$

où

V_1 : effectif des vocables ayant **1** occurrence
V_2 : effectif des vocables ayant **2** occurrences
V_3 : effectif des vocables ayant **3** occurrences
… : … … … … … …
V_n : effectif des vocables ayant **n** occurrences

Considérons, par exemple, notre corpus désigné par **T** qui a l'étendue N = 61177 et que nous voulons réduire à la nuit 0 désignée par **T'** qui a l'étendue N' = 5062. Chaque occurrence d'un vocable a la probabilité $p = \frac{5062}{61177} = 0{,}083$ de se trouver dans **T'** et la probabilité contraire $q = 1 - \frac{5062}{61177} = 0{,}917$ de ne pas s'y trouver.



Le modèle théorique adopté par cette méthode, basé sur la loi binomiale est un enchaînement des termes du binôme $(p+q)^i$ qui correspondent à la probabilité de la distribution des occurrences de vocables pour chaque classe de ces derniers : $(p+q)$ pour les V_1, $(p+q)^2$ pour les V_2, ... $(p+q)^n$ pour les V_n. L'espérance mathématique, notée E(V) ou μ est la moyenne pondérée des vocables de l'ensemble du texte T, avec comme poids de pondération les probabilités $(p+q)^i$; elle est donc calculée en multipliant chaque classe de vocables par sa probabilité et en additionnant, à la fin, tous les produits :

$$\mu = \mathrm{E(V)} = (p+q)\mathrm{V}_1 + (p+q)^2\mathrm{V}_2 + (p+q)^3\mathrm{V}_3 \ldots + (p+q)^n\mathrm{V}_n$$

Désignons maintenant par :
V' **:** le nombre de vocables attendus théoriquement en T'
V_0' **:** le nombre de vocables théoriquement absents de T'
V **:** le nombre de vocables rencontrés réellement dans T ;

nous avons : $V = V' + V_0' \Rightarrow V' = V - V_0'$

Étant donné que $V' = V - V_0'$, pour calculer donc l'espérance mathématique (*i.e.* le nombre) des vocables attendus dans le texte réduit T', c'est-à-dire V', Charles Muller propose de calculer d'abord le nombre de vocables qui seraient absents de T', c'est-à-

dire V_0', et le soustraire par la suite au nombre total des vocables de T, c'est-à-dire V. Ce qui se traduit par les formules suivantes :

$$E(V_0') = qV_1 + q^2V_2 + q^3V_3 \cdots + q^nV_n$$

$$E(V_0') = \sum_{i=1}^{n} q^iV_i$$

$$V' = V - \sum_{i=1}^{n} q^iV_i$$

Ces formules vont donc nous permettre de calculer le vocabulaire théorique de chacune des nuits.

Reprenons notre exemple concernant la réduction de notre corpus (T) à la nuit 0 (T'), nous avons calculé plus haut p = 0,083 et q = 0,917. Pour calculer le vocabulaire théoriquement présent V', nous allons d'abord calculer le vocabulaire théoriquement absent de T' c'est-à-dire V_0'.

$E(V_0') = (0{,}917 \times 3443) + ((0{,}917)^2 \times 1109) + ((0{,}917)^3 \times 563) + \cdots + ((0{,}917)^{7825} \times 1)$

$E(V_0') = (3158) + (933) + (434) + \cdots + (3{,}0934\ (10)^{-294})$

$E(V_0') = 5254{,}19$

d'où **V'** = V - V_0' = 6652 - 5254,19 = **1397,81**

Nous aurons, en outre, besoin de calculer l'écart-type théorique qui est la racine carrée positive de la variance théorique. Les formules de la variance et de l'écart-type théoriques sont les suivantes :

$\sigma^2 = npq$ d'où : $\sigma = \sqrt{npq}$

où

σ^2	:	la variance théorique
σ	:	l'écart-type théorique
n	:	nombre de vocables (= V)
p	:	probabilité des vocables d'être absents (= V_0'/
q	:	probabilité inverse

[*] Il s'agit ici de la probabilité d'un ***vocable*** d'être absent de l'ensemble des vocables V et non de celle d'une ***occurrence*** par rapport à l'ensemble des occurrences N.

Appliquée à notre T' (la nuit 0), cette formule nous donne un écart-type théorique de :

$$\sigma = \sqrt{6652 \times 0{,}79 \times 0{,}21} = \sqrt{1104} = 33{,}23$$

Nous avons maintenant calculé le vocabulaire théoriquement présent (1397,81) ainsi que l'écart-type théorique, sachant que le vocabulaire réel de la nuit 0 est de 1412, nous calculons l'écart absolu entre les deux, nous obtenons 14,19. Nous divisons, à présent cet écart absolu par l'écart-type théorique qui a, pour cette nuit, la valeur de 33,23, nous obtenons l'écart réduit qui a donc la valeur de 0,43. Pour obtenir l'écart translaté nous ajoutons à la valeur de cet écart réduit le nombre 35, ce qui donne 35,43. Toutes ces valeurs, nous les trouvons inscrites dans la première ligne du tableau suivant et qui correspondent à la nuit 0.

Richesse lexicale des nuits d'*al-ÞImtâÝ wa l-muÞânasa* selon la loi binomiale (méthode Muller)

	N	V réelle	V' théorique	Ecart absolu	σ	Ecart réduit	Ecart translaté
Nuit 00	5062	1412	1397,81	14,19	33,23	0,43	35,43
Nuit 01	2478	703	818,02	- 115,02	26,78	- 4,29	30,71
Nuit 02	3115	896	974,36	- 78,36	28,84	- 2,72	32,28
Nuit 03	2004	627	693,82	- 66,82	24,93	- 2,68	32,32
Nuit 04	4248	1133	1229,17	- 96,17	31,66	- 3,04	31,96
Nuit 05	906	340	369,14	- 29,14	18,67	- 1,56	33,44
Nuit 06	7079	1644	1776,87	- 132,87	36,09	- 3,68	31,32
Nuit 07	2569	688	841,04	- 153,04	27,11	- 5,65	29,35
Nuit 08	10788	1967	2373,00	- 406,00	39,07	- 10,39	24,61
Nuit 09	4607	1003	1304,86	- 301,86	32,39	- 9,32	25,68
Nuit 10	9564	1772	2187,81	- 415,81	38,32	- 10,85	24,15
Nuit 13	2427	535	1750,85	- 1215,85	35,92	- 33,85	1,15
Nuit 14	3271	819	1011,07	- 192,07	29,28	- 6,56	28,44
Nuit 15	1773	510	630,37	- 120,37	23,89	- 5,04	29,96
Nuit 16	1286	420	488,85	- 68,85	21,28	- 3,24	31,76

Classement

	Ecart translaté
Nuit 00	35,43
Nuit 05	33,44
Nuit 03	32,32
Nuit 02	32,28
Nuit 04	31,96
Nuit 16	31,76
Nuit 06	31,32
Nuit 01	30,71
Nuit 15	29,96
Nuit 07	29,35
Nuit 14	28,44
Nuit 09	25,68
Nuit 08	24,61
Nuit 10	24,15
Nuit 13	1,15

Tableau 72

Comme nous l'avons fait pour la nuit 0, traitée dans les lignes précédentes, à partir des données du tableau des distributions des fréquences des nuits, nous avons mesuré, selon les formules annoncées plus haut, la part du vocabulaire théoriquement

absent dans chacune des nuits, puis celle du vocabulaire théoriquement présent. Ce vocabulaire attendu est par la suite comparé à celui qu'on trouve en réalité, le vocabulaire observé, et ce dans le but d'évaluer la distance absolue entre les deux. Cet écart absolu est ramené à un écart réduit en le divisant par l'écart-type théorique calculé pour chaque nuit. Pour mieux apprécier cet écart réduit qui comporte des valeurs négatives, nous avons translaté toutes les valeurs en leur ajoutant un même nombre qui soit supérieur à leur plus grande valeur absolue (la plus grande valeur absolue étant de 33,85 , nous avons choisi d'ajouter 35), et c'est cet écart translaté qui représente la richesse lexicale selon la loi binomiale et sur la base duquel nous pouvons comparer les nuits et les classer en fonction de leur richesse lexicale. Ce sont les résultats de toutes ces opérations que nous résume le tableau 7.

En effet, dans le tableau 7, nous avons présenté pour chaque nuit, dans la partie gauche et dans l'ordre, l'étendue de la nuit (N), l'étendue réelle du vocabulaire (V), l'étendue théorique du vocabulaire présent (V')[300] qui est obtenue par la formule $V' = V - \sum_{i=1}^{n} q^i V_i$, l'écart absolu qui est égal à (V - V'), l'écart-type théorique (σ), l'écart réduit qui est égal à $\frac{\text{écart absolu}}{\text{écart - type}}$, et enfin l'écart translaté qui est égal à (l'écart réduit + 35). Dans la partie droite, nous avons présenté le classement des nuits en fonction de la richesse lexicale et basé sur l'écart translaté.

Une remarque importante doit être faite ici concernant les étendues du vocabulaire théorique et celles du vocabulaire réel : exception faite du préambule (nuit 0), toutes les étendues du vocabulaire théorique sont supérieures à celles du vocabulaire réel (Voir figure 9). C'est pour cette raison que, dans le tableau 7, tous les écarts (absolus et réduits) sont négatifs, en dehors bien sûr de celui de la nuit 0. Cela confirme bien une tendance générale qui concerne quasiment tous les textes.

[300] Par souci d'alléger le tableau, nous n'avons pas voulu présenter l'étendue théorique du vocabulaire absent (V_0') que nous avons utilisé comme étape intermédiaire pour le calcul ce V' et calculé en appliquant la formule $E(V_0') = \sum_{i=1}^{n} q^i V_i$.

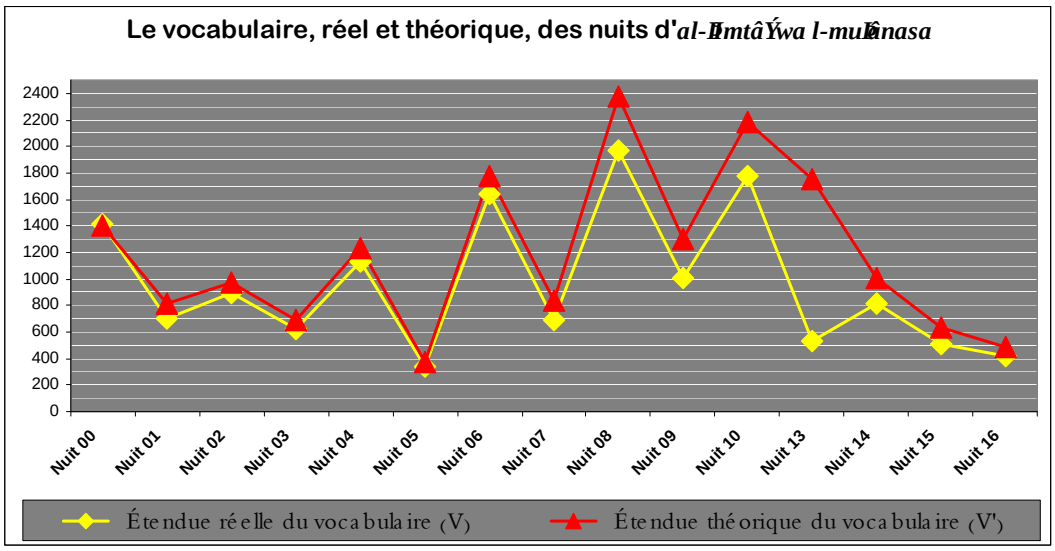


Figure 116

La figure 10 met en représentation graphique le classement des nuits en fonction de la richesse lexicale selon la loi binomiale, classement présenté en chiffres dans la partie droite du tableau 7.

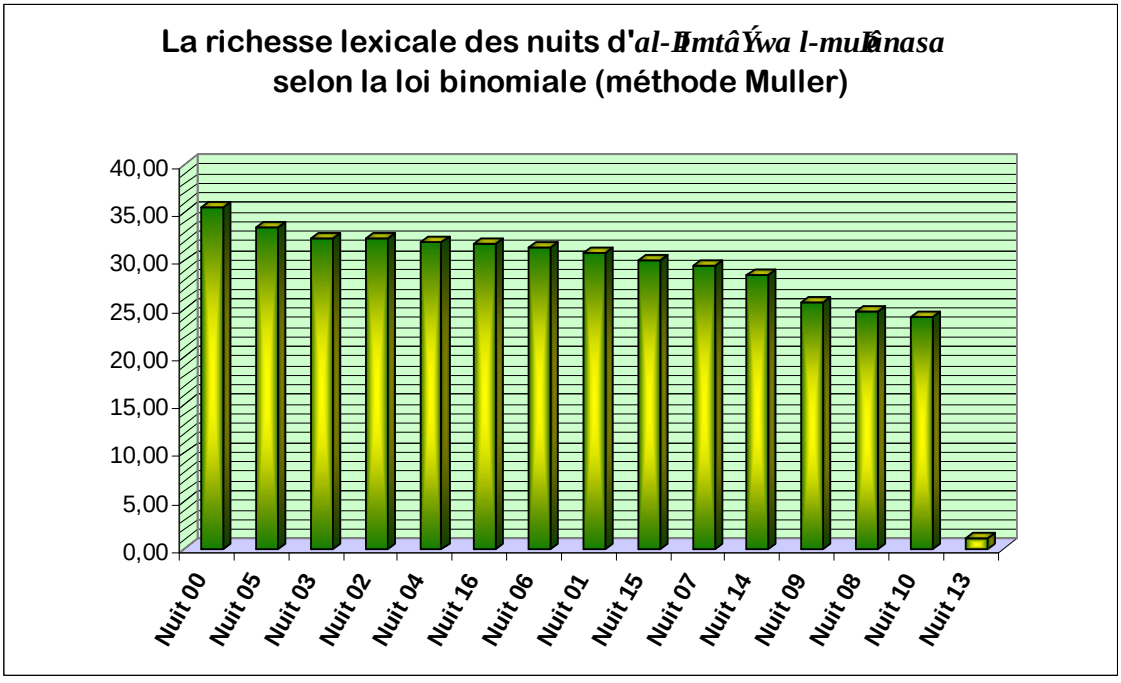


Figure 117

Dans le peloton de tête des nuits les plus riches lexicalement, s'installent les nuits 0, 5 et 3. La nuit 0 renforce sa place de nuit la plus riche lexicalement que toutes

les autres et ce avec les trois méthodes vues jusque là, la méthode Guiraud, l'indice W de Brunet en plus de la loi binomiale dont nous sommes en train d'évaluer les résultats. La nuit 5 que l'indice de Guiraud a placée en avant-dernier rang, et l'indice W de Brunet en cinquième position, occupe ici la deuxième meilleure place au niveau de la richesse lexicale. Le même parcours est remarqué pour la nuit 3 qui passe de la dixième place (Guiraud) à la sixième (Brunet) pour occuper le troisième rang ici.

En ce qui concerne le peloton de queue des nuits les plus pauvres lexicalement, aucune surprise pour la nuit 13 qui confirme bien cette dernière position. La grande surprise vient plutôt des deux nuits ayant les deux plus grandes étendues à savoir la nuit 8 et la nuit 10 qui occupent, respectivement, l'antépénultième et la pénultième places et qui avaient la troisième et la quatrième places selon l'indice de Guiraud, et la septième et la neuvième place selon l'indice W de Brunet ; nous reviendrons sur cette constatation plus bas.

Le groupe des nuits occupant les rangs du milieu quant à lui, il est composé des nuits 6, 1 et 15. Après avoir occupé la 9^{e} place (Guiraud), la nuit 1 a très légèrement varié puisqu'elle est passée à la 8^{e} (Brunet) pour confirmer cette position avec la méthode Muller. La nuit 15, après un mauvais classement selon Guiraud et Brunet (12^{e} et 13^{e} places), passe au neuvième rang selon la loi binomiale de Muller. Le parcours inverse est remarqué pour la nuit 6 qui, après une bonne deuxième place consensuelle entre les deux indices Guiraud et Brunet, est reléguée à la 7^{e} place selon Muller.

À la constatation évoquée plus haut concernant les nuits les plus longues placées en antépénultième et pénultième positions au niveau de la richesse lexicale, s'ajoute celle selon laquelle la nuit la plus courte (la nuit 5) occupe la deuxième place et la quatrième nuit la plus courte (la nuit 3) occupe la troisième place au niveau de la richesse lexicale. Ces observations montrent bien le fait que la mesure de la richesse lexicale selon cette méthode, la loi binomiale, est totalement indépendante de l'influence de la longueur des textes.

Comme pour l'indice W de Brunet et la formule de Guiraud, nous pouvons vérifier cette affirmation en nous appuyant sur le coefficient de corrélation des rangs de

Spearman entre le classement selon l'étendue des nuits et celui selon la richesse lexicale. Pour la méthode binomiale, ce coefficient est de (- 0,332), il est même négatif ; c'est-à-dire qu'il n'y a pas de corrélation directe entre les valeurs de la richesse et celles de l'étendue N des nuits.

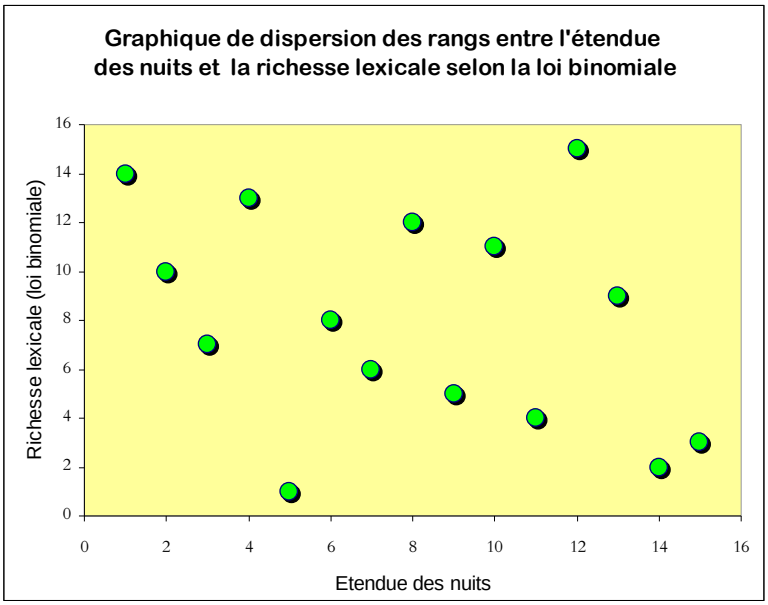


Figure 118

Au seuil donc de signification $\alpha = 0,05$, on ne peut pas rejeter l'hypothèse nulle d'absence de corrélation directe et on ne peut que déclarer la corrélation non significative dans la perspective de notre problématique. Le graphique de la figure 11 montre le nuage, sporadique, des points-nuits dans cette dispersion des rangs avec une tendance à peine dévoilée de corrélation inverse, mais toujours non significative, traduite par un coefficient de corrélation négatif mais faible.

Par ailleurs, au niveau de la répartition des nuits autour de la richesse lexicale moyenne, c'est la méthode binomiale qui place le plus de nuits au-dessus de la moyenne, qui est ici de 28,17, avec 11 nuits contre seulement quatre ayant une richesse lexicale inférieure à cette moyenne et qui sont les nuits 8, 9, 10 et 13.

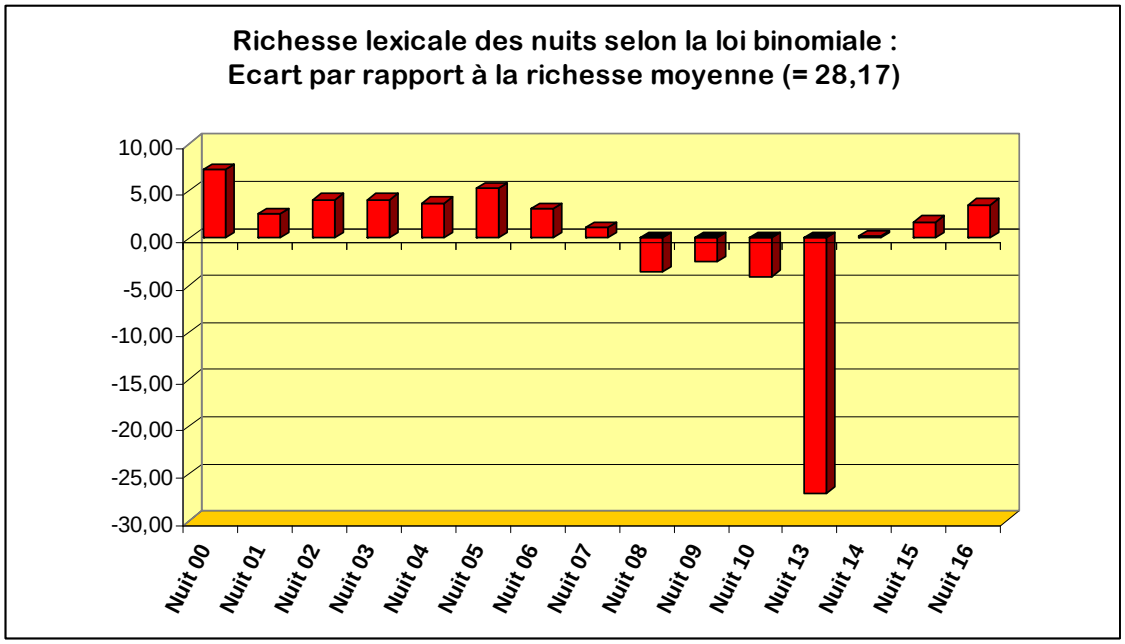


Figure 119

Les nuits ayant une richesse lexicale s'écartant le plus, positivement, de la moyenne sont la nuit 0, la nuit 5 et le nuit 2.

Le plus frappant dans ce graphique, c'est la distance négative exceptionnelle qui existe entre la valeur de richesse lexicale de la nuit 13 et celle de la richesse moyenne. Cette situation a été constatée aussi, avec moins de flagrance, *infra* dans la mesure de richesse lexicale selon l'indice W de Brunet.

La richesse lexicale la plus proche de la moyenne c'est celle de la nuit 14, puis celle de la nuit 7 toutes deux, au-dessus de la moyenne. La nuit 9 a la richesse lexicale la plus proche, négativement, de la richesse lexicale moyenne.

1.5. L'indice V_m de Yule-Herdan

L'indice V_m de Yule-Herdan, ou indice de **V**ariation de la **m**oyenne, est en fait une amélioration apportée en 1955 par Gustav Herdan à la « caractéristique » **K** élaborée en 1944 par G. Udny Yule et considérée, par bon nombre de spécialistes dont Herdan, comme pas sûre, pas stable et dépendante de l'étendue des textes.

Nous n'avons, en effet, pas retenu la « caractéristique » K parce qu'elle ne représente pas un indice de richesse lexicale fiable. Son utilité réside plutôt dans sa capacité à mesurer la concentration de la fréquence autour des vocables, ce qui n'est pas négligeable du point de vue stylistique mais qui reste tout de même en deçà de la mesure de la richesse lexicale globale attendu de ce type d'indice.

L'indice V_m , qui est en fait un coefficient de variation de la fréquence, se présente sous la formule suivante :

$$V_m = \frac{\nu_f}{\sqrt{V}}$$

Au niveau de la mise en pratique et des opérations nécessaires au calcul, cet indice est considéré comme le plus laborieux de tous les indices de richesse lexicale.

Pour décrire la formule et rendre compte des démarches de l'élaboration de l'indice V_m , nous partons, pour chaque nuit, de la fréquence moyenne $\overline{f}$ qui représente le quotient $\frac{N}{V}$. Comme V croît moins vite que N, alors $\overline{f}$ croît avec l'étendue des nuits.

À partir de cette fréquence moyenne, nous calculons l'écart-type qui lui est associé σ_f. L'écart-type est, par définition, la racine carrée positive de la variance :

$$\sigma_f = \sqrt{\frac{\sum_{i=1}^{n}(f_i - \bar{f})^2}{V}}$$

Pour ce faire, nous devons donc avoir toutes les fréquences $\bar{f}_i$ des vocabulaires V_i de chaque nuit que nous prendrons du tableau des distributions des fréquences des nuits.

Nous calculons ensuite l'indice de variation de fréquence $\boldsymbol{\nu_f}$ qui n'est autre que le rapport de l'écart-type à la moyenne :

$$\nu_f = \frac{\sigma_f}{\bar{f}}$$

Étant donné que la moyenne croît à l'allure de V et que l'écart-type croît à l'allure de N, alors le rapport de l'écart-type à la moyenne, l'indice de variation de fréquence $\boldsymbol{\nu_f}$ est fortement influencé par l'étendue des textes. Pour neutraliser cette influence, l'indice de variation de fréquence doit être divisé par la racine carrée de V ; d'où la formule finale de l'indice de variation de la moyenne $\boldsymbol{V_m}$:

$$V_m = \frac{\nu_f}{\sqrt{V}}$$

Nous présentons donc, dans le tableau 8 suivant, les résultats obtenus suite aux opérations que nous avons faites pour le calcul de l'indice $\boldsymbol{V_m}$ et que nous venons de décrire. Devant chaque nuit, nous avons inscrit, dans l'ordre, N, V, V_1, $\overline{f}$, $\boldsymbol{\sigma_f}$, $\boldsymbol{\nu_f}$, et enfin $\boldsymbol{V_m}$.

Richesse lexicale des nuits d'*al-ÞImtâÝ wa l-muÞânasa*
selon l'indice V_m de Yule-Herdan

	N	V	V_1	$\overline{f}$	σ_f	ν_f	V_m
Nuit 00	5062	1412	1048	3,58	28,28	7,89	0,210
Nuit 01	2478	703	498	3,52	19,06	5,41	0,204
Nuit 02	3115	896	649	3,48	22,02	6,33	0,212
Nuit 03	2004	627	470	3,20	18,85	5,90	0,236
Nuit 04	4248	1133	834	3,75	25,00	6,67	0,198
Nuit 05	906	340	265	2,66	14,35	5,39	0,292
Nuit 06	7079	1644	1140	4,31	28,67	6,66	0,164
Nuit 07	2569	688	464	3,73	18,01	4,82	0,184
Nuit 08	10788	1967	1249	5,48	28,97	5,28	0,119
Nuit 09	4607	1003	637	4,59	20,64	4,49	0,142
Nuit 10	9564	1772	1033	5,40	25,49	4,72	0,112
Nuit 13	2427	535	342	4,54	14,97	3,30	0,143
Nuit 14	3271	819	546	3,99	19,54	4,89	0,171
Nuit 15	1773	510	344	3,48	15,53	4,47	0,198
Nuit 16	1286	420	313	3,06	15,28	4,99	0,244

Tableau 73

En lisant bien le tableau 8 et la figure 13, on ne peut que se rendre compte d'un avantage certain de cet indice qui n'apparaissait pas dans l'indice de Guiraud, par exemple, mais qu'on a pu remarquer également dans l'indice W de Brunet et la loi binomiale de Muller. À l'image de ces deux indices, l'avantage de l'indice $\boldsymbol{V_m}$ se manifeste bien dans notre corpus au niveau de l'amortissement de l'influence da la longueur (N) des nuits. C'est ce qui traduit le fait que les deux nuits les plus longues la nuit 8 (10788 occurrences) et la nuit 10 (9564 occurrences) se trouvent en bas du tableau, respectivement, à l'avant-dernière et la dernière place ; la nuit 9 avec ses 4607 occurrences occupe l'antépénultième place.

Cependant, il est vraiment difficile d'interpréter les valeurs de $\boldsymbol{V_m}$ quant à leur traduction de la richesse lexicale. L'avantage de l'amortissement de la longueur des textes dont nous venons de rendre compte se voit, en fait, contrecarré par deux

tendances liées à l'indice V_m et que nous exposons quelques lignes plus bas ; le tout entre alors dans une sorte de tiraillement et de conflit rendant très difficile, si ce n'est impossible, l'interprétation des valeurs de cet indice. C'est d'ailleurs ce qui a poussé N. Ménard à conclure à l'impossibilité de « tirer au clair les liens entre « V_m » et la richesse lexicale »[301] et A. Cossette à considérer que « l'étude de V_m montre aussi qu'il est dangereux de l'interpréter comme un indice de richesse lexicale »[302].

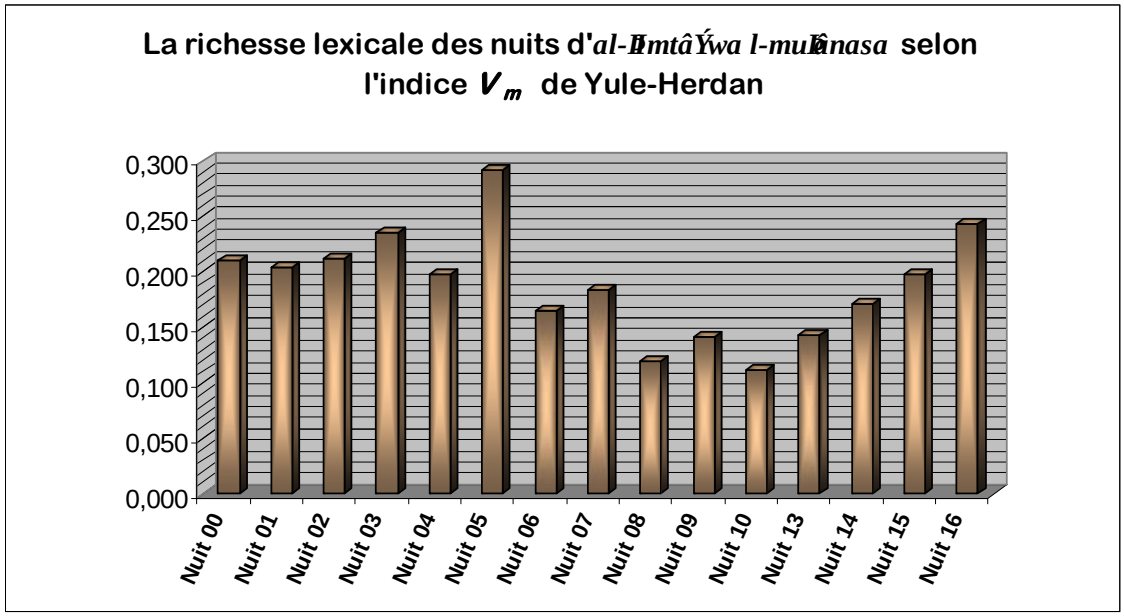


Figure 120

Les conflits de tendances internes à l'indice V_m ont créé, au niveau des résultats que nous avons obtenus pour notre corpus, des problèmes supplémentaires. En effet, par rapport à ce que nous avons observé avec les autres méthodes, nous remarquons un certain nombre de perturbations inattendues au niveau des valeurs de cet indice. Il est clair que le schéma qui s'est dessiné à petits pas par le biais des autres formules, en commençant par la méthode de comparaison des indices, se voit ici perturbé et les classements qu'on a pu obtenir bouleversés. Même les ébauches de classement obtenues par la première méthode et confirmées par la suite, aussi partielles et rudimentaires soient-elles, ne sont pas respectées par cette méthode.

Cela s'explique, nous semble-t-il, par le fait que, même s'il amortit bien les perturbations dues à l'étendue des textes, l'indice V_m révèle deux facteurs majeurs :

[301] Nathan Ménard, *op. cit.*, p. 77.

[302] André Cossette, *op. cit.*, p. 89.

- il croît fortement avec les vocables de fréquence inférieure à $\overline{f}$, et principalement avec ceux de fréquence 1 (V_1), les *hapax*.
- il croît rapidement avec l'effectif et la fréquence des vocables de fréquence bien supérieure à $\overline{f}$, les vocables de forte fréquence.

Le premier facteur est, nous semble-t-il, celui qui a influencé le plus l'indice V_m. Il est frappant de voir, à ce propos, que la proportion de V_1 à V dans les trois nuits placées en tête par cette méthode, les nuits 05, 16 et 03, varie entre 78% et 74%, alors qu'elle est au niveau de l'ensemble du corpus de l'ordre de 52%. C'est visiblement ce qui explique la croissance anormale de V_m pour ces nuits en les propulsant en haut du classement. Pour s'en convaincre, il suffit de comparer les deux classements, celui de $\boldsymbol{V_m}$ et celui du quotient $\frac{V_1}{V}$, que nous avons opposés dans la figure 14 suivante. Nous y voyons, à quelques rangs près, un alignement presque total des rangs des nuits dans les deux classements.

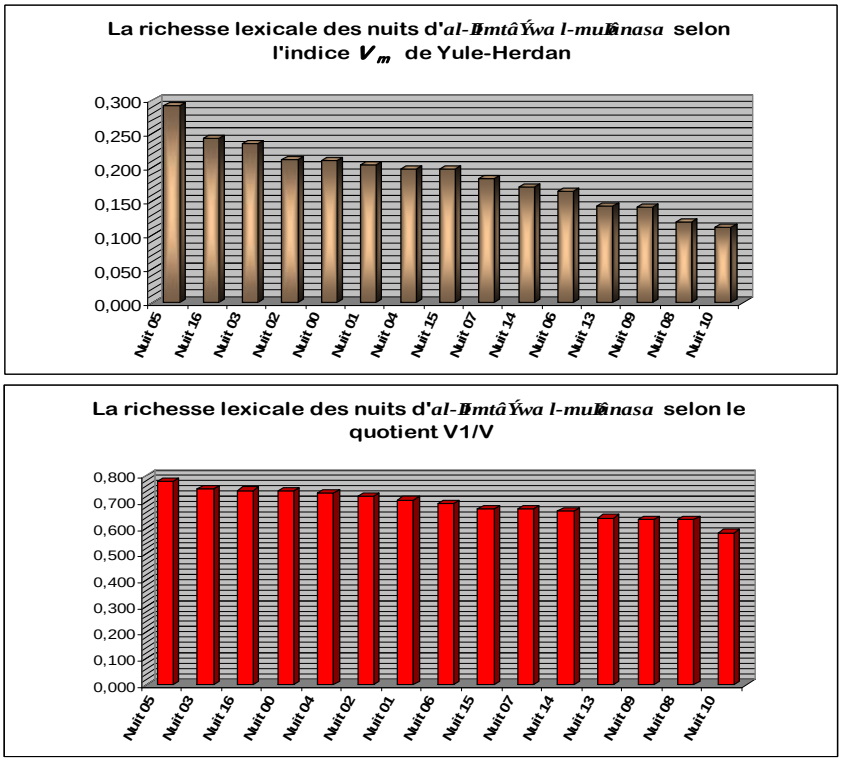


Figure 121
Confirmation de l'influence de V_1 sur l'indice V_m en comparant les deux diagrammes de la richesse lexicale, selon l'indice V_m et le quotient V_1/V.

En outre, le nombre des vocables de fréquence 1 (V_1) ne peut pas mesurer la richesse lexicale ; ce que mesurent ces *hapax* c'est plutôt l'excentricité ou l'hétérogénéité du vocabulaire. En effet, cela est bien vérifié dans nos trois nuits qui sont en tête du classement, et ce, nous semble-t-il, pour des raisons purement thématiques puisque dans chacune de ces nuits quasiment un seul grand thème est traité. La nuit 8 traite principalement de la comparaison de quelques savants et poètes, on y trouve précisément le fameux débat (*munâÛara*) entre le grammairien ÞAbu SaÝîd as-Sîrâfî et le philosophe Mattâ b. Yûnus , débat qui a été évoqué dans plusieurs livres mais TawÎîdî fut le seul à l'avoir relaté entièrement. Dans la nuit 9, on trouve le thème des caractères innés des animaux et des êtres humains. Alors que la nuit 10 traite des étrangetés chez les animaux.

Un autre phénomène important caractérise également négativement l'indice $\boldsymbol{V_m}$: il s'agit de l'inversement de l'influence de l'étendue du texte sur la richesse lexicale. En effet, à force de vouloir bien faire, l'indice $\boldsymbol{V_m}$ a inversé l'influence de l'étendue : voulant neutraliser cette influence, la formule de Yule-Herdan a divisé, comme nous l'avons vu plus haut, l'indice de variation de fréquence par la racine carrée de V, et ce faisant, la richesse lexicale a pris des valeurs inversement corrélées aux valeurs de l'étendue des nuits. Nous nous sommes rendu compte de cette corrélation inverse en appliquant le test de corrélation des rangs de Spearman qui nous a livré un coefficient négatif de (- 0,707) au seuil de signification $\alpha = 0{,}05$. Cette corrélation inverse est très claire dans la disposition du nuage de points dans la figure 15.

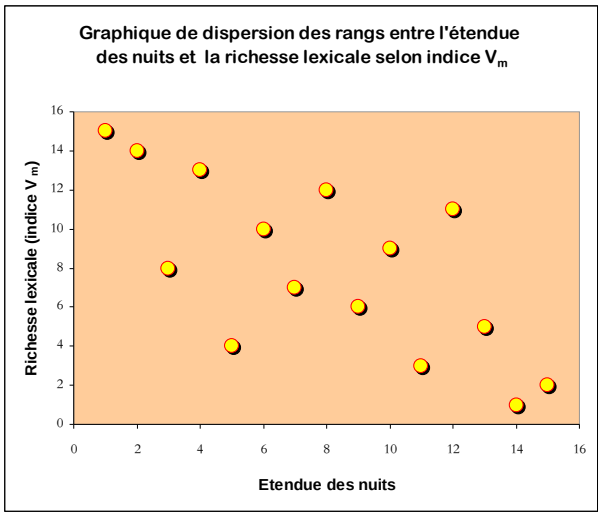


Figure 122

Cette corrélation inverse est également visible au niveau de la répartition des nuits autour de la richesse lexicale moyenne, qui est ici de 0,188, où l'on voit, par exemple, la distance notable par rapport à cette moyenne, négativement des valeurs de la richesse lexicale des nuits 8 et 10, et positivement des valeurs de la richesse lexicale des nuits 5 et 16.

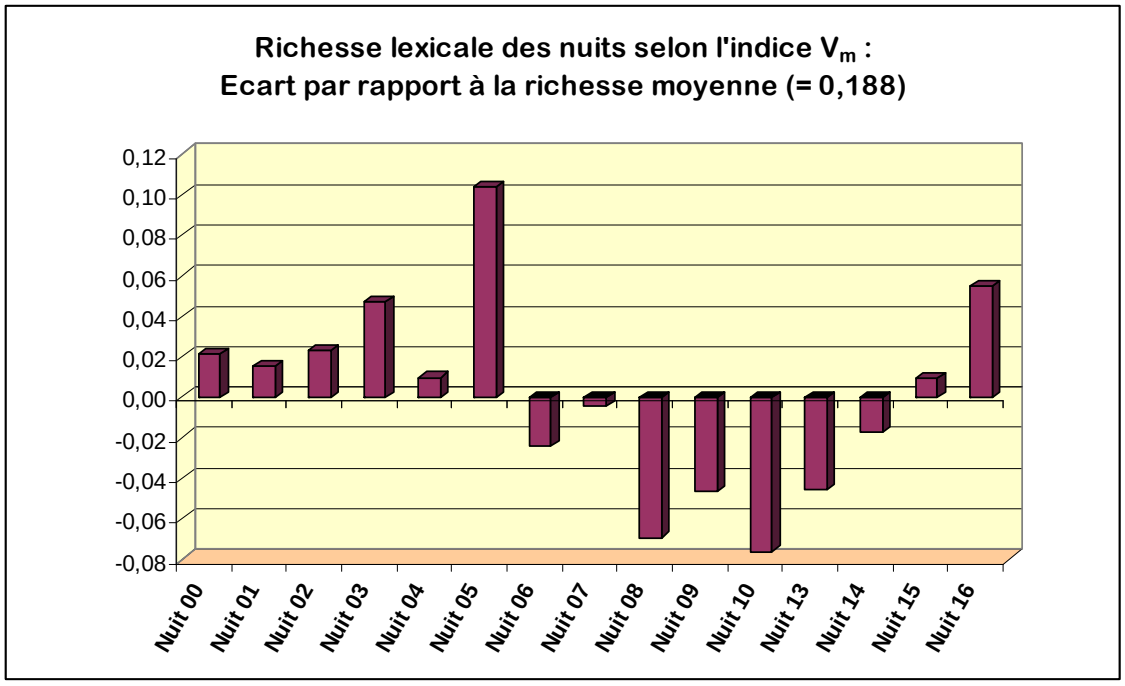


Figure 123

Les nuits ayant la richesse lexicale la plus proche de la moyenne sont, négativement la nuit 7 et , positivement les nuits 4 et 15.

2. Bilan

Nous avons pu, tout au long de ce chapitre, exposer, analyser et appliquer à notre corpus, cinq des plus importantes méthodes de mesure de la richesse lexicale.

Au terme de cette longue chevauchée exploratoire et contrastive, que doit-on retenir des résultats de ces méthodes ? Quelle(s) méthode(s) choisir ou suggérer pour la mesure de la richesse lexicale des textes arabes ? Peut-on (doit-on) toutes les utiliser conjointement ? Et dans ce cas, peut-on se baser sur ces différents classements pour en déduire un qui en soit, en quelque sorte, la résultante ? Avant d'essayer de répondre à ces questions, faisons d'abord une synthèse des résultats fournis par les différentes méthodes et des points communs aux différents classements.

Récapitulatif des classements des nuits d'*al-ÞImtâÝ wa l-muÞânasa* en fonction de la richesse lexicale selon les différentes méthodes utilisées

Nuit	Classement selon :			
	Guiraud ($V/\sqrt{N}$)	Brunet (*W*)	Muller (*Binomiale*)	Yule-Herdan (V_m)
Nuit 00	1	1	1	5
Nuit 01	9	8	8	6
Nuit 02	6	5	4	4
Nuit 03	10	6	3	3
Nuit 04	5	3	5	7
Nuit 05	14	5	2	1
Nuit 06	2	2	7	11
Nuit 07	11	11	10	9
Nuit 08	3	7	13	14
Nuit 09	7	14	12	13
Nuit 10	4	9	14	15
Nuit 13	15	15	15	12
Nuit 14	8	12	11	10
Nuit 15	12	13	9	8
Nuit 16	13	10	6	2

Tableau 74

En guise de récapitulatif, nous avons présenté dans le tableau 9, les classements des nuits d'*al-ÞImtâÝ wa l-MuÞânasa* en fonction de la richesse lexicale obtenus par quatre des cinq méthodes que nous avons retenues (la première méthode, la méthode de comparaison des indices, n'ayant pas fourni de classement complet). Pour une meilleure lecture du tableau, nous avons coloré en rouge les rangs qui sont communs, pour une nuit donnée, à deux ou trois méthodes (aucun rang n'a été commun aux quatre méthodes simultanément). Nous avons aussi coloré en jaune les rangs qui n'ont varié que très légèrement, c'est-à-dire les rangs contigus, autrement dit, quand une nuit donnée ayant un rang R_n selon une méthode, passe, selon une autre méthode soit au rang R_{n-1}, soit au rang R_{n+1}.

Selon ce tableau, l'on voit bien que seules les deux nuits 0 et 13 ont été placées, par trois méthodes sur quatre, à des positions stables qui sont respectivement, le premier et le dernier rang.

Aussi, y voit-on que les deux méthodes qui présentent le plus grand nombre de rangs communs et contigus réunis sont la méthode binomiale de Ch. Muller et l'indice V_m de Yule-Herdan avec deux rangs communs et sept rangs contigus ; nous verrons plus loin que ces deux méthodes présentent les classements les plus corrélés entre eux.

Les deux méthodes présentant le plus grand nombre de rangs communs sont la méthode de Guiraud et l'indice W de Brunet avec quatre rangs communs.

Si nous considérons maintenant le nombre total, pour chaque méthode, des rangs communs ou contigus à ceux de toutes les autres méthodes, nous obtiendrons le tableau suivant :

Nombre total, pour chaque méthode, des rangs communs et contigus à ceux des autres méthodes

	Communs	Ecart / Moy	Contigus	Ecart / Moy
Guiraud	7	*1*	3	*- 3*
Brunet	6	*0*	6	*0*
Muller	8	*2*	9	*3*
Yule-Herdan	2	*- 4*	7	*1*
Moyenne	***6***		***6***	

Tableau 75

La figure 17 de la page suivante nous montre bien les différentes courbes des classements des nuits d'*al-ÞImtâÝ wa l-MuÞânasa* obtenus par les quatre méthodes retenues.

Cette figure confirme graphiquement, les constatations que nous venons de faire. On y voit nettement, par exemple, les trois courbes Guiraud, Brunet et Muller se confondre aux deux points de coordonnées (Nuit 00, 1) et (Nuit 13, 15). Les deux courbes Guiraud et Brunet se confondent en quatre points (Nuit 00, 1), (Nuit 06, 2), (Nuit 07, 11) et (Nuit 13, 15). Si l'on observe bien les deux courbes Muller (courbe jaune) et Yule-Herdan (courbe bleu ciel), l'on se rend compte qu'elles suivent *grosso modo* la même trajectoire en se rejoignant en deux points (Nuit 02, 4) et (Nuit 03, 3), alors qu'elles font un départ, chacune de son côté, en deux points relativement éloignés l'un de l'autre et finissent leurs parcours par le même scénario.

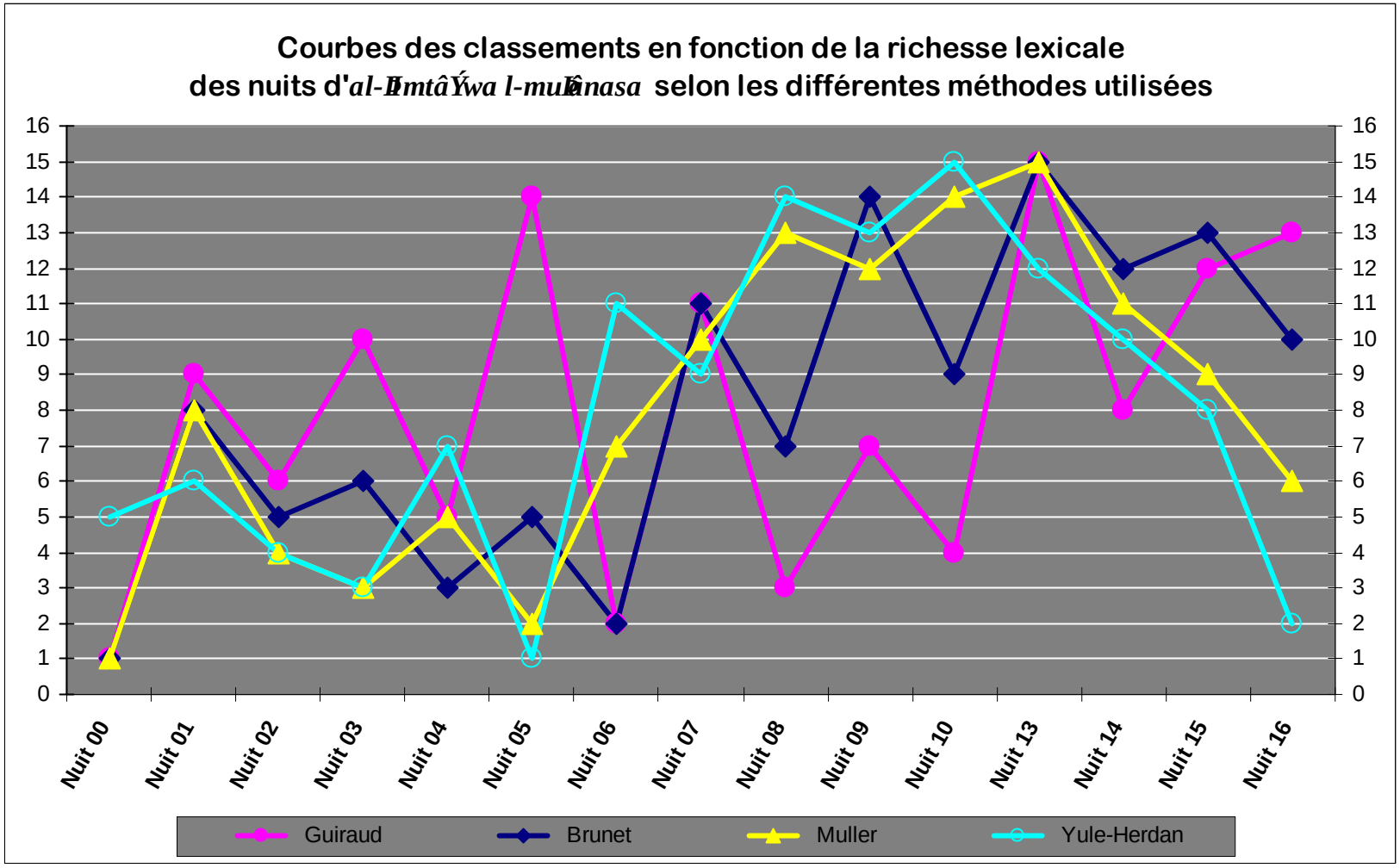

Courbes des classements en fonction de la richesse lexicale
des nuits d'*al-Imtâ' wa l-mu'ânasa* selon les différentes méthodes utilisées
16
15
14
13
12
11
10
9
8
7
6
5
4
3
2
1
0
Nuit 00
Nuit 01
Nuit 02
Nuit 03
Nuit 04
Nuit 05
Nuit 06
Nuit 07
Nuit 08
Nuit 09
Nuit 10
Nuit 13
Nuit 14
Nuit 15
Nuit 16
Guiraud
Brunet
Muller
Yule-Herdan

Figure 124

Une des constatations faites plus haut, concernant le grand nombre des rangs communs et contigus réunis entre les deux méthodes Muller et Yule-Herdan au niveau du tableau 9, se traduit, dans la figure 17, par les trajectoires presque identiques des deux courbes. Ce constat est quantifié dans le tableau suivant où les deux méthodes prennent la valeur la plus importante des coefficients de corrélation des rangs de Spearman 0,871, ce qui veut dire que ces deux classements sont les plus corrélés entre eux. Cette corrélation se traduit, au niveau du graphique, par le fait que les deux courbes ont quasiment la même allure.

En effet, nous avons utilisé le test de Spearman pour évaluer la corrélation des rangs et voir quels liens pourraient exister entre les différents classements selon les méthodes utilisées. Nous résumons ces coefficients de corrélation dans le tableau 11 avec, pour chacun, son interprétation quant à la corrélation entre la méthode en début de chaque ligne et celle à la tête de chaque colonne ; elle est soit significative, soit non significative au seuil de signification $\alpha = 0,05$.

Corrélation des rangs de Spearman entre les classements obtenus par les 4 principales méthodes :
Coefficient et interprétation

	Yule-Herdan	**Muller**	**Brunet**
Guiraud	- 0,364 Corrélation **non significative**	0,079 Corrélation **non significative**	0,590 Corrélation **significative**
Brunet	0,409 Corrélation **non significative**	0,744 Corrélation **significative**	
Muller	**0,871** Corrélation **significative**		

Tableau 76

Le coefficient de corrélation des rangs de Spearman a été détaillé plus haut dans un autre chapitre[303], il est donc inutile d'en exposer encore une fois ici le principe de fonctionnement, l'utilité et la méthode de calcul.

[303] Voir Chapitre 9 « *Les caractéristiques lexicométriques* ».

Au seuil de signification α = 0,05, on peut rejeter l'hypothèse nulle d'absence de corrélation, pour trois couples de méthodes (*Muller, Yule-Herdan*), (*Muller, Brunet*) et (*Brunet, Guiraud*) et conclure donc à une corrélation significative entre les rangs de chacune des méthodes à l'intérieur de ces trois couples, avec un coefficient de corrélation, respectivement, de 0,871, 0,744 et 0,590. La plus forte corrélation existe donc entre la méthode Muller et la méthode Yule-Herdan. La corrélation entre Brunet et Guiraud est à peine significative.

Au seuil de signification, également, α = 0,05, on ne peut pas rejeter l'hypothèse nulle d'absence de corrélation, pour les trois autres couples de méthodes (*Yule-Herdan, Guiraud*), (*Yule-Herdan, Brunet*) et (*Muller, Guiraud*), autrement dit, la corrélation n'est pas significative pour ces trois couples, avec un coefficient de corrélation, respectivement, de (- 0,364), 0,409 et 0,079.

2.1. Analyse factorielle

Les coefficients de corrélation des rangs présentent une information certes importante, mais ne suffisent pas, à eux seuls, de mieux analyser les différents classements et les facteurs qui les ont influencés, positivement ou négativement. C'est pourquoi nous avons fait appel à une méthode d'analyse factorielle qui est l'analyse factorielle des variables latentes combinée à une technique qui permet d'identifier une structure factorielle renfermant des contributions extrêmes des variables. Il s'agit de la technique de rotation orthogonale varimax. Il est à noter ici, qu'entre autres coefficients calculés, l'analyse factorielle des variables latentes utilise les coefficients de corrélation de Spearman, de Pearson ou de Kendall ; nous avons choisi ceux de Spearman que nous avons présentés plus haut dans le tableau 11.

En utilisant l'analyse factorielle des variables latentes avec rotation varimax, nous allons essayer de résumer, la structure de corrélation des classements effectués par les différentes méthodes de mesure de richesse lexicale, en identifiant des facteurs sous-jacents communs aux méthodes et pouvant expliquer une part importante de la

variabilité des classements. Pour cette analyse factorielle des variables latentes, nous avons utilisé le logiciel XLSTAT* version 7.5.3.

La rotation varimax est utilisée pour simplifier l'interprétation des facteurs en minimisant le nombre de variables qui ont des contributions élevées sur chaque facteur. Elle rend ainsi l'interprétation plus aisée en maximisant la variance du carré des coordonnées des variables par colonne.

Afin de prendre en compte plus d'information concernant la dispersion des points-Nuits, nous avons utilisé XLSTAT–3DPlot** pour visualiser, en 3 dimensions, la représentation des nuits en fonction des méthodes.

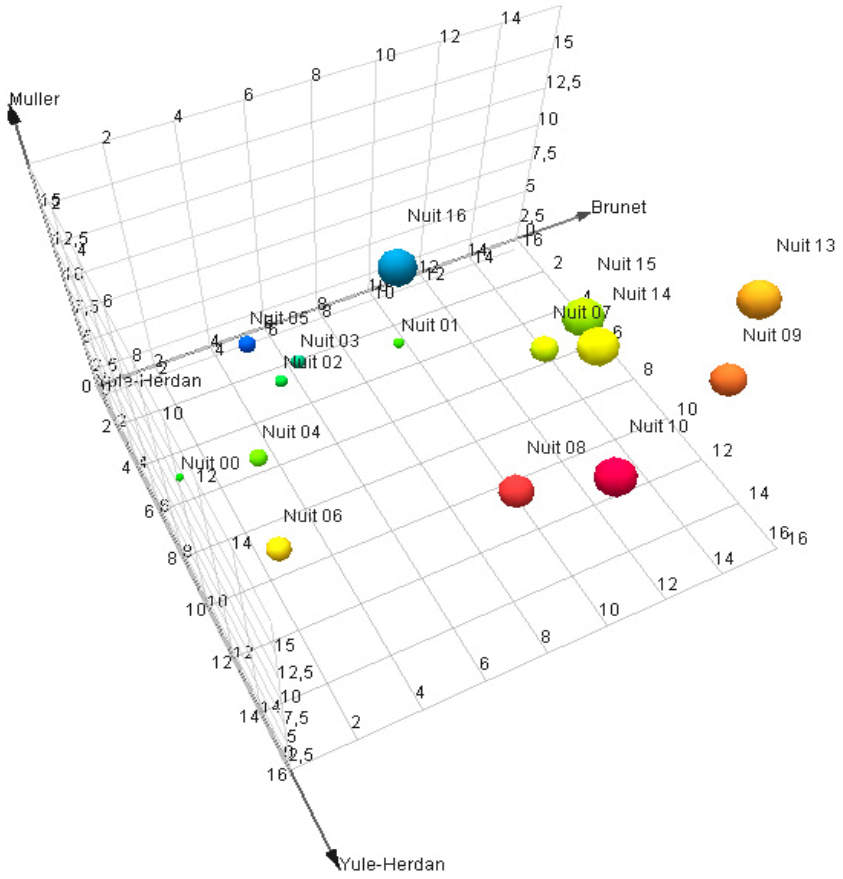


Figure 125
Représentation graphique, en 3 dimensions, du classement des nuits en fonction de la richesse lexicale selon les méthodes

* Le logiciel XLSTAT est une marque déposée de Addinsoft, http://www.xlstat.com

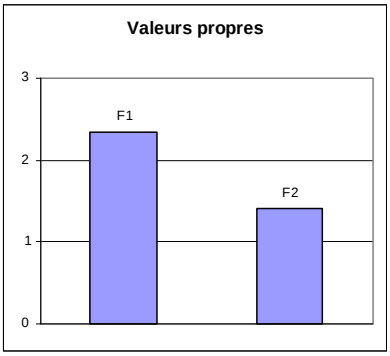
** XLSTAT-3DPlot est un module additionnel de XLSTAT, http://www.xlstat.com/fr/products/xlstat-3dplot/

L'analyse factorielle des variables latentes n'a retenu que deux axes principaux donc deux valeurs propres que nous présentons dans le tableau 12 avec leur valeur numérique, leur valeur en pourcentage total (ou inertie totale) avant et après la rotation varimax, le pourcentage cumulé, le pourcentage de la variance commune et le pourcentage cumulé.

Nous rappelons que nous sommes parti, pour les données initiales, d'un tableau de quatre colonnes (voir tableau 9 plus haut). Avec l'analyse factorielle et les deux valeurs propres retenues, l'on arrive à conserver 93,51 % de la somme des carrés des distances (variance totale) avant rotation varimax et 100 % de variance totale après rotation varimax. On réduit ainsi le tableau de 4 à 2 colonnes en conservant 100 % de l'inertie totale : c'est une situation parfaite pour l'interprétation des résultats de l'analyse.

Valeurs propres
de l'analyse factorielle des variables latentes

		F1	F2
Valeur propre		**2,34**	**1,40**
% variance totale	Avant rotation varimax	58,50	35,01
	Après rotation varimax	**59,75**	**40,24**
% cumulé	Avant rotation varimax	58,50	93,51
	Après rotation varimax	**59,75**	**100,00**
% variance commune		**62,56**	**37,44**
% cumulé		**62,56**	**100,00**



Nombre de valeurs propres triviales supprimées : 2

Tableau 77

Dans la représentation graphique des méthodes de la figure 19, résultat de l'analyse factorielle après rotation varimax, il est clair que le premier axe (axe F1) met en opposition d'une part, les trois méthodes Brunet, Muller et Yule-Herdan qui ont une grande part de rangs contigus traduisant une faible variabilité (variance spécifique entre 0,006 et 0,110, voir tableau 13) et d'autre part, la méthode Guiraud qui est sous représentée à ce niveau avec seulement 3 rangs contigus et donc (-3) points d'écart par rapport à la moyenne (voir tableau 10 plus haut), c'est ce qui traduit le fait qu'elle ait la

plus forte variabilité des quatre méthodes (variance spécifique de 0,116, voir tableau 13). Nous pouvons dire que ce premier axe est l'axe de la contiguïté sur lequel, plus on se déplace vers la droite, plus on trouve des méthodes à forte contiguïté des rangs (ou à faible variabilité).

Coordonnées des méthodes avant et après rotation varimax

		F1	F2	Variance spécifique
Guiraud	Avant rotation varimax	0,171	0,925	0,116
	Après rotation varimax	**-0,147**	**0,929**	
Brunet	Avant rotation varimax	0,808	0,487	0,110
	Après rotation varimax	**0,599**	**0,729**	
Muller	Avant rotation varimax	0,981	-0,098	0,027
	Après rotation varimax	**0,958**	**0,236**	
Yule-Herdan	Avant rotation varimax	0,834	-0,547	0,006
	Après rotation varimax	**0,968**	**-0,237**	

Tableau 78

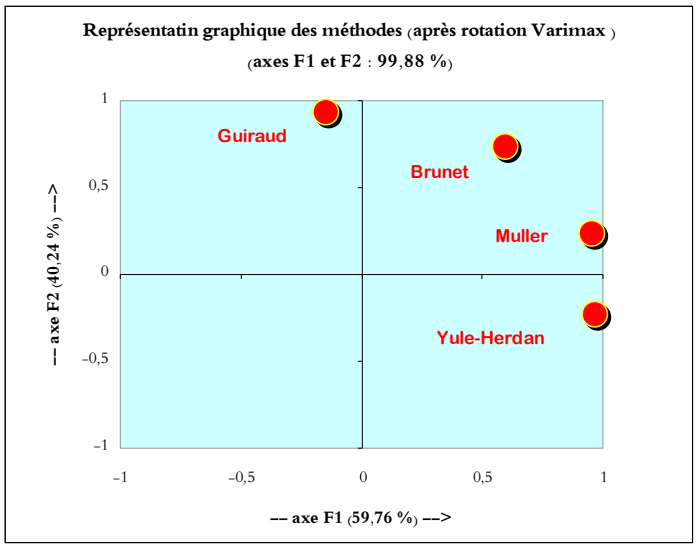


Figure 126
Analyse factorielle des variables latentes du classement des 15 nuits par les quatre méthodes de mesure de richesse lexicale : Représentations graphique des méthodes

Le deuxième axe (axe F2) quant à lui, il oppose les trois méthodes Guiraud, Brunet et Muller à la méthode Yule-Herdan sur la base de la communalité. En effet, alors que les trois méthodes Guiraud, Brunet et Muller ont respectivement, 7, 6 et 8 rangs communs à ceux des autres méthodes, la méthode Yule-Herdan n'en a que 2,

c'est-à-dire avec une sous représentation de (- 4) points d'écart par rapport à la moyenne. Mais ce qui marque le plus cette opposition c'est plutôt l'influence inverse de N et de V_1.

La figure 20 illustre la représentation graphique des nuits résultant de l'analyse factorielle des variables latentes après rotation varimax. Comme toute représentation graphique d'une analyse factorielle, elle constitue la partie la plus riche en information et synthétise un certain nombre de tableaux de chiffres.

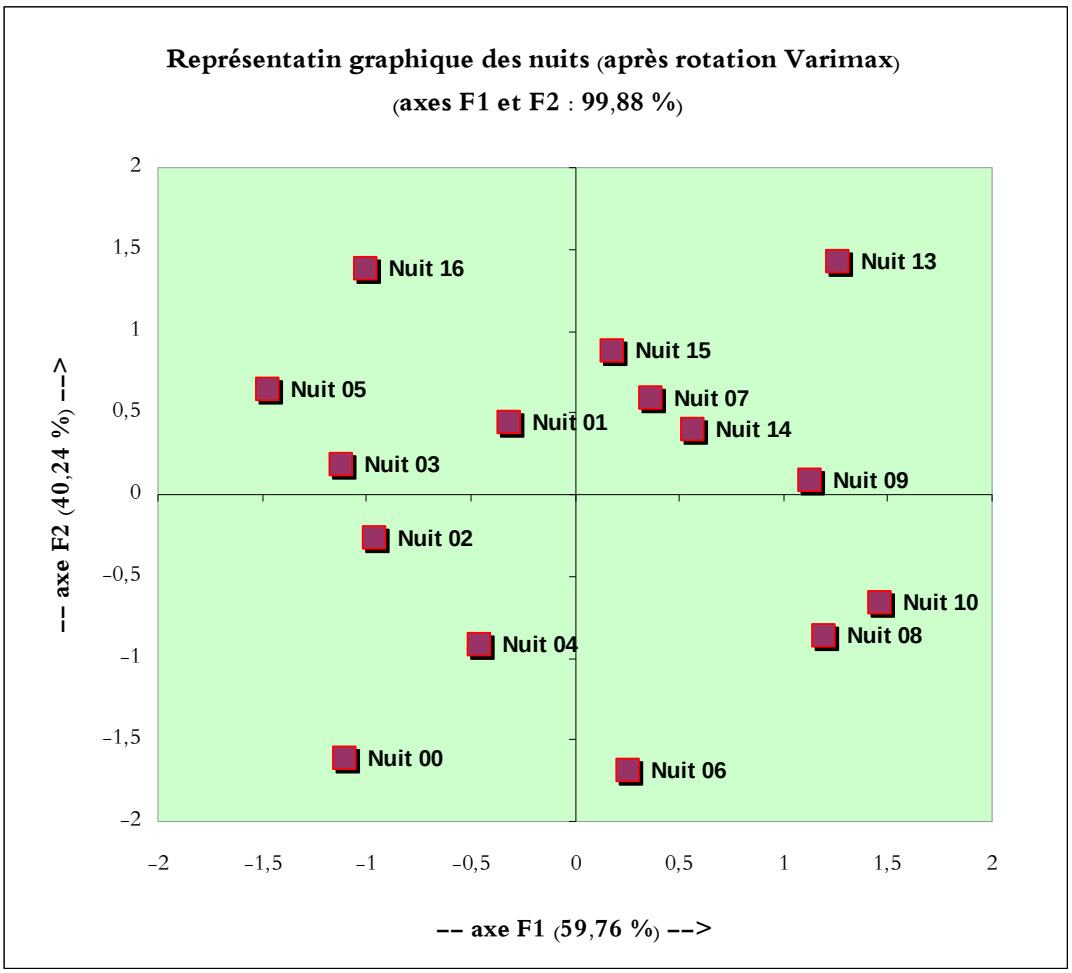


Figure 127
Analyse factorielle des variables latentes du classement des 15 nuits par les quatre méthodes de mesure de richesse lexicale : Représentations graphique des nuits

Pour pouvoir interpréter la représentation graphique résumant l'analyse factorielle, il va falloir étudier le plan engendré par les axes d'inertie F1 et F2 et voir comment se répartissent selon ces deux axes, les points-lignes et les points-colonnes.

Commençons d'abord par faire quelques remarques générales concernant le graphique de la figure 20. Ce graphique est un graphique asymétrique puisque les points-lignes et les points-colonnes sont représentés dans deux échelles différentes. les deux axes F1 et F2 ont des valeurs propres respectivement, de 2,34 et 1,40 et des pourcentages d'inertie de 59,76 % et 40,24 %. Les points-nuits y sont très dispersés à part deux petits groupes où l'on trouve des points-nuits ramassés. Le rapprochement des nuits, telles les nuits 15, 7 et 14 d'un côté, ou les nuits 8 et 10 de l'autre, signifie qu'elles ont un profil similaire.

2.1.1. Interprétation du premier axe :

L'axe F1, qui use de 60 % de l'inertie, oppose nettement les nuits ayant les rangs les plus faibles donc les nuits les plus riches lexicalement, à gauche, aux nuits ayant les rangs les plus élevés donc les nuits les plus pauvres lexicalement, à droite. Plus on se déplace, sur cet axe, vers la droite, plus on trouve des nuits pauvres lexicalement.

En outre, en lisant bien la représentation graphique et surtout les données initiales et les données calculées par l'analyse factorielle, on se rend compte que les méthodes qui "imposent" leurs classements selon ce premier axe (axe F1), ce sont les deux méthodes Muller et Yule-Herdan.

En effet, à gauche du point d'intersection des deux axes (ou centroïde moyen), c'est-à-dire dans la partie négative de l'axe F1, nous trouvons les nuits occupant les rangs de la première moitié du classement selon à la fois la méthode Muller et la méthode Yule-Herdan. Les deux groupes de gauche se confondent donc et leur intersection est égale à leur union.

À droite du centroïde moyen, c'est-à-dire dans la partie positive de l'axe F1, nous trouvons toutes les nuits placées dans la deuxième moitié du classement selon la méthode Yule-Herdan et presque toutes les nuits placées dans la deuxième moitié du

classement selon la méthode Muller à l'exception de la nuit 6 qui est normalement placée au 7ème rang mais propulsée, sur le graphique, à droite du centroïde moyen à cause de l'inertie du classement de la méthode Yule-Herdan selon lequel cette nuit est placée au 11ème rang. L'intersection des deux groupes de droite est composée de toutes les nuits de la deuxième moitié du classement (qui sont les mêmes pour les deux méthodes) excepté la nuit 6.

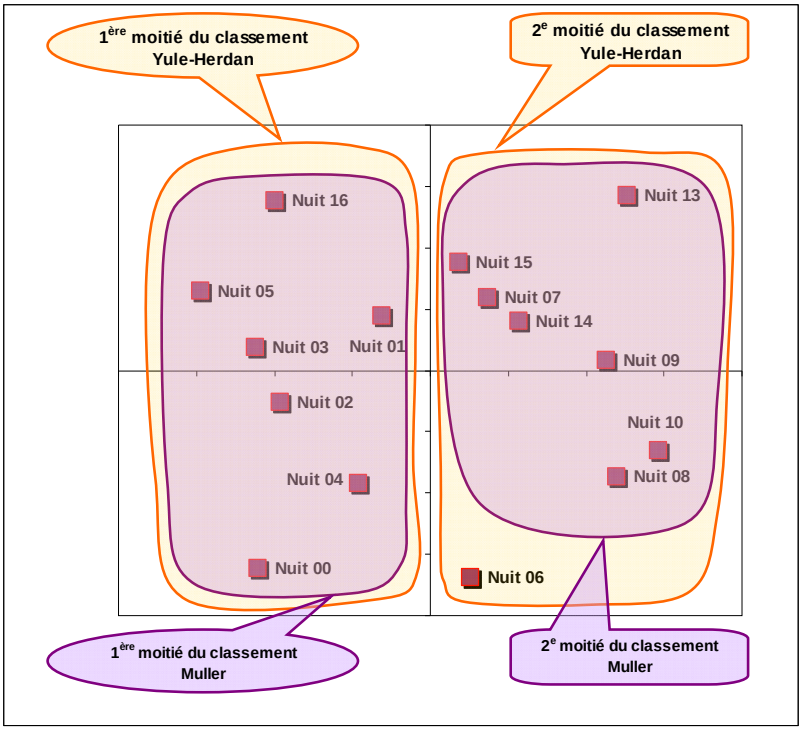


Figure 128
La dispersion des nuits interprétée selon l'axe F1

À l'intérieur de ces deux ensembles, le sens du classement est, bien entendu le même, toujours selon l'axe F1 : plus on se déplace vers la droite, plus on trouve les nuits pauvres lexicalement selon les deux méthodes Muller et Yule-Herdan. Cependant, comment peut-on expliquer l'opposition, par exemple, entre la nuit 0 et la nuit 16 à l'intérieur du premier ensemble, et entre la nuit 6 et la nuit 13 à l'intérieur de deuxième ensemble ? Cela concerne, on s'en doute bien, l'interprétation du deuxième axe. C'est ce que nous allons expliquer tout de suite.

2.1.2. Interprétation du deuxième axe :

L'axe F2 qui n'utilise que 37,44 % de l'inertie, est interprétable quant à lui, comme un axe opposant également les nuits les plus riches lexicalement, dans la partie inférieure du graphique, aux nuits les moins riches lexicalement, dans la partie supérieure de celui-ci, mais selon deux autres méthodes cette fois-ci, la méthode Guiraud et la méthode Brunet. Plus on monte tout le long de l'axe F2, plus on rencontre des nuits de moins en moins riches lexicalement.

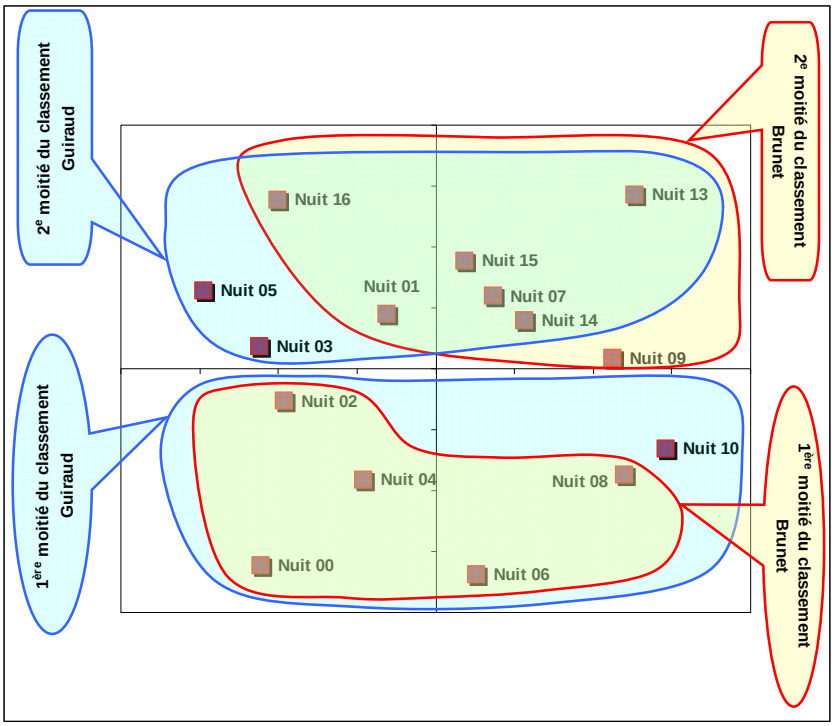


Figure 129
La dispersion des nuits interprétée selon l'axe F2

Au-dessous du centroïde moyen, l'on trouve d'un côté, les nuits qui occupent les rangs de la 1ère moitié du classement selon la méthode Guiraud excepté la nuit 9, et de l'autre, les nuits occupant les rangs de la 1ère moitié du classement selon la méthode Brunet excepté les nuits 3 et 5. La nuit 9 qui est placée au 7ème rang selon Guiraud, est propulsée graphiquement au-dessus de l'axe F1 à cause du poids du rang qu'elle occupe selon la méthode Brunet (14ème rang). Quant aux deux nuits 3 et 5 qui sont placées respectivement, au 6ème et au 5ème rang selon Brunet, elles sont propulsées graphiquement au-dessus de l'axe F1 à cause du poids des rangs qu'elles occupent selon

la méthode Guiraud (10ème et 14ème rang). L'intersection des deux groupes dans la partie inférieure du graphique est l'ensemble regroupant les nuits 0, 2, 4, 6 et 8. Alors que l'intersection des deux groupes dans la partie supérieure du graphique est l'ensemble composé des nuits 16, 1, 15, 7, 14 et 13.

Comme pour les groupes à gauche et à droite du graphique, à l'intérieur des groupes au-dessus et au-dessous du l'axe F1, le sens du classement est le même qu'entre les deux groupes : les nuits les plus riches lexicalement sont placées dans la partie inférieure de chaque groupe, et les nuits les moins riches dans la partie supérieure.

2.1.3. Interprétation globale :

Pour faciliter l'interprétation qui va suivre, désignons maintenant par :

Muller 1 : la 1ère moitié du classement des nuits selon la méthode Muller
Muller 2 : la 2ère ... ... Muller
Yule-Herdan 1 : la 1ère ... ... Yule-Herdan
Yule-Herdan 2 : la 2ère ... ... Yule-Herdan
Guiraud 1 : la 1ère ... ... Guiraud
Guiraud 2 : la 2ère ... ... Guiraud
Brunet 1 : la 1ère ... ... Brunet
Brunet 2 : la 2ère ... ... Brunet

En combinant les deux axes, nous aurons en fin de compte, quatre ensembles répartis de part et d'autre de chacun des deux axes F1 et F2. Géométriquement, ces ensembles correspondent aux quatre quarts du graphique.

Dans le quart inférieur gauche, c'est-à-dire la zone délimitée, en haut et à droite, par les deux parties négatives des axes, nous trouvons trois nuits (nuit 0, nuit 4 et nuit 2) qui sont à l'intersection des 4 groupes représentant la 1ère moitié du classement de chaque méthode : « Guiraud 1 », « Brunet 1 », « Muller 1 » et « Yule-Herdan 1 ». Ces nuits sont donc quadruplement corroborées dans cette position. Nous appellerons ce quart Q1.

Dans le quart diamétralement opposé, c'est-à-dire le quart supérieur droit qui est délimité, en bas et à gauche, par les deux parties positives des axes, nous trouvons 4 nuits (nuit 15, nuit 7, nuit 14 et nuit 13) quadruplement confirmées dans cette position

puisqu'elles sont à l'intersection des 4 groupes représentant la 2ème moitié du classement de chaque méthode : « Guiraud 2 », « Brunet 2 », « Muller 2 » et « Yule-Herdan 2 ». La nuit 9 qui fait aussi partie de ce quart n'est confirmée dans cette position que par trois méthodes, « Brunet 2 », « Muller 2 » et « Yule-Herdan 2 » : la méthode Guiraud l'avait placée dans la 1ère moitié du classement. Nous appellerons ce quart Q4.

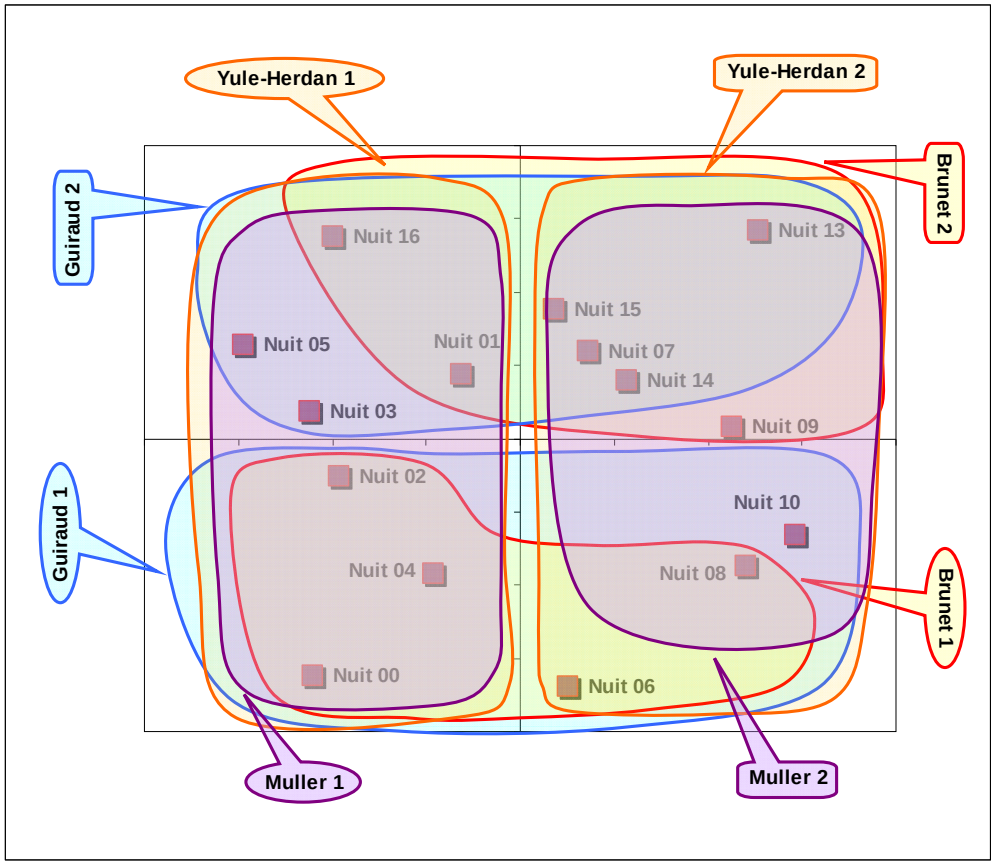


Figure 130
Interprétation globale, selon les deux axes F1 et F2, de la représentation graphique de l'analyse factorielle des variables latentes

Le quart supérieur gauche est le lieu d'intersection de « Guiraud 2 », « Yule-Herdan 1 », « Muller 1 » et « Brunet 2 ». Il est composé de deux couples de nuits, le premier couple (nuit 5 et nuit 3) à l'intersection de « Guiraud 2 », « Yule-Herdan 1 » et « Muller 1 » ; le deuxième couple (nuit 16 et nuit 1) appartient simultanément aux 4 groupes. Nous appellerons ce quart Q2.

Le dernier quart, le quart inférieur droit, est le lieu d'intersection des groupes « Guiraud 1 », « Brunet 1 », « Muller 2 » et « Yule-Herdan 2 ». On y trouve 3 nuits : la

nuit 6 appartenant à « Guiraud 1 », « Brunet 1 » et « Yule-Herdan 2 », la nuit 10 appartenant à « Guiraud 1 », « Muller 2 » et « Yule-Herdan 2 ». La nui 8, quant à elle, appartient aux 4 groupes. Nous appellerons ce quart Q3.

De cette représentation graphique, nous avons pu dégager jusque là, quatre sous-ensembles que nous avons nommés Q1, Q2, Q3 et Q4 se situant, géométriquement, aux quatre quarts du plan du graphique. Considération faite des inerties des deux axes, de l'interprétation de ces deux derniers et du sens de lecture de la dispersion des nuits selon les axes F1 et F2, l'on peut déduire une certaine hiérarchisation de nos quatre sous-ensembles. En effet, pour trouver le sous-ensemble le plus riche lexicalement, il faut se déplacer, sur le graphique, d'abord vers la gauche et ensuite vers le bas ; réciproquement, pour trouver le sous-ensemble le moins riche lexicalement, se déplacer d'abord vers la droite et ensuite vers le haut. De ce fait, le sous-ensemble Q1 est celui qui regroupe les nuits les plus riches lexicalement de tout le corpus. À l'opposé, le sous-ensemble Q4 est celui qui renferme les nuits les plus pauvres lexicalement de tout le corpus. Après Q1, c'est le sous-ensemble Q2 qui regroupe les nuits occupant du 4ème au 7ème rang. Entre Q2, le deuxième sous-ensemble le plus riche et Q4, le sous-ensemble le moins riche lexicalement, nous trouvons Q3 qui est le sous-ensemble constitué des trois nuits occupant du 8ème au 11ème rang.

Nous venons ainsi de déduire graphiquement et à partir de la structure de corrélation des classements rendue possible par l'analyse factorielle, le classement des quatre sous-ensembles facilitant d'un côté, une meilleure synthèse des régularités dans les classements obtenus par les différentes méthodes utilisées, et de l'autre, une première étape d'un éventuel classement résultant de ceux des quatre méthodes. Le schéma directeur de cet éventuel classement "factoriel" est composé de la suite hiérarchisée des quatre sous-ensembles : Q1 > Q2 > Q3 > Q4 (" > " à lire : "plus riche que").

Pour obtenir un classement final de toutes les nuits, inféré graphiquement, nous allons classer les nuits à l'intérieur de chaque sous-ensemble. Pour ce faire, c'est le même raisonnement qui nous a permis de classer les sous-ensembles qui va nous servir de base à ce classement interne. Dans Q1, par exemple, la nuit la plus riche lexicalement

de ce sous-ensemble (mais aussi de tout le corpus étant donné que Q1 est le plus riche des quatre sous-ensembles), sera celle qui se trouve le plus à gauche et le plus en bas de ce quart de plan : c'est donc la nuit 0. De même pour Q4 où la nuit la plus pauvre lexicalement de ce sous-ensemble, et de tout le corpus, sera celle qui se trouve le plus à droite et le plus en haut de ce quart de plan, il s'agit donc de la nuit 13.

Après avoir ainsi classé les quatre sous-ensembles, nous arrivons avec ce raisonnement à classer, de proche en proche, toutes les nuits pour obtenir à la fin un classement qu'on pourrait qualifier de "graphique" ou de "factoriel" résultant de la représentation graphique de l'analyse factorielle des variables latentes, avec rotation varimax, des classements des nuits d'*al-ÞImtâÝ wa l-MuÞânasa* obtenus par les quatre méthodes de mesure de la richesse lexicale : la méthode Guiraud, l'indice W de Brunet, la méthode binomiale de Muller et l'indice V_m de Yule-Herdan. Ce classement est donc le suivant :

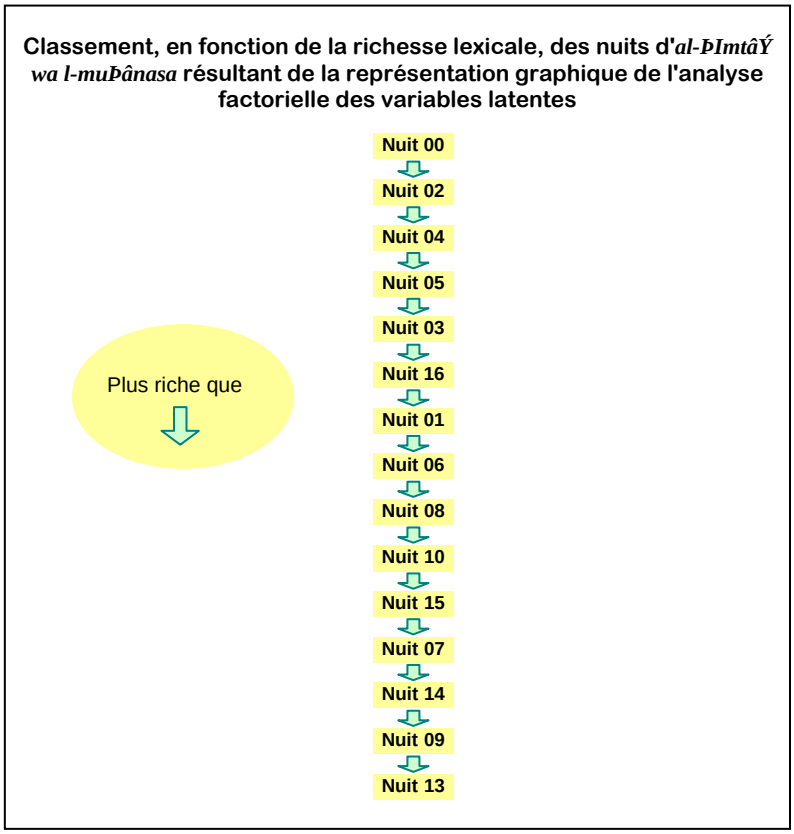


Figure 131

Ce classement "factoriel" étant inféré, nous avons voulu savoir comment se placent, par rapport à lui, les autres classements qui, eux, sont calculés et dont il est la résultante.

Pour ce faire, nous avons calculé la matrice de similarité/dissimilarité, basée sur le coefficient de corrélation de Spearman, des cinq classements (les 4 classements calculés et le classement inféré). La matrice fait apparaître la similarité, ou corrélation qui existe entre chaque couple de ces classements ; mais étant donné que nous connaissons déjà la structure de corrélation entre les quatre classements calculés, nous allons seulement nous intéresser ici aux coefficients de corrélation entre chacun de ces classements et le classement inféré de l'analyse factorielle des variables latentes. C'est ce que nous présentons dans le tableau suivant :

Corrélation entre les classements calculés et le classement inféré

	Méthode "factorielle"	**Corrélation**
Guiraud	0,318	non significative
Brunet	**0,872**	**significative**
Muller	**0,882**	**significative**
Yule-Herdan	0,711	significative

Tableau 79

Il ressort de ces coefficients de corrélation que le classement le plus corrélé au classement "factoriel", c'est celui obtenu selon la méthode binomiale de Muller. Ce couple a le coefficient de corrélation le plus élevé (0,882). En deuxième position, c'est le classement selon la méthode Brunet qui est le deuxième classement le plus corrélé à celui de la méthode "factorielle" avec un coefficient de corrélation de 0,872.

2.2. La richesse lexicale du corpus complet

Nous avons pu jusque là, étudier la richesse lexicale des nuits, classer celles-ci en fonction de la richesse lexicale pour chaque méthode utilisée, et comparer les classements ainsi obtenus, selon les différentes méthodes que nous avons retenues et selon une méthode inférentielle basée sur la représentation graphique de l'analyse factorielle des variables latentes ; mais que peut-on dire à présent, de la richesse lexicale du corpus en entier ?

Notons d'abord que la méthode binomiale n'est pas applicable ici pour mesurer la richesse lexicale du corpus *al-ÞImtâÝ wa l-MuÞânasa* comme un tout. En effet, la loi binomiale, nous l'avons vu dans la présentation de son assise théorique, est utilisée pour mesurer la richesse lexicale des parties d'un corpus et non d'un corpus isolé ne faisant lui-même pas partie d'un autre corpus plus étendu que lui. La méthode binomiale permet d'abord d'établir la distribution théorique de chacune des parties à partir de la distribution de tout le corpus. Ensuite, elle permet d'évaluer la richesse lexicale des parties du corpus en comparant leurs vocabulaires théoriques respectifs au vocabulaire réel du corpus.

Faute d'un corpus plus large dont *al-ÞImtâÝ wa l-MuÞânasa* serait une partie, nous ne pouvons par conséquent, évaluer la richesse lexicale de notre corpus en entier, selon la méthode binomiale, puisqu'il ne constitue pas une partie d'un tout[304].

Nous avons, en revanche, tous les éléments nécessaires à la mesure de la richesse lexicale d'*al-ÞImtâÝ wa l-MuÞânasa* en utilisant la méthode de Guiraud, celle de Brunet et celle de Yule-Herdan.

[304] Nous entendons ici un corpus constitué, organisé et expoitable par ordinateur, qu'il soit annoté ou non.

2.2.1. La richesse lexicale de tout le corpus selon la méthode Guiraud

L'étendue d'*al-ÞImtâÝ wa l-MuÞânasa* étant de N = 61 177, l'étendue de son vocabulaire de V = 6 652, nous calculons la racine carrée de N, $\sqrt{N}$ = 247,340, puis nous divisons l'étendue du vocabulaire V par la racine carrée de N que nous venons de calculer, nous aurons l'indice de richesse lexicale selon Guiraud :

$$\frac{V}{\sqrt{N}} = \frac{6\,652}{247{,}34} = \mathbf{26{,}894}$$

La richesse lexicale de tout *al-ÞImtâÝ wa l-muÞânasa* **selon la formule de Guiraud**

	N	***V***	$\sqrt{N}$	$V/\sqrt{N}$
Al-ÞImtâÝ wa l-MuÞânasa	61 177	6 652	247,340	**26,894**

Tableau 80

2.2.2. La richesse lexicale de tout le corpus selon la méthode Brunet

Avec N = 61 177, V = 6 652 et l'exposant fractionnaire α = 0,172, nous calculons $V^{\alpha} = 6\,652^{0{,}172} = 4{,}5451$, puis $V^{-\alpha} = \frac{1}{V^{\alpha}} = \frac{1}{4{,}5451} = 0{,}22$. Nous calculons enfin, $W = N^{V^{-\alpha}} = 61177^{0{,}22} = 11{,}301$.

Comme l'indice W est un indice évoluant en raison inverse de la richesse lexicale, pour avoir un indice qui évolue dans le même sens que celle-ci, nous calculons

$$\mathrm{R} = \frac{25 - W}{15} = \frac{25 - 11{,}301}{15} = \mathbf{0{,}913}$$

La richesse lexicale de tout *al-ÞImtâÝ wa l-muÞânasa*
selon la l'indice W de Brunet

	N	***V***	V^{α}	$V^{-\alpha}$	$W = N^{V^{-\alpha}}$	***R***
Al-ÞImtâÝ wa l-MuÞânasa	61177	6652	4,5451	0,22	11,301	**0,913**

Tableau 81

2.2.3. La richesse lexicale de tout le corpus selon la méthode Yule-Herdan

Comme nous l'avons vu pour le calcul de la richesse lexicale des nuits, le calcul de cet indice V_m est laborieux. Après avoir calculé la fréquence moyenne $\overline{f}$ qui est égale au rapport $\frac{N}{V} = \frac{61\,177}{6\,652} = 9{,}20$, nous partons du tableau de la distribution des fréquences pour calculer, pour chaque V_i, la distance $\left(f_i - \bar{f}\right)^2$. Ensuite, après avoir porté au carré chacune de ces distances, nous calculons la somme des carrés des distances :

$$\sum_{i=1}^{n}\left(f_i - \bar{f}\right)^2 = 13\,516\,058{,}28$$

Laquelle est divisée par V :

$$\frac{\sum_{i=1}^{n}\left(f_i - \bar{f}\right)^2}{V} = 2031{,}879$$

Puis vient le moment de calculer l'écart-type σ_f qui est la racine carrée de la variance que nous venons de calculer :

$$\sigma_f = \sqrt{\frac{\sum_{i=1}^{n}(f_i - \bar{f})^2}{V}} = 45{,}08$$

Cet écart-type est ensuite divisé par la fréquence moyenne pour obtenir l'indice de variation de fréquence ν_f :

$$\nu_f = \frac{\sigma_f}{\bar{f}} = \frac{45{,}08}{9{,}20} = 4{,}90.$$

En fin de compte l'indice V_m de Yule-Herdan est calculé en divisant ν_f par la racine carrée de V :

$$V_m = \frac{\nu_f}{\sqrt{V}} = \frac{4{,}90}{81{,}56} = \mathbf{0{,}060}$$

La richesse lexicale de tout *al-ÞImtâÝ wa l-muÞânasa* **selon la l'indice V_m de Yule-Herdan**

	N	***V***	$\bar{f}$	$\Sigma(f_i - \bar{f})^2$	$\Sigma(f_i - \bar{f})^2 / V$	σ_f	ν_f	V_m
Al-ÞImtâÝ wa l-MuÞânasa	61 177	6 652	9,20	13 516 058,28	2031,879	45,08	4,90	**0,060**

Tableau 82

Nous récapitulons dans le tableau suivant, la richesse lexicale de notre corpus *al-ÞImtâÝ wa l-MuÞânasa* complet calculée selon les trois méthodes, Guiraud, Brunet et Yule-Herdan :

Récapitulatif de la mesure de la richesse lexicale de tout *al-ÞImtâÝ wa l-muÞânasa* **selon les trois formules**

	Formule de Guiraud	**Indice W de Brunet**	**Indice V_m de Yule-Herdan**
Al-ÞImtâÝ wa l-MuÞânasa	**26,894**	**0,913**	**0,060**

Tableau 83

2.3. La richesse lexicale d'*al-ÞImtâÝ wa l-MuÞânasa* comparée à celle d'autres corpus

Dans l'absolu, les indices que nous venons de résumer dans le tableau 18 n'ont aucune signification en eux-mêmes. De par le caractère relatif de la notion de richesse lexicale que nous évoquions dans l'introduction de ce chapitre ; pour que ces indices soient significatifs, il faut les comparer à d'autres indices de richesse lexicale calculés selon les mêmes méthodes. Encore faut-il que les règles suivies dans le dépouillement lexicologique et la quantification soient les mêmes ou que la norme lexicologique soit au moins clairement énoncée.

Malheureusement, parmi les rares travaux en lexicométrie arabe, encore plus rares sont ceux dont les auteurs traitent de la question de richesse lexicale, et pour ceux qui le font, à l'image de Chabir Ayadi (Ayadi, 1998), l'énoncé de la norme lexicologique qu'ils ont adoptée est souvent très expéditif et laisse beaucoup à désirer.

Il convient de noter ici que, pour son étude de la richesse lexicale des *Maqâmât al-HamaÆânî*, Ch. Ayadi a utilisé comme unité de décompte, la forme et non le lemme. De plus, il n'a utilisé pour la mesure de la richesse lexicale, ni la méthode binomiale, ni l'indice W de Brunet, ni l'indice V_m de Yule-Herdan. Parmi les méthodes que nous avons présentées, commentées et appliquées à notre corpus, Ch. Ayadi n'a utilisé que la formule de Guiraud pour laquelle il a conclu à sa dépendance du facteur étendue. Avant d'utiliser cette formule, il a appliqué à son corpus deux des touts premiers rapports utilisés entre V_1 et V (pour lui F_1 et F), $\frac{V_1}{V}$ et $\frac{V_1^3}{V^2}$, dont on sait depuis longtemps déjà qu'ils ne sont pas valables pour évaluer la richesse lexicale d'un texte du fait de leur grande sensibilité à l'étendue du corpus, pour conclure en fin de compte que « force est de constater qu'aucun de ces rapports n'est réellement constant dans un texte donné, donc indépendant de son étendue »[305]. C'est pourquoi il s'est mis à la recherche d'une formule nouvelle qu'il a baptisée "*coefficient de richesse lexicale*" en se posant comme

[305] Ayadi, Chabir, *op. cit.*, p. 125.

objectif « d'établir un coefficient de richesse lexicale indépendante de l'étendue et qui permet de comparer des textes de longueur différente »[306]. Nous ne somme, à vrai dire, pas du tout convaincu ni des principes de base de ce "*coefficient de richesse lexicale*", ni des deux paramètres qui ont servi à l'élaboration de son "texte-modèle"[307] et que nous jugeons chargés de contradictions.

Nous allons comparer dans ce qui suit, les deux corpus, le nôtre *al-ÞImtâÝ wa l-MuÞânasa* et celui de Chabir Ayadi *Maqâmât al-HamaÆânî* ; pour cela :

- Nous avons d'abord calculé, à partir des données quantitatives de Ch. Ayadi, la richesse lexicale des *Maqâmât al-HamaÆânî* selon les trois méthodes que nous avons utilisées pour notre corpus, la méthode Guiraud, l'indice W de Brunet et l'indice V_m de Yule-Herdan.
- Ensuite, considération faite que les données quantitatives présentées par Ch. Ayadi pour son corpus étaient basées sur la forme et non le lemme, nous étions donc obligé de refaire, pour notre corpus, tous les calculs et les manipulations nécessaires, avec tout ce que cela représente comme investissement en termes de temps et d'énergie, en prenant cette fois-ci la forme comme unité de décompte et non le lemme, et ce pour que nous puissions procéder à des comparaisons entre les deux corpus en fonction de la richesse lexicale. Faute de quoi il n'y aurait aucun sens à comparer des valeurs de richesse lexicale ayant été calculées sur des bases différentes.

Mais avant de comparer la richesse lexicale de notre corpus à celle des *Maqâmât al-HamaÆânî* selon les trois méthodes, nous devons signaler d'abord que la méthode de comparaison des indices nous donne une combinaison d'indices du type ❶ c'est-à-dire (1-1-1-1). Ce qui veut dire qu'aucune conclusion ne peut être tirée et la comparaison entre la richesse lexicale d'*al-ÞImtâÝ wa l-MuÞânasa* et celle de *Maqâmât al-HamaÆânî* reste non résolue.

[306] *Idem*, *ibidem*, p.126.

[307] *Idem*, *ibidem*, p.126 et suivantes.

En effet, en comparant, deux à deux, les indices pour chacun des corpus, nous avons trouvé que ceux d'*al-ÞImtâÝ wa l-MuÞânasa* sont à chaque fois supérieurs à ceux des *Maqâmât al-HamaÆânî*.

	V	V_1	$\overline{f}$	q_1
Al-ÞImtâÝ wa l-MuÞânasa :	10 174	6 763	6,01	0,661
Maqâmât al-HamaÆânî :	5 219	3 171	5,71	0,392
Combinaison obtenue →	**(1 - 1 - 1 - 1)**			

En ce qui concerne, maintenant, les trois autres méthodes de mesure de la richesse lexicale, après les différents calculs nous obtenons les données suivantes :

Pour la méthode Guiraud :

	N	V	$\sqrt{N}$	$V/\sqrt{N}$
Al-ÞImtâÝ wa l-MuÞânasa	61 177	10 174	247,340	**41,134**
Maqâmât al-HamaÆânî	29 820	5 219	172,685	**30,223**

Pour l'indice W de Brunet :

	N	V	V^{α}	$V^{-\alpha}$	$W = N^{V^{-\alpha}}$	**R**
Al-ÞImtâÝ wa l-MuÞânasa	61 177	10 174	4,8898	0,2045	9,526	**1,032**
Maqâmât al-HamaÆânî	29 820	5 219	4,3594	0,2294	10,627	**0,958**

Pour l'indice V_m de Yule-Herdan :

	N	V	$\overline{f}$	σ_f	ν_f	V_m
Al-ÞImtâÝ wa l-MuÞânasa	61 177	10 174	6,01	68,85	11,45	**0,114**
Maqâmât al-HamaÆânî	29 820	5 219	5,71	45,94	8,04	**0,111**

Nous résumons les valeurs de la richesse lexicale obtenues , pour chacun des deux corpus, par les trois méthodes dans le tableau suivant.

La richesse lexicale entre *al-ÞImtâÝ wa l-muÞânasa* **et** *Maqâmât al-HamaÆânî* **(sur la base des formes)**

	Formule de Guiraud	Indice W de Brunet	Indice V_m de Yule-Herdan
Al-ÞImtâÝ wa l-MuÞânasa	**41,134**	**1,032**	**0,114**
Maqâmât al-HamaÆânî	**30,223**	**0,958**	**0,111**

Tableau 84

Il ressort donc très clairement de ces valeurs de richesse lexicale qu'*al-ÞImtâÝ wa l-MuÞânasa* est plus riche lexicalement que *Maqâmât al-HamaÆânî* et ce selon chacune des trois méthodes de mesure utilisées.

2.4. Quelle(s) méthode(s) pour la mesure de la richesse lexicale des corpus arabes ?

En guise de synthèse de ce bilan et en nous inscrivant dans une perspective normative dans le cadre de la lexicométrie arabe, voici ce que nous recommandons pour la mesure de la richesse lexicale des corpus arabes :

↳ Commencer, tout d'abord, par la méthode de comparaison des indices. C'est une méthode primordiale, elle doit être utilisée dans tous les cas, surtout de prime abord. Même dans le cas où cette méthode n'arrive pas à résoudre toutes les comparaisons binaires, les cas résolus (surtout les cas sûrs) pourront ouvrir la voie par la suite, aux autres méthodes de mesure de richesse lexicale qu'on aura choisies d'utiliser en dernier recours. En effet, sur les 22 comparaisons binaires résolues par cette méthode

dans le cadre de l'étude de notre corpus, 21 sont confirmées, à la fois par les quatre méthodes utilisées ultérieurement et par le classement "factoriel" résultant de la représentation graphique de l'analyse factorielle. En outre, la méthode de comparaison des indices est la plus logique, la plus directe et la plus simple à calculer même sans avoir recours à l'ordinateur.

↳ S'il ne faut choisir qu'une seule méthode de mesure de la richesse lexicale, nous suggérons, sans hésitation aucune, l'utilisation de la méthode binomiale proposée par Charles Muller. Ce choix est dicté par deux considération au moins : d'un côté, la « logique irréprochable » de cette méthode et son bien-fondé théorique ; elle permet, en effet, d'évaluer la richesse d'un fragment de texte en comparant son vocabulaire théorique à son vocabulaire réel, et ce après avoir, bien entendu, établi la distribution théorique du fragment à partir de la distribution du texte complet. D'un autre côté, la méthode binomiale a pu séduire un très grand nombre de spécialistes de lexicométrie qui l'ont adoptée et utilisée simultanément ou à la place même, pour certains, de leur propre méthode, à l'image d'Etienne Brunet qui suggère que « si l'on se méfie de l'indice W, qui n'est guère qu'une approximation empirique, on peut recourir à la logique irréprochable de la loi binomiale »[308]. Le modèle binomial est donc un modèle très stable et très fiable qui est « resté jusqu'à aujourd'hui le principal outil dont on se sert pour évaluer la richesse lexicale »[309].

↳ À défaut d'utiliser la méthode binomiale et si l'on tient toujours à n'utiliser qu'une seule méthode, nous suggérons l'utilisation de l'indice W de Brunet. En effet, nous avons pu établir dans ce chapitre consacré à la mesure de la richesse lexicale, que les deux méthodes Muller et Brunet faisaient preuve d'une grande stabilité. Les deux autres méthodes, quant à elles, étaient fortement sensibles à la longueur des textes, soit directement (la méthode Guiraud), soit inversement (la méthode Yule-Herdan). Cela nous permet d'éliminer les deux méthodes dont les résultats ont été ternis par les excentricités et la forte variabilité : la méthode Guiraud et l'indice Vm de Yule-Herdan. Faire confiance aux deux méthodes les plus stables et les plus sûres, la méthode Muller et la méthode Brunet, nous semble une décision d'une évidence indiscutable.

[308] Brunet Etienne, *Le vocabulaire de Victor Hugo*, 1988, p. 25.

[309] Cossette André, *op. cit.*, p. 19.

D'autant plus que les classements obtenus avec ces deux méthodes ont, comme nous l'avons vu plus haut, les deux plus grands (et presque équivalents) coefficients de corrélation (0,882 et 0,872) avec le classement "factoriel" résultant de l'analyse factorielle de tous les classements que nous avons obtenus pour notre corpus.

↳ Si, finalement, aucune des trois solutions précédentes n'est retenue, l'on peut imaginer une utilisation conjointe des quatre méthodes (voire même plus), ou d'une partie d'entre elles, dans le but de comparer les classements obtenus. À la suite de cette comparaison, l'on serait amené soit, à privilégier une méthode précise, soit à utiliser l'analyse factorielle des variables latentes, avec rotation varimax, pour mieux résumer la structure de corrélation des différents classements. L'analyse factorielle des variables latentes peut tout à fait, servir d'outil d'aide à la prise de décision dans le choix de l'une ou l'autre des méthodes. Dans le cas contraire, l'on peut déduire de la représentation graphique, comme nous l'avons fait plus haut pour notre corpus[310], un classement qui serait la résultante de tous les classements obtenus avec les différentes méthodes.

[310] Nous l'avons fait aussi et surtout, dans un but méthodologique et normatif.

Chapitre 11

L'accroissement du vocabulaire

Nous venons de voir, au chapitre précédent, l'un des faits de la structure lexicale d'*al-ÞImtâÝ wa l-MuÞânasa* observée d'un angle statique : la richesse lexicale.

La structure lexicale peut également être observée d'un autre point de vue, dynamique cette fois-ci : l'accroissement du vocabulaire.

Loin donc d'une conception statique des textes que nous avons adoptée jusqu'à maintenant et qui considère le corpus soumis à l'analyse quantitative comme « une chose achevée », la notion d'accroissement du vocabulaire met l'accent sur le caractère dynamique du corpus à étudier.

Il est donc question, avec la notion d'accroissement du vocabulaire, d'étudier « la façon dont le vocabulaire d'un texte se constitue quand on suit ce texte du premier mot au point final »[311]. Autrement dit, ce qui sera étudié dans ce chapitre, ce n'est plus l'étendue du corpus N, considérée comme un ensemble clos, prise dans sa relation avec l'étendue du vocabulaire pour évaluer la richesse lexicale, mais plutôt la façon dont l'étendue du vocabulaire V se constitue au fur et à mesure que le texte croît jusqu'à son dernier mot.

L'accroissement du vocabulaire ou accroissement lexical d'un fragment déterminé d'un corpus, correspond au nombre des vocables qui y apparaissent pour la première fois. Et l'évaluation de l'accroissement du vocabulaire de tout le corpus se fait par l'accroissement cumulé de tous les fragments composant le corpus, autrement dit par le nombre cumulé des vocables nouveaux qui apparaissent, à chaque fois, dans les parties consécutives du corpus.

[311] Muller Charles, *Etude...*, *op. cit.*, p. 67

1. Accroissement réel du vocabulaire

Au début d'un texte, chaque mot apporte normalement un vocable, avant que les répétitions n'interviennent pour ôter à ces mots le statut de vocable et leur donner celui d'occurrence de vocable. Ainsi quand on parcourt la première page, par exemple, d'*al-ÞImtâÝ wa l-MuÞânasa*, on compte 265 occurrences représentant 118 vocables qui sont tous des vocables nouveaux puisque l'on est à la première page de tout le corpus. À la 2ème page, on enregistre 336 occurrences pour 154 vocables dont 124 y paraissent pour la première fois. Pour la 3ème page, l'on compte 220 occurrences, 105 vocables et 69 vocables nouveaux. L'accroissement du vocabulaire est donc de 118 pour la 1ère page, 124 pour la 2ème et 69 pour la 3ème et ainsi de suite. Mais ce qui nous intéresse ici c'est l'accroissement du vocabulaire du corpus réparti en nuits et non en pages. Pour cela, étudier l'accroissement du vocabulaire des nuits revient à enregistrer pour chacune d'entre elles le nombre des vocables n'ayant pas encore apparu dans l'une ou l'autre des nuits précédentes.

Aussi, peut-on évaluer pour chaque nuit l'accroissement du vocabulaire dans chaque classe de fréquence. Cela nous permet de mesurer l'effectif des vocables nouveaux selon leur fréquence d'apparition dans chaque nuit. Cette répartition de l'apport lexical selon chaque classe de fréquence va nous aider à mieux considérer la structure lexicale d'*al-ÞImtâÝ wa l-MuÞânasa*. Nous allons donc mesurer d'une part l'accroissement général du vocabulaire et de l'autre, l'accroissement du vocabulaire par classe de fréquence.

1.1. L'accroissement général du vocabulaire

En pratique, pour calculer l'accroissement du vocabulaire réel de notre corpus, nous partons de l'étendue du vocabulaire de la première partie, le préambule, qui est V = 1 412 vocables ; vu que cette partie est la première du corpus, tous ses vocables

sont logiquement des vocables nouveaux. Nous comptons (l'ordinateur le fait pour nous) par la suite, les vocables qui apparaissent dans la nuit suivante, la nuit 1, mais qui sont absents de la partie précédente, autrement dit des vocables qui apparaissent pour la première fois dans la nuit 1 : nous trouvons 404 vocables nouveaux. La même opération est pratiquée pour la nuit suivante, la nuit 2, dans laquelle nous comptons 474 vocables nouveaux, c'est-à-dire des vocables n'appartenant ni à la nuit 1, ni au préambule. Et ainsi de suite, jusqu'à la dernière partie, la nuit 16, dans laquelle nous n'enregistrons que 78 vocables nouveaux. Nous obtenons en fin de compte, les valeurs suivantes :

	N00	**N01**	**N02**	**N03**	**N04**	**N05**	**N06**	**N07**	**N08**	**N09**	**N10**	**N13**	**N14**	**N15**	**N16**
Accr^{t}	1412	404	474	293	565	113	796	215	791	285	869	88	175	94	78
Accr^{t} cumulé	1412	1816	2290	2583	3148	3261	4057	4272	5063	5348	6217	6305	6480	6574	6652

Ces valeurs de l'accroissement réel du vocabulaire dans *al-ÞImtâÝ wa l-MuÞânasa* sont représentées graphiquement dans la figure suivante :

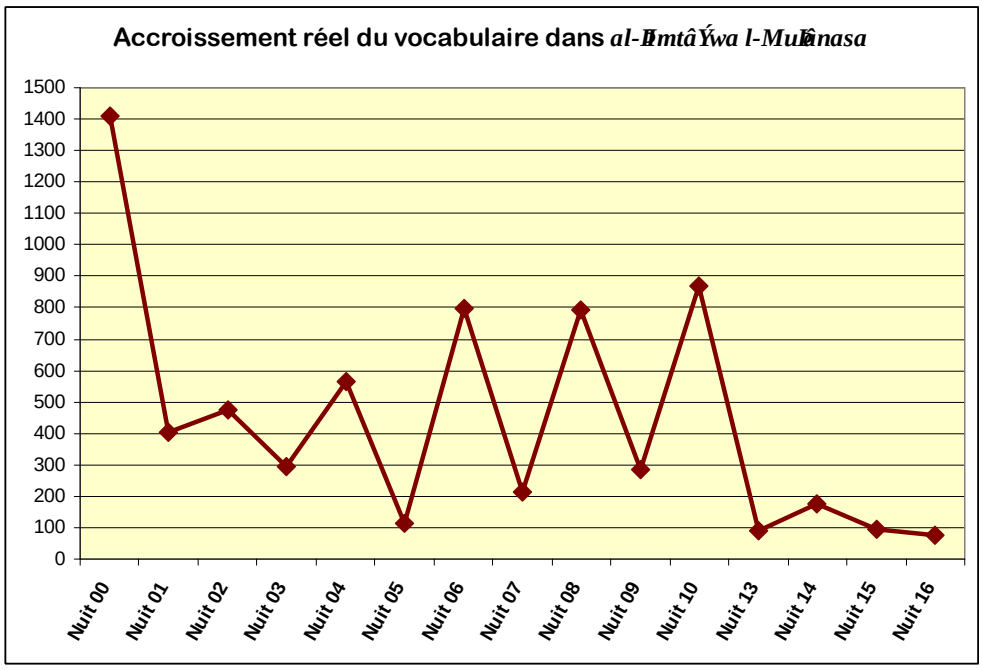


Figure 132

Les fluctuations qui donnent à cette courbe une forme en dents de scie traduisent le fait qu'il y ait des nuits où, pour des raisons principalement thématiques, l'apport lexical augmente ; alors que dans d'autres nuits, les répétitions dominent et l'apport s'y trouve donc faible.

Par ailleurs, il est clair des valeurs de l'accroissement lexical et de la courbe de la figure 1 que la nuit accusant le plus grand apport lexical c'est la nuit 10 avec 869 vocables nouveaux[312] ; suivie par les nuits 6 et 8 avec respectivement 796 et 791 vocables nouveaux.

1.2. L'accroissement du vocabulaire par classe de fréquence

Pour que la vision, dynamique, de la structure lexicale d'*al-ÞImtâÝ wa l-MuÞânasa* soit plus complète, nous avons mesuré l'accroissement lexical de chaque nuit dans chaque classe de fréquence. Étant donné que l'effectif des vocables nouveaux devient très faible au-delà de la fréquence 5, nous nous sommes arrêté à cette fréquence et avons marqué seulement la somme des effectifs de toutes les classes de fréquences supérieures à 5. Nous présentons le résultat de ces mesures dans le tableau à double entrée suivant dont les lignes désignent les différentes nuits de notre corpus et les colonnes les fréquences étudiées. À l'intersection de chaque ligne et de chaque colonne, on trouve l'effectif des vocables nouveaux détectés dans la nuit inscrite au début de la ligne et ayant la fréquence notée en tête de la colonne.

[312] Nous faisons abstraction ici du préambule qui compte 1412 vocables et qui sont tous considérés comme nouveaux du fait que c'est la toute première partie du corpus.

Accroissement réel du vocabulaire par classe de fréquence dans *al-ÞImtâÝ wa l-muÞânasa*

	Accroissement par classe de fréquences						Accroissement	Accr. Cumulé
	Freq.1	Freq.2	Freq.3	Freq.4	Freq.5	Freq.>5		
N00	1 048	188	60	34	10	72	**1 412**	1 412
N01	338	46	12	4	3	1	**404**	1 816
N02	411	49	7	2	3	2	**474**	2 290
N03	270	20	1	1	1	0	**293**	2 583
N04	514	40	10	0	1	0	**565**	3 148
N05	107	4	2	0	0	0	**113**	3 261
N06	672	83	30	3	4	4	**796**	4 057
N07	188	20	5	2	0	0	**215**	4 272
N08	672	93	16	6	2	2	**791**	5 063
N09	247	29	4	4	1	0	**285**	5 348
N10	621	115	61	22	18	32	**869**	6 217
N13	70	14	0	0	3	1	**88**	6 305
N14	145	23	3	3	1	0	**175**	6 480
N15	82	7	4	1	0	0	**94**	6 574
N16	70	7	0	0	1	0	**78**	6 652
Total	**5 455**	**738**	**215**	**82**	**48**	**114**	**6 652**	

Tableau 85

La dissymétrie de ce tableau est facilement remarquable : il y aune concentration des effectifs les plus forts dans les premières colonnes. Ce qui veut dire que l'accroissement du vocabulaire est de plus en plus faible à mesure que la fréquence augmente.

Hormis, la nuit 0 où tous les vocables sont considérés comme nouveaux, les effectifs de la première colonne, celle de la classe de fréquence 1, accusent une distribution assez régulière dans laquelle les vocables nouveaux de fréquence 1 sont quasiment proportionnels à l'étendue du vocabulaire de chaque nuit. La distribution de la classe de fréquence 2 est elle aussi, avec un degré moindre, presque régulière. Ce n'est qu'à partir de la classe de fréquence 3 que l'on observe une disparité nette et où

l'on remarque, par exemple, que la nuit 10 et la nuit 6 ont le plus grand effectif de vocables nouveaux que les autres nuits.

Un fait marquant ici concerne la nuit 10 : outre le fait qu'elle soit la nuit ayant le plus grand apport lexical au niveau général, cette nuit présente également une distribution qui la place en tête de toutes les autres nuits au niveau de chaque classe de fréquence à l'exception de la classe de fréquence 1 où elle se contente de la deuxième place.

La figure 2 met en représentation graphique cette distribution de l'accroissement du vocabulaire selon les classes de fréquences. On y voit d'une façon plus parlante les remarques et les constatations que nous venons de formuler.

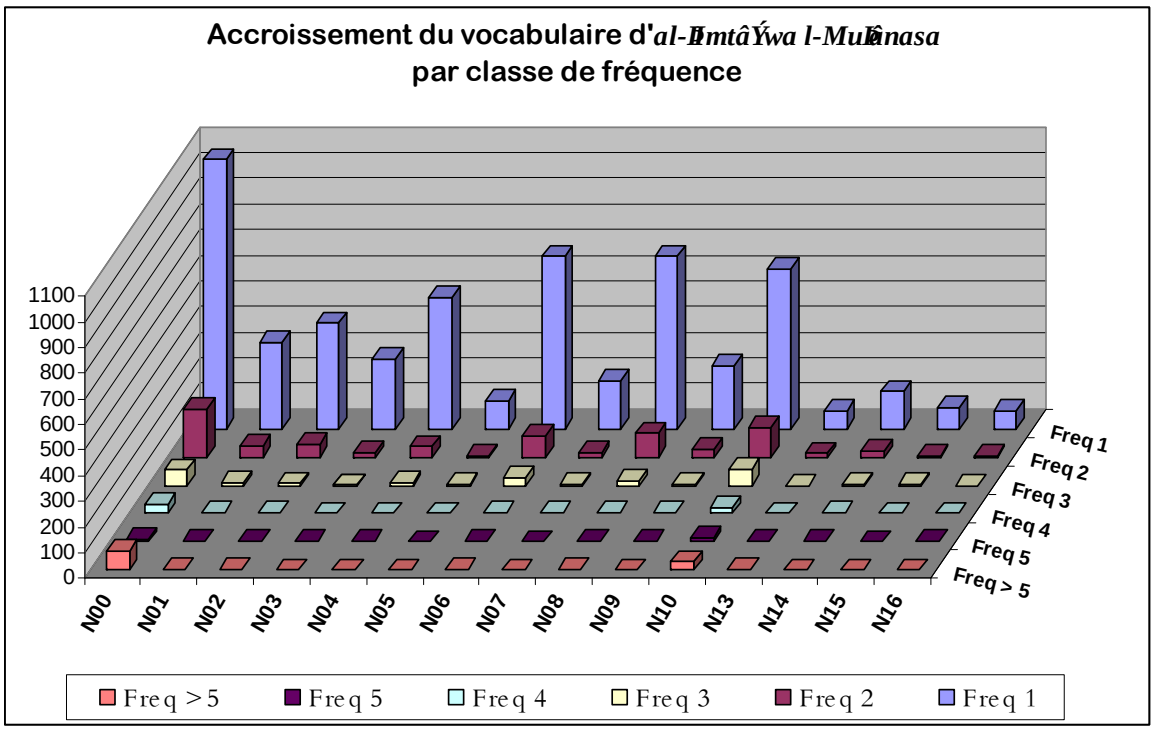


Figure 133

2. Accroissement théorique du vocabulaire

Nous venons de voir comment se renouvelle réellement le vocabulaire dans *al-ÞImtâÝ wa l-MuÞânasa*, d'une nuit à une autre et même d'une classe de fréquence à une autre ; mais comment peut-on évaluer le mouvement de ce renouvellement et comment peut-on interpréter les fluctuations qui peuvent affecter cet accroissement lexical. D'autant plus que des caractéristiques stylistiques ou thématiques peuvent influencer plus ou moins le mouvement de l'accroissement du vocabulaire ; surtout quand on sait qu'à longueur égale, c'est une constante stylistique ou thématique qui détermine l'accroissement du vocabulaire.

Pour repérer donc les mouvements stylistiques ou thématiques qui affectent le processus de renouvellement lexical, il faut pouvoir comparer l'accroissement réellement observé dans le corpus avec un modèle qui décrirait l'évolution "normale" de cet accroissement lexical.

2.1. Modèle de calcul de l'accroissement théorique du vocabulaire

Nous allons présenter dans ce qui suit, la méthode de calcul de l'accroissement théorique du vocabulaire selon la loi binomiale[313] proposée par Charles Muller.

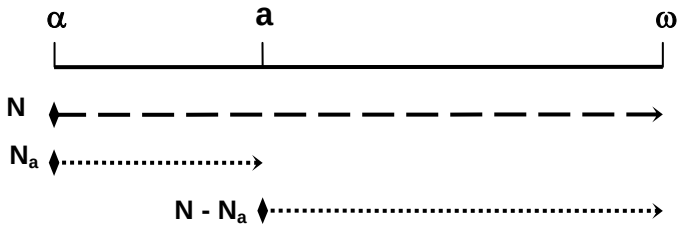
Nous commençons par dire que ce modèle est en quelque sorte lié aux données condensées du corpus dont on veut calculer l'accroissement théorique du vocabulaire et

[313] Pour présenter cette méthode, nous nous sommes inspiré de la présentation détaillée faite par Charles Muller dans son *Initiation à la statistique linguistique*, 1968, pp. 183-190. Voir aussi, Charles Muller, *Etude de statistique lexicale*, 1967, pp. 67-79.

ce parce qu'il utilise la distribution des fréquences de ce même corpus pour mesurer cet apport lexical théorique.

Connaissant l'étendue du vocabulaire V ainsi que les composantes de cette étendue associées aux fréquences 1, 2, 3, **...** n représentant les effectifs V_1, V_2, V_3, **...** V_n , cette méthode se propose de déterminer la probabilité pour un vocable de fréquence i d'avoir sa première occurrence soit avant un point "a" du texte, soit entre un point "a" et un point "b" de ce texte. Autrement dit, la question qui est posée est de savoir où se trouve la première occurrence de chacun des V vocables.

Pour répondre à cette question, considérons le texte comme une suite continue et orientée de début "α" et de fin "ω". Situons sur cette ligne un point "a", N_a est alors le nombre de mots qui sont à gauche de "a" et qui forment le segment "α → a". Nous aurons le schéma suivant :



La longueur relative du fragment "α → a" est exprimé par $\frac{N_a}{N}$ et celle du texte extérieur à ce fragment, c'est-à-dire du fragment "a → ω" par $\frac{N - N_a}{N}$.

Ces deux rapports expriment aussi la probabilité pour une occurrence quelconque de se trouver à gauche ou à droite de"a".

Nous voulons calculer le nombre des vocables nouveaux, c'est-à-dire des occurrences apparues pour la première fois dans le fragment "a → ω". Pour cela il faut que ces occurrences se trouvent exclusivement à droite de "a".

La probabilité donc pour qu'une occurrence d'un vocable se trouve à droite de "a" est :

$$\frac{N - N_a}{N}$$

Pour i occurrences d'un même vocable, on a donc :

$$\rho = \left(\frac{N - N_a}{N} \right)^i .$$

Pour les V_i vocables qui ont la fréquence i, l'effectif théorique de ceux qui ne se trouvent pas dans "$\alpha \rightarrow$ a" est de :

$$E\left(V_{0_i}\right) = V_i \left(\frac{N - N_a}{N} \right)^i$$

Cette formule calcule donc l'effectif théorique des vocables de fréquence i qui, jusqu'au point "a", ont une fréquence 0.

Pour l'ensemble des fréquences représentées dans le texte, le calcul de l'effectif théorique de tous les vocables qui n'ont aucune occurrence dans "$\alpha \rightarrow$ a", se fait par la formule suivante :

$$E\left(V_0\right) = \sum_{i=1}^{n} V_i \left(\frac{N - N_a}{N} \right)^i$$

Considérons maintenant un point "b" à droite de "a". L'accroissement théorique entre le point "a" et le point "b" est calculé par la formule suivante :

$$E\left(V_b \_ V_a\right) = \sum_{i=1}^{n} V_i \left[\left(\frac{N - N_a}{N} \right)^i - \left(\frac{N - N_b}{N} \right)^i \right]$$

2.2. L'accroissement théorique du vocabulaire dans *al-ÞImtâÝ wa l-MuÞânasa*

En nous basant sur le modèle de calcul de l'accroissement théorique du vocabulaire présenté plus haut et construit sur la loi binomiale, nous avons donc calculé pour chaque nuit, une valeur théorique d'accroissement lexical. Autrement dit ce que nous avons calculé c'est ce que chaque nuit devrait apporter comme vocables nouveaux si le lexique de TawÎîdî était resté stable pendant toute la période d'écriture d'*al-ÞImtâÝ wa l-MuÞânasa*. Le résultat de ce calcul est présenté dans la partie droite du tableau 1 de la page suivante.

Nous avons ensuite évalué la distance, pour chaque nuit, entre l'accroissement réel et l'accroissement théorique. Cette distance est évaluée par l'écart réduit qui est, rappelons-le, l'écart absolu divisé par l'écart-type. Ces écarts réduits nous ont permis de classer les nuits d'*al-ÞImtâÝ wa l-MuÞânasa* d'après les déviations observées par rapport à l'accroissement théorique du vocabulaire ; le 1er rang est attribué à la nuit qui a l'excédent le plus fort (la nuit 0 avec le seul excédent qui est de 1,29), le dernier rang à celle ayant le déficit le plus fort (la nuit 13 avec un déficit de -6,65).

Nous allons comparer ce classement obtenu sur la base de l'accroissement du vocabulaire à ceux que fournissent d'un côté l'ordre chronologique des nuits, de l'autre côté l'étendue relative du vocabulaire. L'ordre chronologique des nuits étant l'ordre présumé d'écriture des nuits, du préambule (nuit 0) jusqu'à la nuit 16 (avec l'absence des nuits 11 et 12) pour notre corpus qui est le premier volume d'*al-ÞImtâÝ wa l-MuÞânasa*, nous avons aussi classé les nuits sur la base de l'étendue relative du vocabulaire.

Accroissement du vocabulaire dans *al-ÞImtâÝ wa l-muÞânasa*

	Étendue		Vocabulaire					Accroissement				
	N	N cumulé	Réel		Théorique		Écart	Réel		Théorique		Écart réduit
			Voc.	Voc. cumulé	Voc.	Voc. cumulé		Acc.	Acc. cumulé	Acc.	Acc. cumulé	
N00	5 062	5 062	1 412	1 412	1 398	1 398	14	1412	1 412	1 203,75	1 203,75	1,29
N01	2 478	7 540	703	1 816	818	2 216	- 115	404	1 816	630,81	1 839,40	- 2,78
N02	3 115	10 655	896	2 290	974	3 190	- 78	474	2 290	793,05	2 634,10	- 2,86
N03	2 004	12 659	627	2 583	694	3 884	- 67	293	2 583	572,34	3 208,26	- 3,42
N04	4 248	16 907	1 133	3 148	1 229	5 113	- 96	565	3 148	975,96	4 156,57	- 3,25
N05	906	17 813	340	3 261	369	5 482	- 29	113	3 261	321,94	4 481,50	- 3,52
N06	7 079	24 892	1 644	4 057	1 777	7 259	- 133	796	4 057	1 306,82	5 636,57	- 3,29
N07	2 569	27 461	688	4 272	841	8 100	- 153	215	4 272	619,78	6 222,48	- 4,87
N08	10 788	38 249	1 967	5 063	2 373	10 473	- 406	791	5 063	1 361,78	6 992,83	- 4,20
N09	4 607	42 856	1 003	5 348	1 305	11 778	- 302	285	5 348	842,20	7 606,47	- 5,82
N10	9 564	52 420	1 772	6 217	2 188	13 966	- 416	869	6 217	1 225,39	7 513,42	- 2,81
N13	2 427	54 847	535	6 305	1 751	15 717	- 1216	88	6 305	465,74	7 763,13	- 6,65
N14	3 271	58 118	819	6 480	1 011	16 728	- 192	175	6 480	713,97	7 725,92	- 6,08
N15	1 773	59 891	510	6 574	630	17 358	- 120	94	6 574	461,17	7 725,72	- 5,34
N16	1 286	61 177	420	6 652	489	17 847	- 69	78	6 652	399,71	8 125,44	- 4,94

Tableau 86

Corrélation des Rangs entre les classements des nuits selon l'ordre chronologique, l'étendue relative du vocabulaire et l'accroissement lexical

	Classement		
	Chronologique	**Étendue relative**	**Accroissement**
Nuit 00	1	4	1
Nuit 01	2	9	2
Nuit 02	3	7	4
Nuit 03	4	11	7
Nuit 04	5	5	5
Nuit 05	6	15	8
Nuit 06	7	3	6
Nuit 07	8	10	10
Nuit 08	9	1	9
Nuit 09	10	6	13
Nuit 10	11	2	3
Nuit 13	12	12	15
Nuit 14	13	8	14
Nuit 15	14	13	12
Nuit 16	15	14	11

Coefficient de Spearman

	Accroissement
Étendue relative	0,446
Chronol.	**0,782**

Tableau 87

Pour la corrélation entre le classement chronologique et celui selon l'accroissement du vocabulaire, le coefficient de Spearman est de 0,782, ce qui veut dire qu'au seuil de signification $\alpha = 0,05$, on peut rejeter l'hypothèse nulle d'absence de corrélation. Autrement dit, la corrélation est bien significative, et la disperssion des points sur le graphique de la partie gauche de la figure 2 le montre très clairement.

En revanche, pour l'autre couple, celui du classement selon l'accroissement du vocabulaire et du classement selon l'étendue relative du vocabulaire, le coefficient de corrélation des rangs de Spearman est de 0,446, ce qui nous permet de conclure qu'au seuil de signification $\alpha = 0,05$, on ne peut pas rejeter l'hypothèse nulle d'absence de corrélation. En d'autres termes, la corrélation n'est pas significative, et la dispersion très étalée des points sur le graphique de la partie droite de la figure 3 le dit davantage.

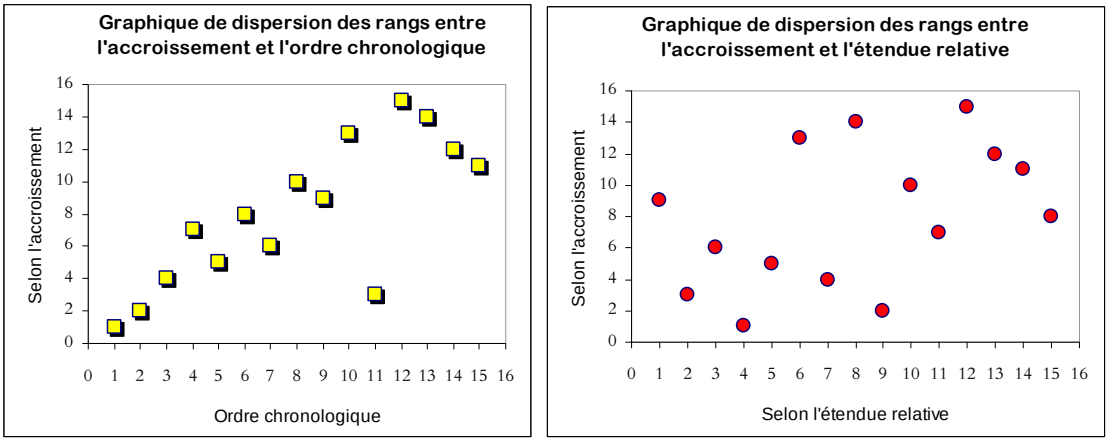


Figure 134
Corrélation des rangs entre les classements selon
l'accroissement lexical, l'étendue relative et l'ordre chronologique

Le contraste, voire même l'opposition qui existe entre les écarts des rangs pour les deux couples de classement, se traduit selon Ch. Muller, par le changement de thème (Muller, 1967 : 85). En effet, les nuits qui sont plus riches en vocables qu'en vocables nouveaux ne montrent pas de renouvellement de vocabulaire. En revanche, celles qui présentent un vocabulaire plutôt restreint et où l'apport en vocables nouveaux est moins déficitaire, accusent un renouvellement lexical notable.

Ce constat peut être confirmé en examinant de plus près l'évolution de la courbe de l'accroissement du vocabulaire sur tout le corpus d'*al-ÞImtâÝ wa l-MuÞânasa*, c'est-à-dire l'allure de la courbe de l'accroissement cumulé qui va, en croissant, de celui de la première nuit, correspondant à l'étendue de son vocabulaire, jusqu'à celui de la dernière nuit pour atteindre en fin de parcours l'étendue V du vocabulaire de tout le corpus.

Il est clair du graphique de la figure 4 que la courbe de l'accroissement cumulé du vocabulaire dans *al-ÞImtâÝ wa l-MuÞânasa* bien qu'elle commence légèrement au-dessus de celle de l'accroissement théorique, reste tout au long de sa trajectoire au-dessous de cette dernière. Ce qui traduit un renouvellement lexical beaucoup moins important que ce qu'on aurait pu avoir théoriquement. Cet écart entre l'accroissement réel et l'accroissement théorique, n'est pas constant à mesure qu'on avance dans le corpus ; il rétrécit à certains endroits du graphique comme, par exemple, entre les deux points de coordonnées (52 420, 6 217) et (52 420, 7 513). L'écart entre les deux courbes

est d'autant plus variable que les courbes elles-mêmes fluctuent plus ou moins amplement.

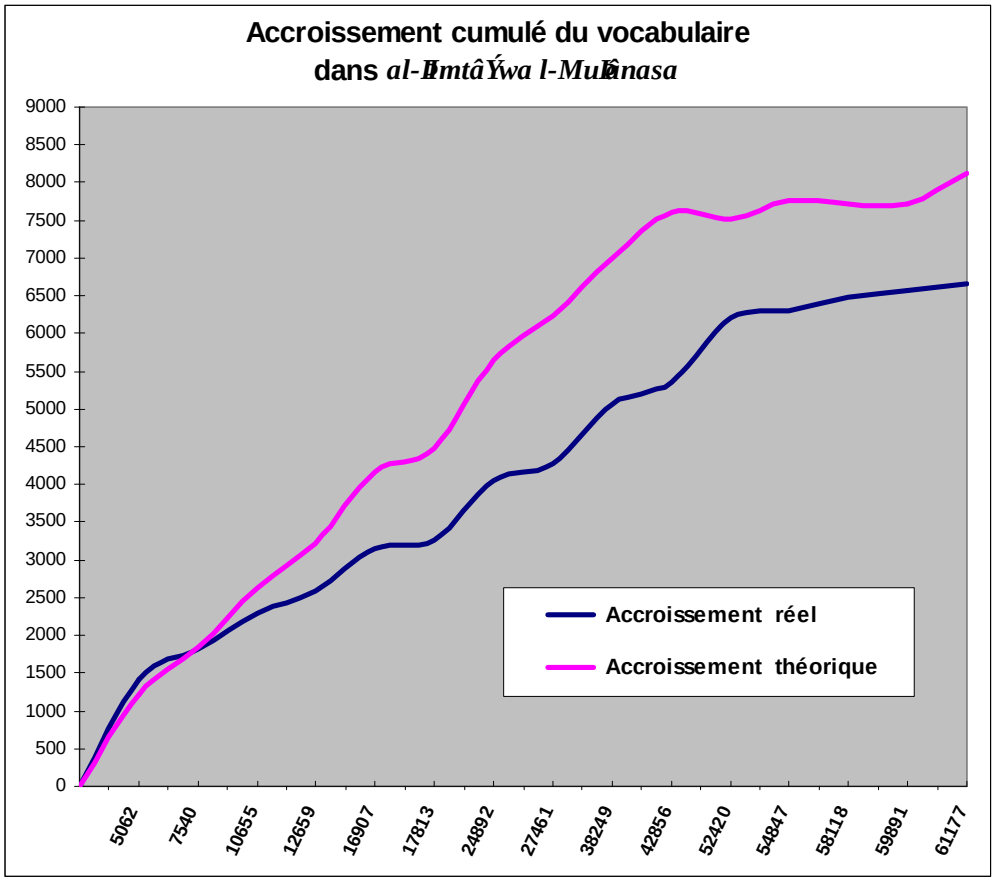


Figure 135

Cette fluctuation apparente tant au niveau de la courbe d'accroissement réel cumulé qu'au niveau de l'accroissement théorique cumulé, qui traduit une irrégularité dans le renouvellement lexical, s'explique par deux facteurs : le premier est la double disparité qui existe, d'un côté entre les étendues de certaines nuits contiguës comme, par exemple, la nuit 4 et la nuit 5 ou entre celle-ci et la nuit 6, et de l'autre entre les étendues relatives du vocabulaire de certaines nuits contiguës dont nous avons rendu compte plus haut. Le deuxième facteur est une sorte d'alternance entre rupture et continuité thématiques où, pour certaines nuits mitoyennes, passant d'une nuit à une autre, l'on change totalement de thème comme pour les nuits 6 et 7, les nuits 8 et 9 ou les nuits 1 et 2 ; mais dans d'autres endroits, l'on remarque une quasi unité thématique

entre quelques autres comme les nuits 2 et 3, les nuits 4 et 5 ou les nuits 13, 14 et 15. Pour ces dernières nuits, la continuité thématique se manifeste, au niveau de la courbe d'accroissement réel du vocabulaire, par des pentes très douces entre les points de coordonnées (10 655, 2 290) et (12 659, 2 583), (16 907, 3 148) et (17 813, 3 261), (54 847, 6 305) et (58 118, 6 480), et enfin (58 118, 6 480) et (59 891, 6 574).

À l'examen de la figure 5 représentant les deux courbes d'accroissement réel et théorique du vocabulaire, nous remarquons que, mis à part de l'accroissement du préambule (nuit 0) où les deux courbes chevauchent et de celui des deux dernière nuits où l'on observe un léger changement d'amplitude, les deux courbes suivent pratiquement le même mouvement et les mêmes oscillations ; la seule différence notable entre les deux courbe est le décalage au niveau des valeurs de l'accroissement lexical où la courbe de l'accroissement réel reste en deçà de celle de l'accroissement théorique.

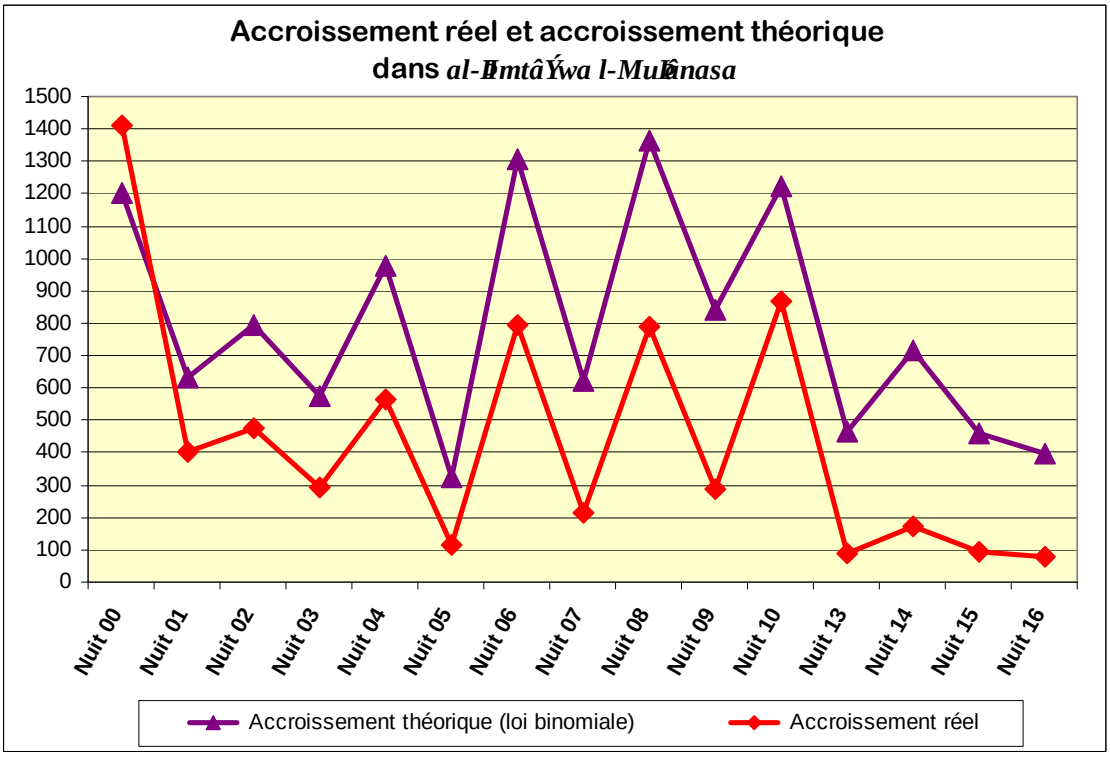


Figure 136

Cette similitude entre les deux courbes au niveau de l'allure générale qui traduit le mouvement de l'accroissement du vocabulaire d'*al-ÞImtâÝ wa l-MuÞânasa*, nous oblige à émettre deux hypothèses pour pouvoir l'expliquer : soit l'accroissement du vocabulaire d'*al-ÞImtâÝ wa l-MuÞânasa* est parfaitement constitué vu la structure lexicale du corpus et le renouvellement thématique qui lui est propre, soit le modèle théorique est trop influencé par les données réelles du corpus sur lesquelles sont basés les calculs théoriques. Ce lien direct entre les valeurs calculées théoriquement et les données réelles du texte étudié a d'ailleurs poussé Charles Muller, lui-même, au moment de la mise en pratique de son modèle, à dire qu'il ne faut pas crier victoire en précisant : « N'oublions pas que nos effectifs théoriques sont calculés en partant d'une distribution réelle, celle du texte entier, qui contient en elle les effectifs réels de la partie du texte »[314].

[314] Charles Muller, *Initiation ...*, *op. cit.*, p.186

Chapitre 12

Les catégories lexicales

Les faits de la structure lexicale tels que la richesse lexicale, l'accroissement du vocabulaire, la distribution des fréquences etc., sont importants pour étudier le vocabulaire d'un auteur ; mais la répartition des mots en catégories lexicales en représente un élément d'une importance capitale puisqu'elle ajoute à ce panorama structurel un aspect relatif au contenu lexical, contenu défini par la nature des unités qui forment le vocabulaire. Le choix, par un écrivain, de ces unités n'est pas arbitraire, c'est un choix réfléchi et conscient. L'étude de la distribution des catégories lexicales permet donc de saisir cette division consciente et authentique où « l'on ne peut pas ne pas rejoindre le choix conscient de l'écrivain »[315].

Mais pour pouvoir saisir la partition de ces unités composant le vocabulaire, il faut les définir d'abord pour les reconnaître et les délimiter ensuite. Les critères de reconnaissance des éléments du vocabulaire et de leur affectation à telle ou telle catégorie lexicale sont multiples et peuvent varier, ils peuvent se fonder, exclusivement ou conjointement, sur tel ou tel aspect morphologique, sur des considérations sémantiques, sur les relations logiques, sur la distribution des unités au niveau de l'axe syntagmatique ...

Nos choix concernant l'affectation de tel ou tel vocable à telle ou telle catégorie lexicale ont été présentés plus haut. Les choix de catégorisation, qui font partie de la norme de lemmatisation au sens large, sont évidemment étroitement liés à ceux faits dans le cadre de la norme de segmentation ; ces deux normes sont exposées dans la norme lexicologique que nous avons proposée au chapitre 5.

En outre, la question des limites entre les catégories lexicales est une question qui reste toujours posée dans ce type d'entreprise qui est la catégorisation. En effet, les frontières, par exemple, entre le participe actif et le nom construit sur le même schème par un procédé de substantivation, restent toujours floues surtout lorsque il est question de textes arabes classiques. Ce qui met en exergue le problème de l'ambiguïté

[315] Brunet Etienne, *Le vocabulaire français de 1789 à nos jours d'après les données du Trésor de la langues française*, 1981, t. 1, p. 295.

polycatégorielle et donc celui de la non exclusivité des partitions lexicales ; cela a pour conséquence de perturber, si l'on n'y prend pas garde, l'opération de catégorisation.

Mais, même si les catégories lexicales ne forment pas toujours des partitions exclusives et même si les risques d'ambiguïtés mono- et polycatégorielles, virtuelles et effectives, guettent à tout moment cette opération de catégorisation, nous avons pu, grâce à la solidité et à la stabilité de la norme lexicologique que nous avons adoptée, sortir indemne de la turbulence catégorielle et avons pu mener à bien notre lemmatisation contextuelle tant au niveau de l'identification qu'au niveau de la catégorisation.

Nous avons pu, en effet, résoudre facilement toutes les difficultés rencontrées lors de la lemmatisation avec ses deux volets, l'identification et la catégorisation, grâce notamment à une bonne concordance entre la norme de lemmatisation et la définition, claire, des catégories lexicales qui nous a permis d'affecter à chaque vocable une seule et unique catégorie lexicale ; surtout que toute étude des catégories lexicales « ne pourra s'intégrer à une analyse lexicale que si l'on définit les catégories de telle façon que chaque unité de lexique appartienne à une catégorie et une seule »[316].

La définition, les subdivisions et la composition de chaque catégorie lexicale ont été présentées en détail dans la partie concernant la catégorisation[317]. Mais avant d'étudier la répartition des catégories lexicales, rappelons ici sommairement ces catégories qui font l'objet d'un traitement quantitatif :

- Les catégories de **mots lexicaux** : Elles comportent les verbes, les noms (primitifs et dérivés) et les adjectifs.
- Les catégories de **mots fonctionnels** : Les mots fonctionnels ou mots-outils sont composés des particules, des noms-outils, des verbes fonctionnalisés et des mots-outils composés.

[316] Ch. Muller, *Principes et méthodes de statistique lexicale*, 1992, p. 170.

[317] Voir, dans le chapitre 5 « *Norme de dépouillement* », la section 4 « *catégorisation* » pp. 260-302.

1. Lexicalité et fonctionnalité

La distribution, dans le discours, des unités lexicales et des unités fonctionnelles est une question qui a toujours suscité une attention particulière chez les lexicologues comme chez les lexicométriciens.

Étant donné que le choix du locuteur porte sur un éventail considérable, une liste ouverte, des mots lexicaux, alors que le nombre des mots fonctionnels est limité, leur liste étant fermée, l'on doit s'attendre à ce que, dans le discours, le nombre (en occurrences ou en vocables) des mots lexicaux soit relativement moins important que celui des mots fonctionnels ; dans le meilleur des cas ils se partagent, à parts égales, le nombre total des mots d'un texte. Ce constat est vérifié dans grand nombre de langues ayant connu des études lexicométriques. Voyons comment se présente le rapport entre lexicalité et fonctionnalité en arabe pour *al-ÞImtâÝ wa l-MuÞânasa*, aussi bien au niveau du corpus en entier qu'au niveau de ses parties : les Nuits.

1.1. Lexicalité et fonctionnalité au niveau du corpus

Au niveau du corpus, les mots lexicaux sont au nombre de 24 016 occurrences représentant près de 40 % de l'ensemble des occurrences d'*al-ÞImtâÝ wa l-MuÞânasa*. Ces 24 016 occurrences sont fournies par 6 414 vocables, c'est-à-dire avec une fréquence moyenne de 3,74. Ils enregistrent 3 380 *hapax* et une fréquence maximale de 714 occurrences correspondant au lemme قَالَ « *qâla* ».

	N	% tage	V	*Hapax*	Freq. max.	Lemme
Mots lexicaux :	24 016	*39,26 %*	6 414	3 380	714	قَالَ
Mots-Outils :	37 161	*60,74 %*	241	65	7 825	ال

Les mots fonctionnels, quant à eux, ils atteignent 37 161 occurrences dépassant ainsi de peu les 60 % du nombre total des occurrences du corpus. Il n'y a que 241 vocables qui ont fourni ce grand nombre de 37 161 occurrences ; ce qui nous donne une fréquence moyenne de 154,20. Les mots fonctionnels ne comptent que 65 *hapax* et la fréquence maximale enregistrée est de 7 825 correspondant à l'article ال « *al* ».

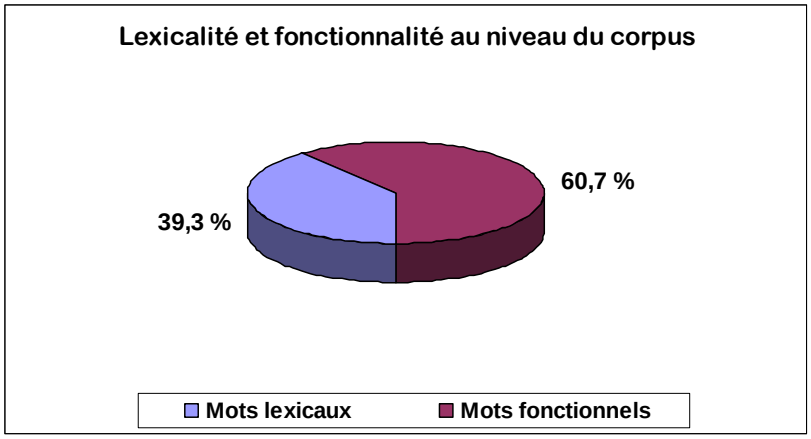


Figure 137

Dans *Maqâmât al-HamaÆânî*[318], les rapports sont de 44 % pour les mots lexicaux, et de 56 % pour les mots fonctionnels. Ces proportions sont données à titre indicatif car la norme lexicologique de Ch. Ayadi n'est pas clairement énoncée.

1.2. Lexicalité et fonctionnalité au niveau des nuits

Au niveau des Nuits, le coefficient de lexicalité varie de 37,57 % (Nuit 14) à 43,28 % (Nuit 10) avec une moyenne de 38,93 %.

Celui de fonctionnalité varie de 56,72 % (Nuit 10) à 62,43 % (Nuit 14) avec une moyenne de 61,07 %.

Les coefficients de lexicalité et de fonctionnalité de chacune des Nuits sont présentés dans le tableau suivant.

[318] Voir : Ch. Ayadi, *op. cit.*

Mots lexicaux et mots fonctionnels dans les nuits d'*al-ÞImtâÝ wa l-MuÞânasa*				
	Mots lexicaux	**Mots-Outils**	***lexicalité***	***Fonctionnalité***
N00	1 945	3 117	***38,42%***	***61,58%***
N01	947	1 531	***38,22%***	***61,78%***
N02	1 201	1 914	***38,56%***	***61,44%***
N03	792	1 212	***39,52%***	***60,48%***
N04	1 628	2 620	***38,32%***	***61,68%***
N05	350	556	***38,63%***	***61,37%***
N06	2 724	4 355	***38,48%***	***61,52%***
N07	997	1 572	***38,81%***	***61,19%***
N08	4 141	6 647	***38,39%***	***61,61%***
N09	1 774	2 833	***38,51%***	***61,49%***
N10	4139	5 425	***43,28%***	***56,72%***
N13	972	1 455	***40,05%***	***59,95%***
N14	1 229	2 042	***37,57%***	***62,43%***
N15	688	1 085	***38,80%***	***61,20%***
N16	489	797	***38,02%***	***61,98%***
	24 016	**37 161**	***39,26%***	***60,74%***

Tableau 88

Nous présentons le rapport des deux coefficients pour chaque Nuit dans le graphique suivant :

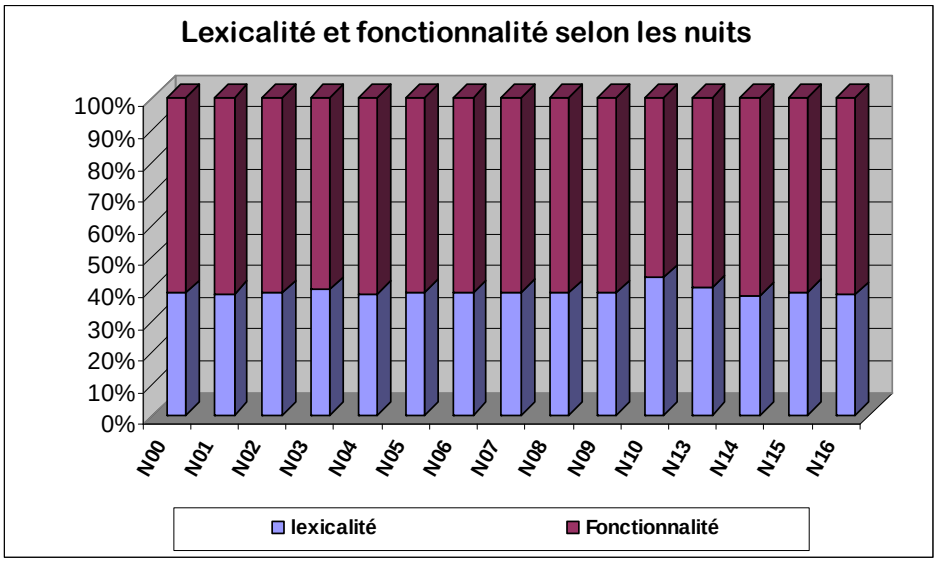


Figure 138

Par rapport à la valeur moyenne de lexicalité, la grande majorité des Nuits, 12 Nuits sur 15, soit 80 % présentent un coefficient de lexicalité inférieur au coefficient moyen qui est de 38,93 % ; autrement dit 80 % des Nuits sont déficitaires en mots lexicaux. Seulement 3 Nuits, soit 20 % des Nuits ont un coefficient de lexicalité supérieur au coefficient moyen, c'est-à-dire qu'elles sont excédentaires en mots lexicaux ; il s'agit des Nuits 3, 10 et 13.

Comme lexicalité et fonctionnalité sont complémentaires, la tendance observée au niveau de la lexicalité est l'inverse de celle au niveau de la fonctionnalité. Par conséquent, 80 % des Nuits sont excédentaires en mots fonctionnels et seulement 20 % y sont déficitaires.

La figure 3 représente l'écart de lexicalité et de fonctionnalité des Nuits par rapport à la moyenne. On y voit également la complémentarité des deux mesures représentée par la symétrie des deux courbes autour de la moyenne, et où la tendance déficitaire d'une courbe est contrecarrée par la tendance excédentaire de l'autre et *vice versa*.

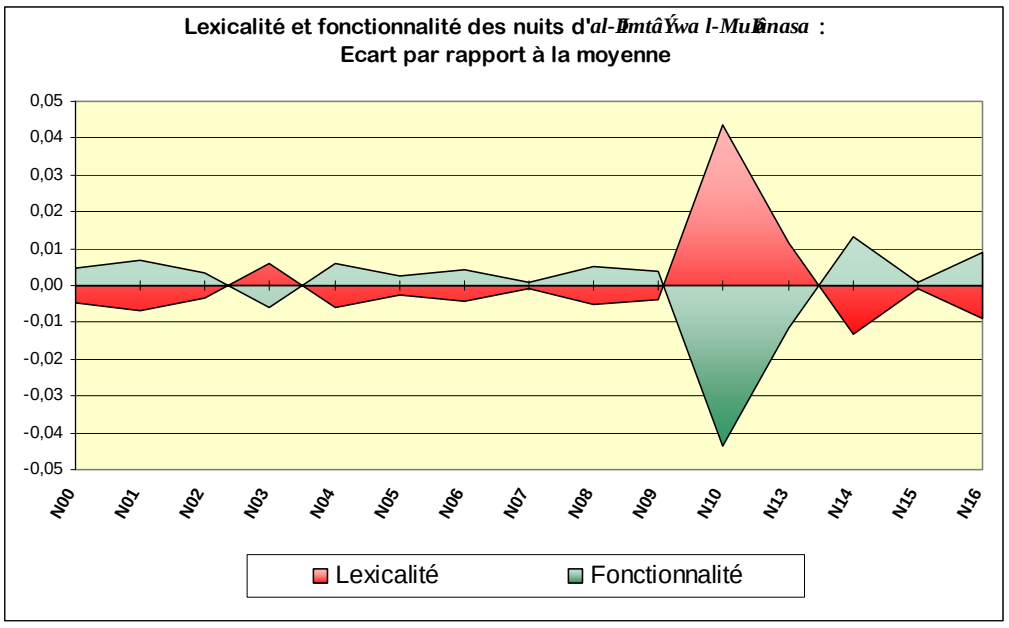


Figure 139

2. Catégories lexicales au niveau du corpus

Nous venons de voir comment se répartissent les mots d'*al-ÞImtâÝ wa l-MuÞânasa* en fonction de leur lexicalité ou de leur fonctionnalité. Mais quelle peut être leur répartition selon les catégories lexicales au sein de l'univers lexical et de l'univers des mots fonctionnels ?

Dans l'univers lexical, la plume de TawÎîdî était favorable à la catégorie nominale puisque celle-ci présente un effectif de 10 560 occurrences représentant ainsi 43,97 % des mots lexicaux et 17,23 % de l'ensemble des occurrences du corpus. Les 10 560 occurrences nominales correspondent à 2 203 vocables. Chaque vocable enregistre ainsi une fréquence moyenne de 4,79. Les *hapax* sont au nombre de 1 007 et la fréquence maximale est de 211 correspondant au nom نفْس « *nafs* »[319].

Effectifs des catégories lexicales dans *al-ÞImtâÝ wa l-MuÞânasa*

	Effectif	% à l'ensemble	% interne
Mots lexicaux			
Verbes	6997	11,44%	29,13%
Noms primitifs	10560	17,23%	43,97%
Noms dérivés	2015	3,29%	8,39%
Adjectifs	4444	7,26%	18,50%
Total des Mots lexicaux	**24016**	**39,26%**	**100,00%**
Mots fonctionnels			
Particules	26343	43,06%	70,89%
Noms-Outils	10524	17,20%	28,32%
Verbes fonctionnalisés	172	0,28%	0,46%
Mots-Outils composés	122	0,20%	0,33%
Total des Mots fonctionnels	**37161**	**60,74%**	**100,00%**
Etendu du corpus	**61177**	**100,00%**	

Tableau 89

[319] On trouvera dans l'annexe G les principales caractéristiques lexicométriques de toutes les catégories et sous-catégories lexicales.

Après les noms primitifs, la plume de TawÎîdî cultivait les verbes. Ils regroupent 6 997 occurrences, c'est-à-dire 29,13 % des mots lexicaux et 11,44 % de l'ensemble des occurrences d'*al-ÞImtâÝ wa l-MuÞânasa*. Cette catégorie compte 1 731 vocables ayant fourni les 6 997 occurrences verbales. Autrement dit, chaque vocable présente une fréquence moyenne de 4,04. Les *hapax* sont au nombre de 896 et la fréquence maximale de 714 correspondant au verbe قَالَ « *qâla* ».

En troisième position, les adjectifs comptent 4 444 occurrences, soit 18,5 % des mots lexicaux et 7,26 % du total des occurrences. La fréquence moyenne des adjectifs est de 2,73 puisqu'il y a 1 627 vocables qui ont fourni les 4 444 occurrences adjectivales. Les *hapax* sont au nombre de 962 et la fréquence maximale de 93 correspondant à l'adjectif صاحِب « *ÒâÎib* ».

Les noms dérivés, eux, occupent la dernière place des mots lexicaux puisqu'ils ne regroupent que 2 015 occurrences, soit 8,39 % des mots lexicaux et seulement 3,29 % de l'ensemble des occurrences d'*al-ÞImtâÝ wa l-MuÞânasa*. Les noms dérivés ne sont pas très répétés puisque les 2 015 occurrences qu'ils comptent sont fournies par 853 vocables ; ce qui donne une fréquence moyenne de 2,36 , c'est-à-dire de moitié inférieure à celle des noms primitifs. Les *hapax* des noms dérivés sont au nombre de 515 et la fréquence maximale de 89 correspondant au déverbal مَعنًى « *maÝnâ* ».

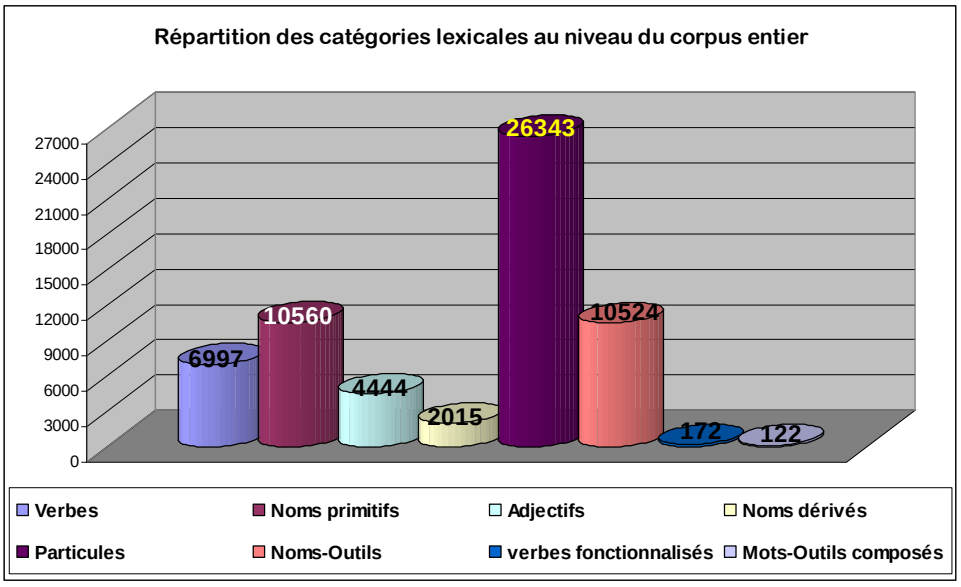


Figure 140

Dans l'univers des mots fonctionnels, la part du lion revient aux particules qui comptent 26 343 occurrences représentant 70,89 % des mots fonctionnels et 43,06 % de l'étendue totale du corpus. Par ailleurs, ces 26 343 occurrences ne sont fournies que par 57 vocables, ce qui donne à la fréquence moyenne une valeur très élevée de 462,16. On ne trouve, dans les particules, que 7 *hapax*. La fréquence maximale des particules est celle à la fois de tous les mots fonctionnels et du corpus entier, c'est-à-dire la fréquence de 7 825 correspondant à l'article défini ال « *al* ».

Les noms-outils arrivent en deuxième position avec 10 524 occurrences fournies par 159 vocables. Ils représentent 28,32 % des mots fonctionnels et 17,20 % de l'ensemble du corpus. La fréquence moyenne est de 66,19. Les noms-outils enregistrent 44 *hapax* et une fréquence maximale de 2 936 correspondant au pronom affixe de 3ème personne du singulier masculin هُ « *hu* ».

Les deux autres catégories de mots fonctionnels, les verbes fonctionnalisés et les mots-outils composés, n'enregistrent que, respectivement, 172 et 122 occurrences fournies par 7 et 18 vocables, ce qui donne une fréquence moyenne respectivement, de 24,57 et 6,78. Elles comptent 3 et 11 *hapax* et enregistrent une fréquence maximale de 148 (ليس « *laysa* ») et 58 (إنّما « *Þinnamâ*»).

2.1. Les verbes

Quant à la distribution interne à la catégorie des verbes dont le tableau 3 présente les effectifs et les pourcentages, la prédominance est aux verbes trilitères simples qui représentent à eux seuls, 69,80 % de l'ensemble des verbes avec 4 884 occurrences parmi elles 329 *hapax*. Nous y comptons 784 vocables fournissant les 4 884 occurrences, ce qui nous donne une fréquence moyenne de 6,23. Ce sont les verbes trilitères qui fournissent et à l'ensemble des verbes et aux mots lexicaux la fréquence maximale qui est de 714 et correspondant au verbe trilitère concave قَالَ/يَقُولُ « *qâla/yaqûlu* ».

Effectif des verbes dans *al-ÞImtâÝ wa l-MuÞânasa*			
Verbes	**Effectif**	**% à l'ensemble**	**% interne**
Trilitère simple	**4884**	***69,80%***	
Quadrilitère simple	**10**	***0,14%***	
Trilitère augmenté	**2097**	***29,97%***	
Forme II فعَّلَ	327	*4,67%*	*15,59%*
Forme III فَاعَلَ	143	*2,04%*	*6,82%*
Forme IV أَفْعَلَ	624	*8,92%*	*29,76%*
Forme V تَفَعَّلَ	250	*3,57%*	*11,92%*
Forme VI تَفَاعَلَ	143	*2,04%*	*6,82%*
Forme VII اِنْفَعَلَ	81	*1,16%*	*3,86%*
Forme VIII اِفْتَعَلَ	391	*5,59%*	*18,65%*
Forme X اِسْتَفْعَلَ	138	*1,97%*	*6,58%*
Quadrilitère augmenté	**6**	***0,09%***	
Forme IV-2 تَفَعْلَلَ	3	*0,04%*	*50,00%*
Forme IV-3 اِفْعَنْلَلَ	2	*0,03%*	*33,33%*
Forme IV-4 اِفْعَلَلَّ	1	*0,01%*	*16,67%*
Total des Verbes	**6997**	***100,00%***	

Tableau 90

Les verbes trilitères augmentés occupent le deuxième rang avec leurs 2 097 occurrences fournies par 933 vocables donnant ainsi une fréquence moyenne de 2,25. Les verbes trilitères augmentés représentent 29,97 % de l'ensemble des verbes. Le nombre des hapax y est de 554 et la fréquence maximale y est de 45 correspondant au verbe trilitère augmenté de forme IV أَرَادَ « *Þarâda* ».

Alors que les verbes trilitères (simples et augmentés) totalisent 6 981 occurrences (4 884 + 2 097) représentant 99,77 % des verbes, les verbes quadrilitères (simples et augmentés) ne totalisent que 16 occurrences soit 0,23 %.

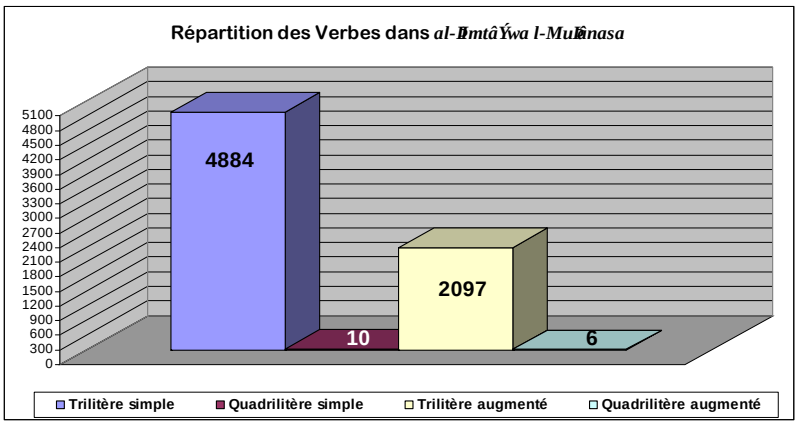


Figure 141

À l'intérieur des verbes trilitères augmentés, la forme IV أَفْعَلَ « *Þaf Ýala* » occupe la 1ère place puisqu'elle représente 29,76 % avec 624 occurrences. La 2ème place est consacrée à la forme VIII اِفْتَعَلَ « *ÞiftaÝala* » avec ses 391 occurrences et 18,65 %. Suivie de la forme II فَعَّلَ « *faÝÝala* » avec un effectif de 327 occurrences et un pourcentage de 15,59 %.

Les formes les moins représentées chez TawÎîdî sont la forme VII اِنْفَعَلَ « *ÞinfaÝala* » et la frome X اِسْتَفْعَلَ « *Þistaf Ýala* » avec des effectifs respectifs de 81 et 138 occurrences et des pourcentages de 3,86 % et 6,58 %.

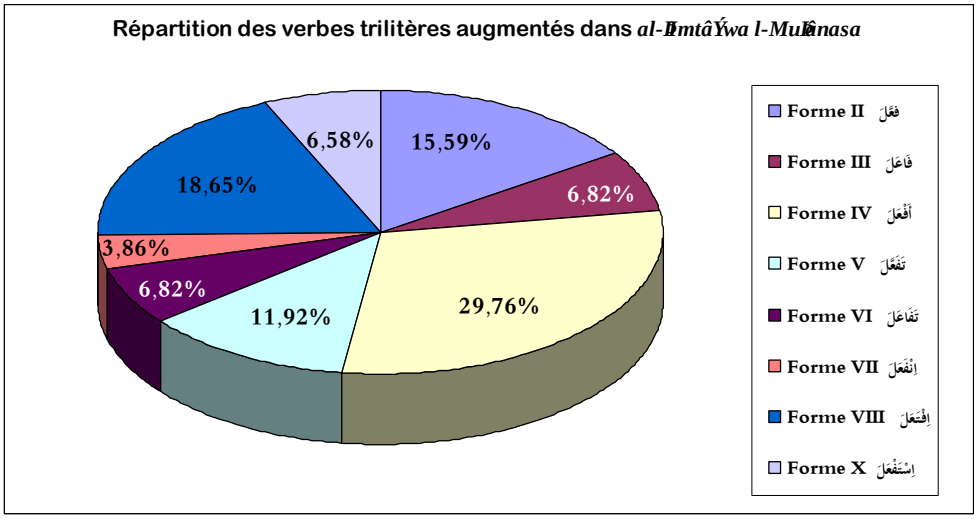


Figure 142

2.2. Les noms primitifs

Dans la catégorie des mots lexicaux la plus importante dans le corpus, la catégorie des noms primitifs, les *maÒdar* primitifs sont les plus fréquents avec 3 936 occurrences représentant 37,27 % de tous les noms primitifs ; à ces occurrences correspondent 952 vocables, ce qui donne une fréquence moyenne de 4,13. Le nombre des *hapax* est de 426, alors que la fréquence maximale est de 79 correspondant au *maÒdar* أمْر « *Þamr* ».

Effectif des noms primitifs dans *al-ÞImtâÝ wa l-MuÞânasa*

Noms primitifs	**Effectif**	***Pourcentage***
Trilitère simple	3808	*36,06%*
Quadrilitère simple	107	*1,01%*
Pentalitère simple	10	*0,09%*
MaÒdar primitif	3936	*37,27%*
Nom d'une fois	50	*0,47%*
Nom de manière	16	*0,15%*
Nom augmenté	2612	*24,73%*
Noms composés	21	*0,20%*
Total des Noms primitifs	**10560**	***100,00%***

Tableau 91

L'effectif des noms trilitères simples n'est pas moins important que celui des *maÒdar* primitifs. Avec une fréquence moyenne de 8,81 , les 432 vocables fournissent 3 808 occurrences, soit 36,06 % de tous les noms primitifs d'*al-ÞImtâÝ wa l-MuÞânasa*. Les *hapax* sont au nombre de 160 et la fréquence maximale de 211 correspondant au nom trilitère simple نفْس « *nafs* ».

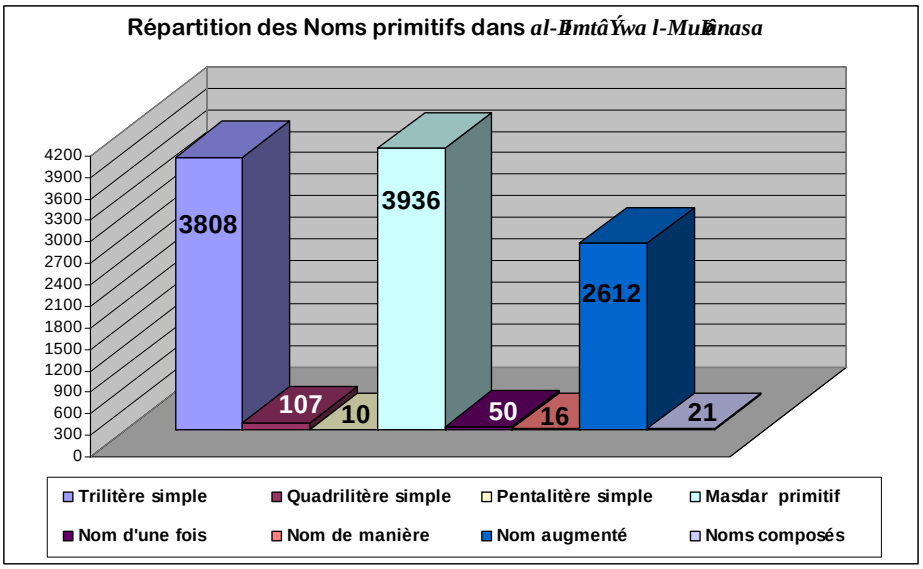


Figure 143

En troisième position des noms primitifs les plus fréquents, se trouvent les noms augmentés ; ils ont un effectif de 2 612 occurrences, soit 24,73 %, fournies par 733 vocables. La fréquence moyenne y est de 3,56 , les *hapax* de 366 et la fréquence maximale de 112 correspondant au nom augmenté إنسان « *Þinsân* ».

Les noms primitifs les moins représentés sont les noms pentalitères simples (10 occ.), le nom de manière (16 occ.) et les noms composés (21 occ.).

2.3. Les adjectifs

Les adjectifs viennent, nous l'avons vu, en 3ème position des mots lexicaux les plus fréquents. À l'intérieur de cette catégorie, les adjectifs les plus fréquents sont les *Òifat mušabbahat* qui représentent 35,80 % de l'ensemble des adjectifs avec un effectif de 1 591 occurrences fournies par 334 vocables, donc avec une fréquence moyenne de 4,76. L'adjectif le plus fréquent dans cette catégorie est la *Òifat mušabbahat* صاحِب « *ÒâÎib* » avec une fréquence maximale de 93. Les *hapax* sont au nombre de 140.

Effectif des adjectifs dans *al-ÞImtâÝ wa l-MuÞânasa*		
Adjectifs	**Effectif**	***Pourcentage***
Participe actif	1376	*30,96%*
Participe passif	581	*13,07%*
Intensif	101	*2,27%*
Elatif	534	*12,02%*
Ñifa mušabbaha	1591	*35,80%*
adjectif de relation	261	*5,87%*
Total des Adjectifs	**4444**	***100,00%***

Tableau 92

En utilisant 30,96 % des occurrences des adjectifs, les participes actifs arrivent en deuxième position avec 1 376 occurrences fournies par 625 vocables et donc une fréquence moyenne de 2,20. Les participes actifs ayant la fréquence 1 (*hapax*) sont au nombre de 399. Le participe actif مُختلِف « *muÌtalif* » est celui qui a la fréquence maximale de 22.

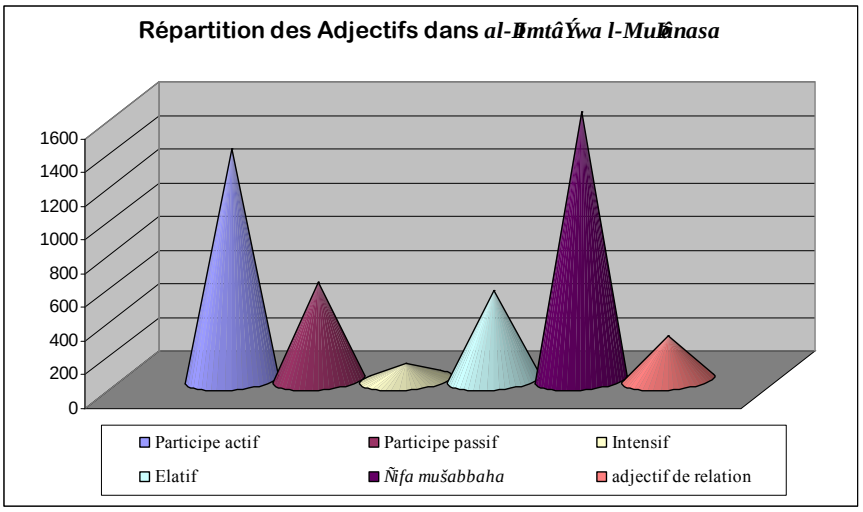


Figure 144

Au milieu du classement, on trouve deux sous-catégories : les participes passifs et les élatifs. Les premiers avec 581 occurrences et 300 vocables, donc une fréquence

moyenne de 1,93 , une fréquence maximale de 23 correspondant au participe passif معروف « *maÝrûf* » et un effectif des *hapax* de 197. Les seconds avec 534 occurrences fournies par 194 vocables, une fréquence moyenne de 2,75 , des *hapax* au nombre de 119 et une fréquence maximale de 87 correspondant à l'élatif آخَر « *ÞâÌar*».

Les adjectifs les moins représentés sont les intensifs avec 101 occurrences et les adjectifs de relation avec 261 occurrences.

2.4. Les noms dérivés

Les *maÒdar* trilitères augmentés dominent la catégorie des noms dérivés en utilisant 64,81 % de l'ensemble des noms dérivés avec un effectif de 1 306 occurrences fournies par 640 vocables. Ils présentent une fréquence moyenne de 2,04, une fréquence maximale de 46 correspondant au *maÒdar* trilitère augmenté عادة « *Ýâda* ». Les *hapax* sont au nombre de 389.

Effectif des noms dérivés dans *al-ÞImtâÝ wa l-MuÞânasa*

Noms dérivés	**Effectif**	***Pourcentage***
Noms de temps/lieu	271	*13,45%*
Nom d'instrument	22	*1,09%*
Diminutif	3	*0,15%*
MaÒdar mîmî	336	*16,67%*
MaÒdar sinâᶜî	59	*2,93%*
MaÒdar trilitère augmenté	1306	*64,81%*
MaÒdar quadrilitère simple	11	*0,55%*
MaÒdar quadrilitère augmenté	7	*0,35%*
Total des Noms dérivés	**2015**	***100,00%***

Tableau 93

Il existe un écart important entre l'effectif des *maÒdar* trilitères augmentés qui occupent la première place et celui des *maÒdar mîmî* qui occupent la deuxième place avec un pourcentage de 16,67 % et un effectif de 336 occurrences fournies par 73 vocables. La fréquence moyenne en est de 4,60 , la fréquence maximale de 89 correspondant au *maÒdar mîmî* معنى « *maÝnâ* » et les *hapax* de 41.

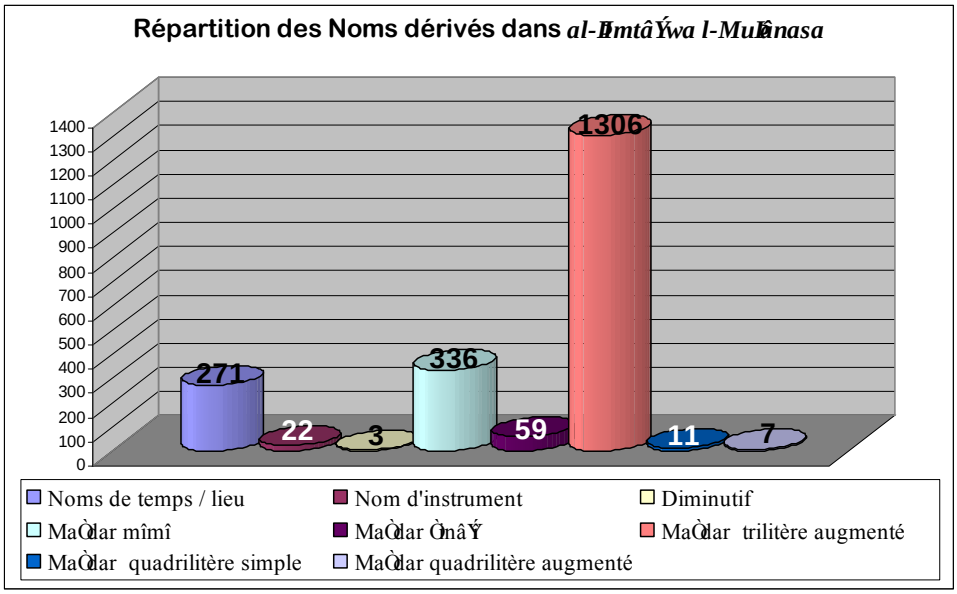


Figure 145

En troisième position, on trouve les noms de temps et de lieu qui ont un effectif de 271 occurrences fournies par 86 vocables, donc une fréquence moyenne de 3,15. Les *hapax* sont au nombre de 52 et la fréquence maximale est de 39 ; elle correspond au nom de lieu مَجلِس « *majlis* ».

Les diminutifs, les *maÒdar* quadrilitères simples et les *maÒdar* quadrilitères augmentés sont très nettement sous-représentés avec seulement 3, 7 et 11 occurrences.

2.5. Les particules

Dans la catégorie des particules, un peloton de tête formé de 3 sous-catégories domine largement toutes les autres : il s'agit des coordonnants, de l'article défini et des prépositions.

Effectif des particules dans *al-ÞImtâÝ wa l-MuÞânasa*		
Particules	**Effectif**	***Pourcentage***
Article défini	7825	*29,70%*
Prépositions	7257	*27,55%*
Coordonnants	8221	*31,21 %*
Particules de négation	910	*3,45 %*
Particules de comparaison	32	*0,12 %*
Particule d'attente ou de certitude	246	*0,93 %*
Particules du futur	8	*0,03 %*
Particules de condition	390	*1,48 %*
Particules de corroboration	773	*2,93 %*
Particule d'espérance et d'apitoiement	10	*0,04 %*
Particules de "*maÒdarité"*	288	*1,09 %*
Particules d'appel	46	*0,17 %*
Particules de réponse	23	*0,09 %*
Particules interrogatives	60	*0,23 %*
Particules d'exception	144	*0,55 %*
Particules d'ouverture	1	*0,004 %*
Particules d'incitation et de remords	16	*0,06 %*
Particules assimilées au verbe	53	*0,20 %*
Particules d'alternative	29	*0,11 %*
Particules d'explication	11	*0,04 %*
Total des Particules	**26343**	***100 %***

Tableau 94

Avec leurs 8 221 occurrences représentant 31,21 % des particules, les coordonnants prennent la tête de ce peloton. Par ailleurs, seulement 7 vocables fournissent les 8 221 occurrences, ce qui donne une fréquence moyenne extrêmement élevée de 1 174,43 ; elle est deux fois et demi supérieure à la fréquence moyenne de

toutes les particules et sept fois et demi à celle de tous les mots-outils. Les coordonnants n'enregistrent aucun *hapax*. Leur fréquence maximale est de 6 718 correspondant au coordonnant و « *wa* ».

Occupant 29,70 % des particules, le deuxième dans le peloton de tête est l'article défini ال « *al* », seul vocable de la sous-catégorie et qui fournit les 7 825 occurrences correspondant à la fréquence maximale à la fois des particules, des mots fonctionnels et de tout le corpus.

Quant à la troisième place du peloton de tête qui s'approprie 27,55 % des particules, elle revient aux prépositions qui regroupent 13 vocables fournissant 7 257 occurrences, ce qui donne une fréquence moyenne de 558,23. La préposition la plus fréquente est بِ « *bi* » avec une fréquence de 1 503. Les prépositions enregistrent 2 *hapax*.

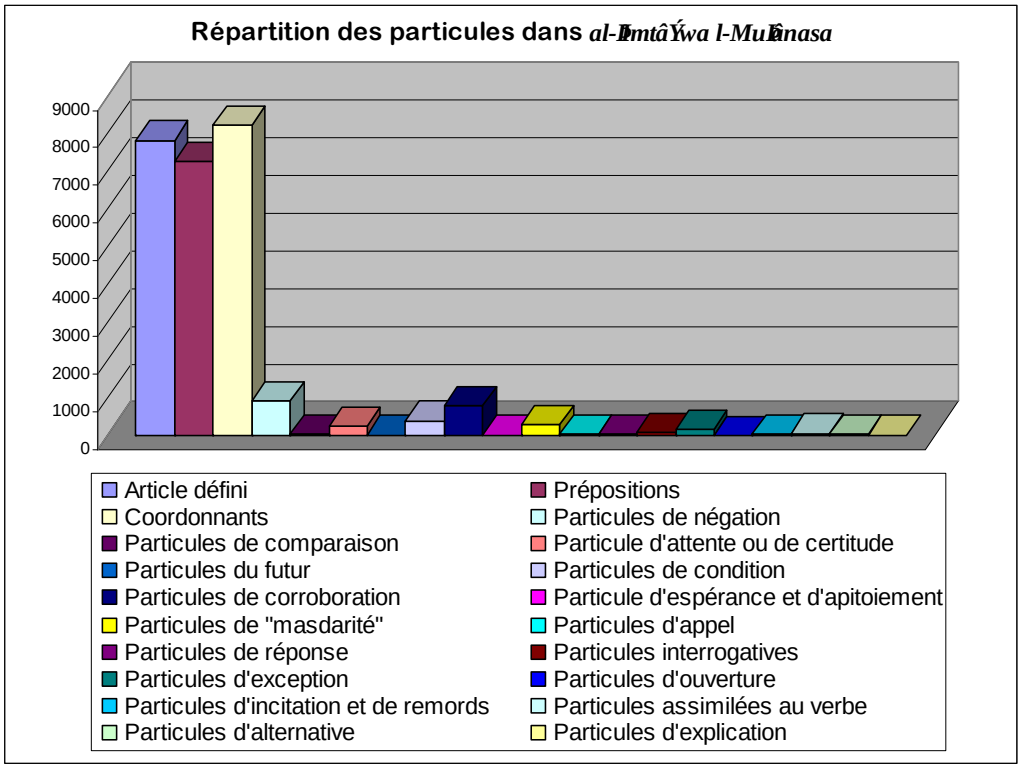


Figure 146

Entre ce peloton de tête et le peloton de queue formé d'un grand nombre de sous-catégories à effectifs faibles, deux sous-catégories intermédiaires : les particules de négation et les particules de corroboration. Les particules de négation enregistrent 910 occurrences, 2 vocables, aucun *hapax*, et une fréquence maximale de 738 correspondant à la particule de négation لا « *lâ* ». Les particules de corroboration enregistrent, quant à elles, 773 occurrences, 4 vocables, un seul *hapax*, et une fréquence maximale de 500 correspondant à la particule de corroboration أنّ « *Þanna* ».

Le peloton de queue regroupe toutes les autres sous-catégories, et qui sont au nombre de 15. Elles enregistrent des effectifs relativement faibles décroissant de 390, pour les particules de condition, jusqu'à 1 pour les particules d'ouverture. Ces 15 sous-catégories réunies enregistrent un effectif moyen de 86,69 occurrences.

2.6. Les noms-outils

La prédominance, au niveau des noms-outils, est à la sous-catégorie des pronoms personnels qui se taillent à eux seuls la part de 59,27 % des noms-outils avec 6 238 occurrences fournies par seulement 19 vocables, ce qui donne une fréquence moyenne de 328,32. Les pronoms personnels n'enregistrent qu'un seul *hapax* ; le pronom le plus fréquent est le pronom affixe de troisième personne du singulier masculin هُ « *hu* » avec 2 936 occurrences.

Loin derrière, on trouve 3 sous-catégories à effectifs rapprochés, les démonstratifs, les adverbes et les relatifs.

Les démonstratifs enregistrent 1 142 occurrences produites par 13 vocables, une fréquence moyenne de 87,85 , deux *hapax* et une fréquence maximale de 562 correspondant au démonstratif هذا « *hâÆâ* ».

Pour les adverbes, on relève 1 064 occurrences fournies par 39 vocables, une fréquence moyenne de 27,28 , 6 *hapax* et une fréquence maximale de 321 correspondant à l'adverbe إذا « *ÞiÆâ* ».

Effectif des noms-outils dans *al-ÞImtâÝ wa l-MuÞânasa*

Noms-Outils	**Effectif**	***Pourcentage***
Démonstratifs	1 142	*10,85%*
Noms interrogatifs	100	*0,95%*
Relatifs	1 050	*9,98%*
Noms de verbes	9	*0,09%*
Nombres cardinaux	275	*68,24%*
Nombres ordinaux	127	*31,51%*
Nombres fractionnaires	1	*0,25%*
Circonlocutifs	32	*0,30%*
Adverbes	1 064	*10,11%*
Pronoms personnels	6 238	*59,27%*
Annectifs	486	*4,62%*
Total des Noms-Outils	**10 524**	***28,32%***

Tableau 95

Les relatifs, quant à eux, enregistrent 7 vocables fournissant 1 050 occurrences, une fréquence moyenne de 150, un seul *hapax* et une fréquence maximale de 664 correspondant au relatif ما « *mâ* ».

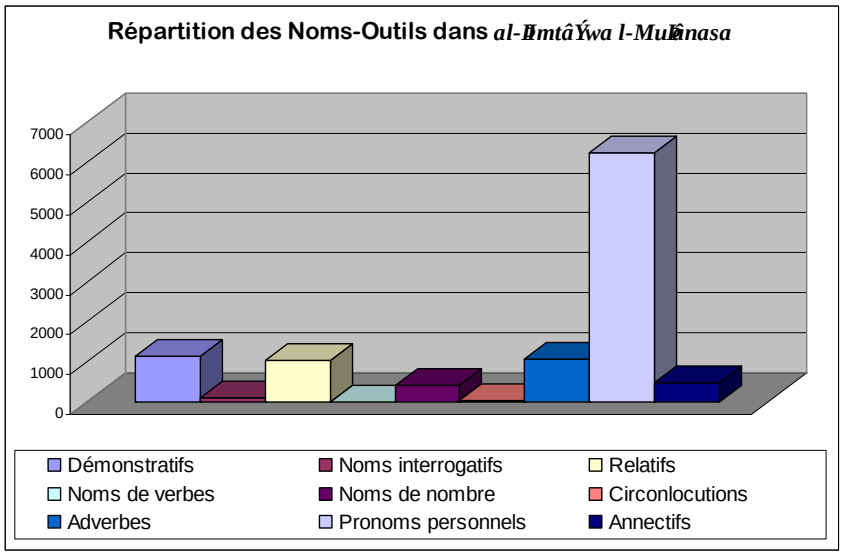


Figure 147

Dans le peloton de queue, on trouve les noms interrogatifs, les circonlocutions et les noms de verbes. Ils enregistrent respectivement, 100, 32 et 9 occurrences. Ces trois sous-catégories sont précédées par deux autres, les annectifs et les noms de nombre.

Les annectifs enregistrent 486 occurrences fournies par 7 vocables ; ils ont une fréquence moyenne de 69,42 , aucun *hapax* et une fréquence maximale de 259 correspondant à l'annectif كُلّ « *kull* ».

Les noms de nombres regroupent 403 occurrences fournies par 60 vocables ; ils ont une fréquence moyenne de 6,72 , 29 *hapax* et une fréquence maximale de 85 correspondant à l'ordinal أَوَّلٌ « *Þawwal*».

2.7. Les verbes fonctionnalisés et les mots-outils composés

Les verbes fonctionnalisés et les mots-outils composés sont, nous l'avons vu plus haut, les deux catégories les moins représentées dans *al-ÞImtâÝ wa l-MuÞânasa*. Elles ne représentent que, respectivement, 0,46 % et 0,33 % de l'étendue du corpus.

Au niveau des verbes fonctionnalisés, la part du lion est réservée aux verbes figés à l'accompli avec 157 occurrences représentant 91,28 % des verbes fonctionnalisés. Les 157 occurrences sont fournies par 5 vocables dont le plus fréquent est le verbe لَيْسَ « *laysa*» avec 148 occurrences ; il n'y a aucun *hapax* pour les verbes figés à l'accompli.

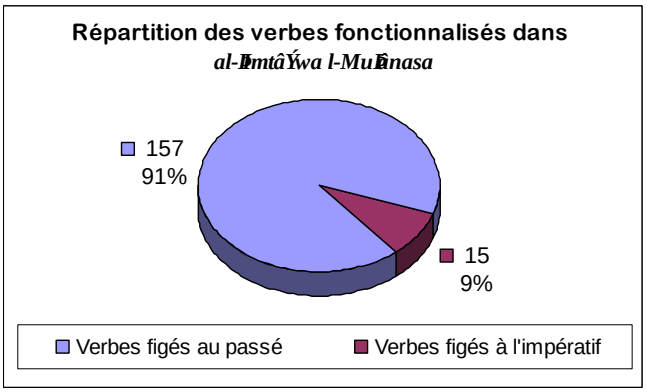


Figure 148

Les 8,72 % restants des verbes fonctionnalisés correspondent aux verbes figés à l'impératif qui représentent seulement 15 occurrences fournies par 2 vocables, un *hapax* et un verbe ayant 14 occurrences, il s'agit du verbe هَاتِ « *hâti*».

Quant aux mots-outils composés, ils enregistrent 155 occurrences fournies par 18 vocables, 11 *hapax* et une fréquence maximale de 58 correspondant au mot-outil composé إنّما « *Þinnamâ*».

3. Catégories lexicales au niveau des nuits

La prédominance des mots fonctionnels sur les mots lexicaux détectée au niveau du corpus entier est bien confirmée au niveau de ses parties, au niveau des Nuits.

Non seulement la domination des mots-outils sur les mots lexicaux est corroborée, mais aussi toute la hiérarchisation des mots lexicaux révélée au niveau du corpus est reproduite au niveau de chaque Nuit où l'on trouve le classement suivant, dans l'ordre croissant des effectifs : les noms dérivés, les adjectifs, les verbes et enfin les noms primitifs.

En plus du tableau 9 présentant la répartition des catégories lexicales selon les Nuits, la figure 12 de la page suivante montre très clairement la hiérarchisation des mots lexicaux et la prédominance des mots-outils.

Répartition des catégories lexicales selon les nuits dans *al-ÞImtâÝ wa l-MuÞânasa*

	Noms dérivés	Adjectifs	Verbes	Noms primitifs	Mots-Outils	Totaux
N00	169	415	519	842	3 117	**5 062**
N01	92	152	319	384	1 531	**2 478**
N02	91	265	313	532	1 914	**3 115**
N03	55	146	277	314	1 212	**2 004**
N04	150	332	521	625	2 620	**4 248**
N05	28	89	111	122	556	**906**
N06	216	505	829	1 174	4 355	**7 079**
N07	92	213	267	425	1 572	**2 569**
N08	461	832	1 224	1 624	6 647	**10 788**
N09	163	313	422	876	2 833	**4 607**
N10	157	576	1 287	2 119	5 425	**9 564**
N13	52	174	256	490	1 455	**2 427**
N14	167	237	308	517	2 042	**3 271**
N15	64	117	206	301	1 085	**1 773**
N16	58	78	138	215	797	**1 286**
Totaux	**2 015**	**4 444**	**6 997**	**10 560**	**37 161**	**61 177**

Tableau 96

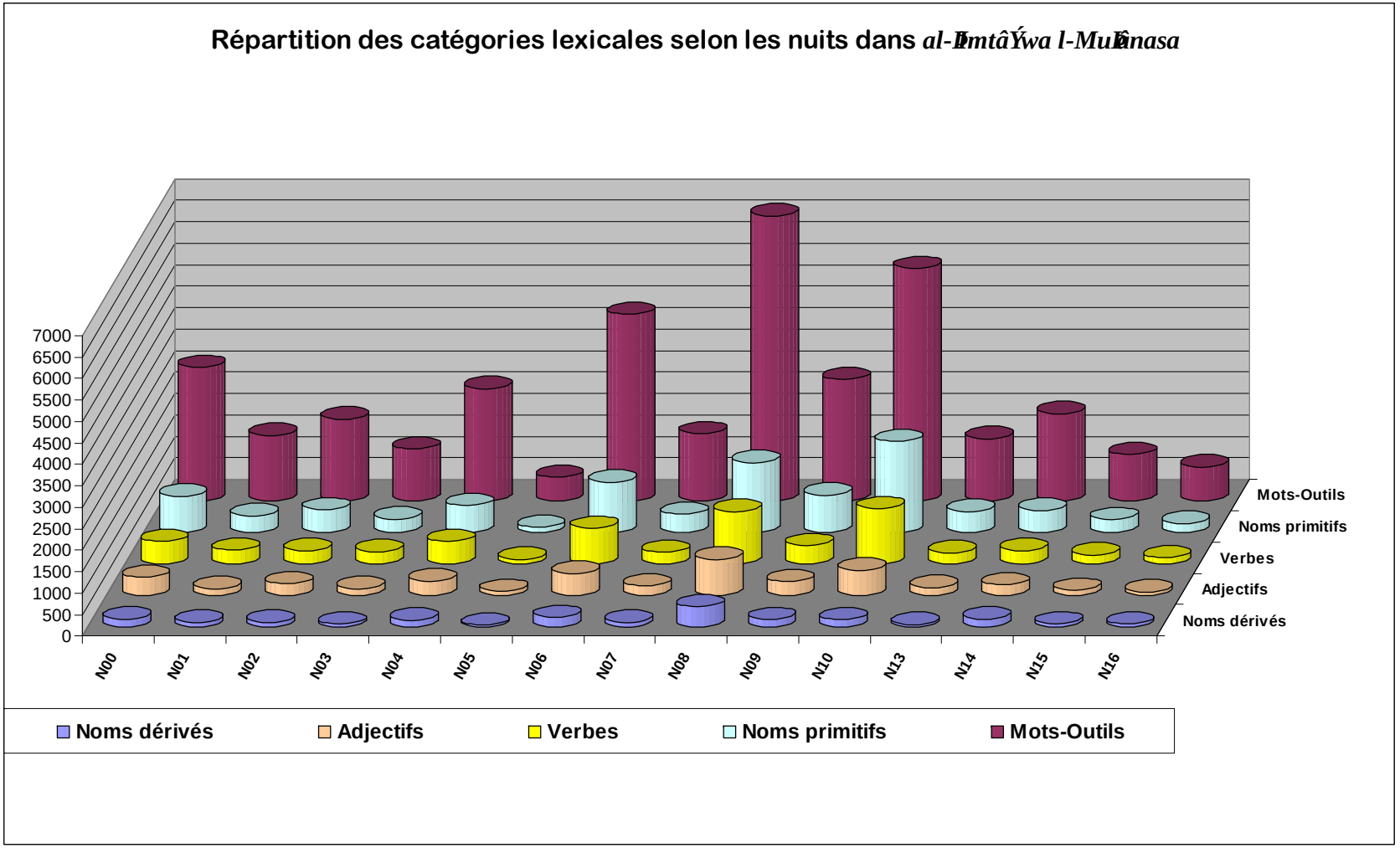

Répartition des catégories lexicales selon les nuits dans al-Imtâ' wa l-Mu'ânasa
7000
6500
6000
5500
5000
4500
4000
3500
3000
2500
2000
1500
1000
500
0
N00
N01
N02
N03
N04
N05
N06
N07
N08
N09
N10
N13
N14
N15
N16
Mots-Outils
Noms primitifs
Verbes
Adjectifs
Noms dérivés
Noms dérivés
Adjectifs
Verbes
Noms primitifs
Mots-Outils

Figure 149

3.1. Les verbes

Au niveau des verbes, la répartition confirme bien, dans chaque Nuit, la prédominance des verbes trilitère simples suivis par les verbes trilitères augmentés.

Répartition des verbes selon les nuits dans *al-ÞImtâÝ wa l-MuÞânasa*

	N0	N1	N2	N3	N4	N5	N6	N7	N8	N9	N10	N13	N14	N15	N16	
V. tri. simple	335	232	222	196	364	80	560	177	849	315	894	196	206	157	101	**4 884**
V. quadri. simple	0	0	0	0	1	0	0	1	3	0	2	0	2	0	1	**10**
Forme II	21	11	19	19	22	10	49	15	70	14	52	9	9	3	4	**327**
Forme III	7	9	3	6	12	4	17	6	30	5	27	4	9	3	1	**143**
Forme IV	69	33	22	26	51	7	74	23	115	27	129	16	15	6	11	**624**
Forme V	19	10	17	8	22	2	22	9	37	22	49	10	12	6	5	**250**
Forme VI	14	5	2	3	13	1	38	5	23	3	16	3	13	4	0	**143**
Forme VII	6	1	3	1	5	1	4	3	12	6	14	4	9	7	5	**81**
Forme VIII	37	11	16	13	22	6	48	19	63	19	78	8	28	14	9	**391**
Forme X	10	7	9	5	9	0	17	9	21	11	23	6	5	6	0	**138**
tafaÝlala	0	0	0	0	0	0	0	0	1	0	1	0	0	0	1	**3**
'if Ýanlala	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	**2**
'if Ýalalla	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**1**
Verbes	**519**	**319**	**313**	**277**	**521**	**111**	**829**	**267**	**1 224**	**422**	**1 287**	**256**	**308**	**206**	**138**	**6 997**

Tableau 97

À l'intérieur de cette dernière sous-catégorie, se sont les verbes de forme IV qui prennent la tête du classement dans 12 Nuits sur 15. Dans la Nuit 5, les verbes de forme IV cèdent la première place à ceux de forme II et, dans les Nuits 14 et 15, à ceux de forme VIII.

Les 3 Nuits qui enregistrent le plus grand effectif de verbes sont dans l'ordre, la Nuit 10, la Nuit 8 et la Nuit 6 avec respectivement, 1 287, 1 224 et 829 occurrences.

Les 3 Nuits les plus pauvres en verbes sont la Nuit 5, la Nuit 16 et la Nuit 15 avec respectivement, 111, 148 et 206 occurrences.

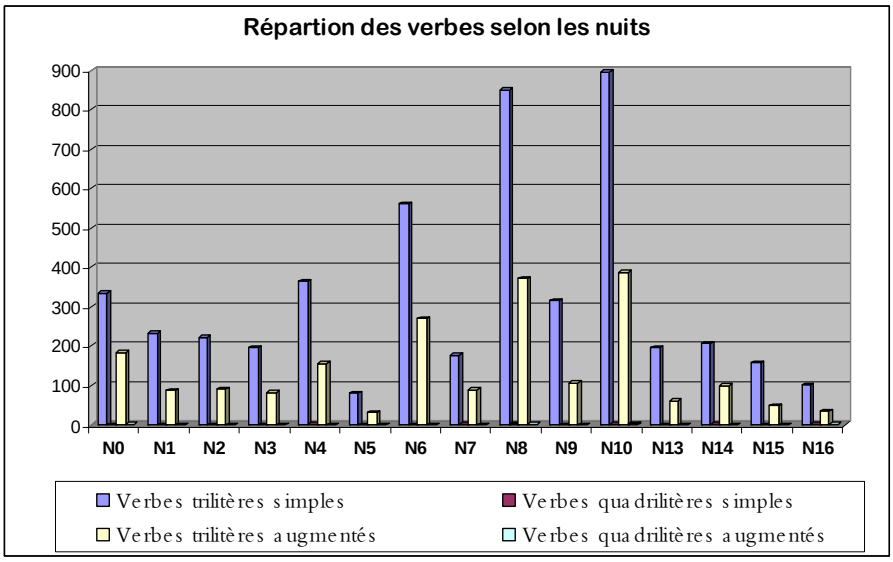


Figure 150

3.2. Les noms primitifs

La répartition des noms primitifs selon les Nuits place les *maÒdar* primitifs en tête du classement dans 12 Nuits sur 15. Ils sont devancés par les noms trilitères simples dans la Nuit 3, la Nuit 13 et la Nuit 10. Dans cette dernière Nuit, ils sont même relégués à la 3ème place derrière les noms trilitères simples et les noms augmentés. Les noms trilitères simples enregistrent un effectif anormalement élevé dans la Nuit 10.

Répartition des noms primitifs selon les nuits dans *al-ÞImtâÝ wa l-MuÞânasa*

	N0	N1	N2	N3	N4	N5	N6	N7	N8	N9	N10	N13	N14	N15	N16	
Nom composé	1	3	0	0	0	0	4	1	0	0	11	0	0	0	1	**21**
***MaÒdar* primitif**	409	164	266	110	271	53	475	187	622	428	355	154	210	135	97	**3 936**
Nom d'une fois	8	3	5	0	4	1	5	1	5	2	12	1	2	1	0	**50**
Nom de manière	4	1	0	1	1	0	4	0	1	1	0	0	1	0	2	**16**
Nom augmenté	179	75	122	76	144	36	293	121	413	210	568	98	146	76	55	**2 612**
Nom tri. simple	239	134	135	126	201	32	389	115	573	233	1113	216	156	88	58	**3 808**
N. quadri. simple	1	4	4	1	3	0	4	0	10	2	52	21	2	1	2	**107**
N. penta. simple	1	0	0	0	1	0	0	0	0	0	8	0	0	0	0	**10**
Noms primitifs	**842**	**384**	**532**	**314**	**625**	**122**	**1 174**	**425**	**1 624**	**876**	**2 119**	**490**	**517**	**301**	**215**	**10 560**

Tableau 98

Les Nuits les plus riches en noms primitifs sont dans l'ordre, les Nuits 10, 8 et 6 avec respectivement, 2 119, 1 624, et 1 174 occurrences.

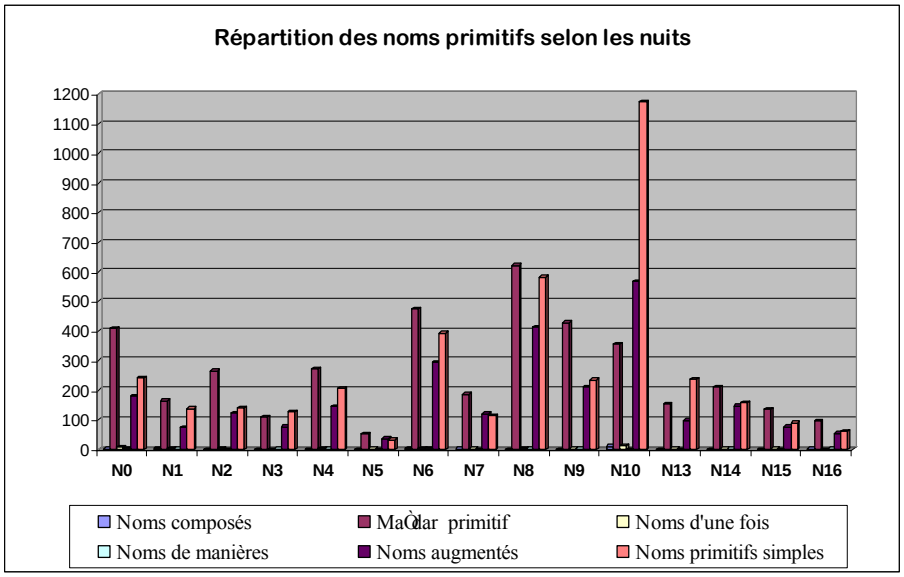


Figure 151

Les Nuits les plus pauvres en noms primitifs sont les Nuits 5, 16 et 15 avec respectivement, 122, 215, et 301 occurrences.

3.3. Les adjectifs

Concernant des adjectifs, la 1ère place des *Òifat mušabbahat* relevée au niveau du corpus entier est contestée ici dans 5 Nuits sur 15. En effet, dans les Nuits 8, 13, 14, 15 et 16 ; ce sont les participes actifs, 2èmes au niveau du corpus, qui arrachent cette première place aux *Òifat mušabbahat*.

Répartition des adjectifs selon les nuits dans *al-ÞImtâÝ wa l-MuÞânasa*

	N0	N1	N2	N3	N4	N5	N6	N7	N8	N9	N10	N13	N14	N15	N16	
Participe actif	124	26	61	50	101	21	146	63	282	77	184	72	73	55	41	**1 376**
Participe passif	54	15	49	13	51	14	71	32	112	52	35	23	43	9	8	**581**
Intensif	9	6	6	5	12	0	6	5	12	6	32	0	0	1	1	**101**
Elatif	48	11	27	16	49	21	72	33	94	47	69	15	21	8	3	**534**
Ñifa mušabbaha	170	92	111	57	110	31	169	71	277	105	229	53	58	36	22	**1 591**
Adjectif de relation	10	2	11	5	9	2	41	9	55	26	27	11	42	8	3	**261**
Adjectifs	**415**	**152**	**265**	**146**	**332**	**89**	**505**	**213**	**832**	**313**	**576**	**174**	**237**	**117**	**78**	**4 444**

Tableau 99

Les Nuits les plus riches en adjectifs sont dans l'ordre, la Nuit 8, la Nuit 8 et la Nuit 6 avec respectivement, 832, 576, et 505 occurrences.

Les Nuits les plus pauvres en adjectifs sont dans l'ordre, les Nuits 16, 5 et 15 avec respectivement, 78, 89, et 117 occurrences.

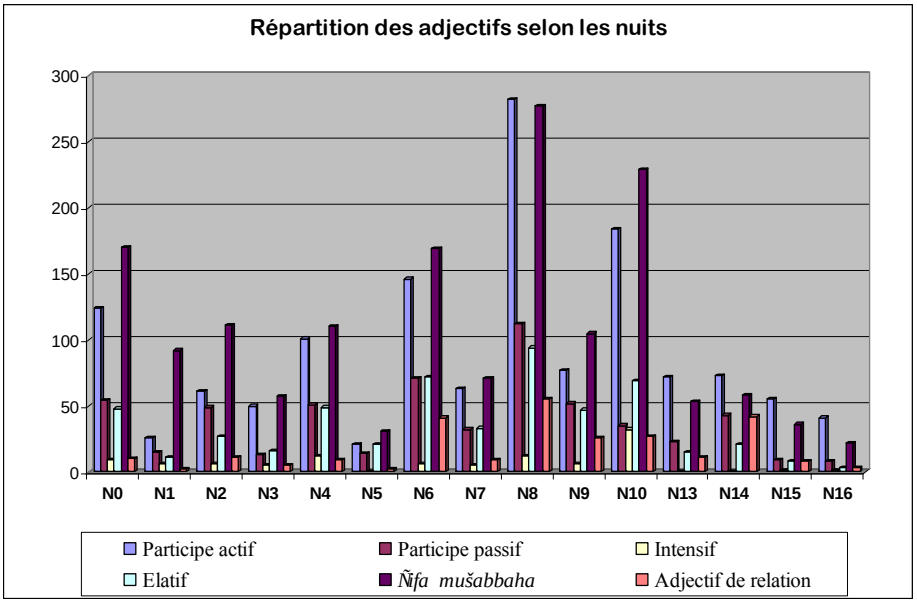


Figure 152

3.4. Les noms dérivés

Sans surprise, la prédominance des *maÒdar* dérivés et surtout, parmi eux, celle des *maÒdar* trilitères augmentés, constatée au niveau du corpus, est largement installée au niveau des Nuits.

La deuxième place est partagée entre les noms de temps et de lieu (Nuits 2, 5, 7, 9, 10, 15 et 16) et les *maÒdar mîmî* (Nuits 0, 1, 4, 6, 8, 13 et 14) ; les deux sous-catégories sont *ex aequo* dans la Nuit 3 avec 11 occurrences.

Répartition des noms dérivés selon les nuits dans *al-ÞImtâÝ wa l-MuÞânasa*

	N0	N1	N2	N3	N4	N5	N6	N7	N8	N9	N10	N13	N14	N15	N16	
Noms de temps / lieu	11	11	20	11	13	5	32	26	43	17	48	5	12	9	8	271
N. d'instrument	2	0	1	0	3	0	0	1	2	0	13	0	0	0	0	22
Diminutif	0	0	0	0	0	0	0	0	0	0	2	0	0	1	0	3
MaÒdar mîmî	15	12	11	11	19	3	40	11	150	13	8	11	21	8	3	336
MaÒdar ÒinâÝî	2	1	3	0	1	0	9	0	21	4	4	0	13	0	1	59
***MaÒdar* tri. augm.**	138	68	54	33	113	19	131	53	242	126	81	36	120	46	46	1306
***MaÒdar* quadri. simple.**	1	0	1	0	1	0	3	0	3	0	1	0	1	0	0	11
***MaÒdar* quadri. augm.**	0	0	1	0	0	1	1	1	0	3	0	0	0	0	0	7
Noms dérivés	**169**	**92**	**91**	**55**	**150**	**28**	**216**	**92**	**461**	**163**	**157**	**52**	**167**	**64**	**58**	**2 015**

Tableau 100

Les Nuits 8, 6 et 0 sont les Nuits les plus riches en noms dérivés avec respectivement, 461, 216 et 169 occurrences. Alors que celles les plus pauvres en noms dérivés sont les Nuits 5, 13, et 3 avec respectivement, 28, 52 et 55 occurrences.

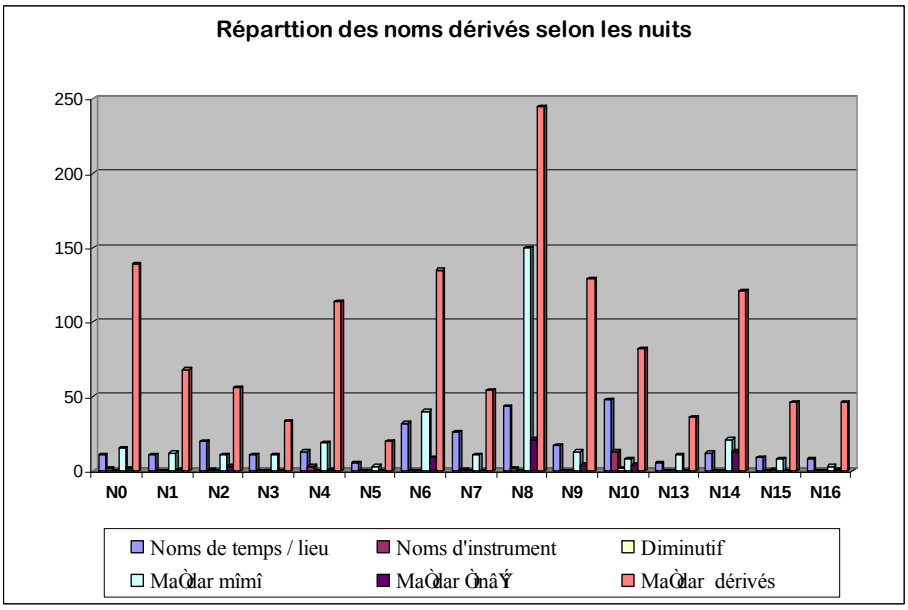


Figure 153

3.5. Les mots-outils

Répartition des mots-outils selon les nuits dans *al-ÞImtâÝ wa l-MuÞânasa* : Catégories																
	N0	**N1**	**N2**	**N3**	**N4**	**N5**	**N6**	**N7**	**N8**	**N9**	**N10**	**N13**	**N14**	**N15**	**N16**	
Particule	2 115	1 067	1 326	797	1 835	381	3 141	1 173	4 726	2 200	3 647	1 063	1 506	793	573	**26 343**
Nom-Outil	982	459	575	404	758	164	1 186	391	1 864	608	1 741	369	517	290	216	**10 524**
V. fonctionnalisé	13	2	10	4	18	3	17	5	31	17	15	21	10	2	4	**172**
Mot-Outil composé	7	3	3	7	9	8	11	3	26	8	22	2	9	0	4	**122**
Mots-Outils	**3 117**	**1 531**	**1 914**	**1 212**	**2 620**	**556**	**4 355**	**1 572**	**6 647**	**2 833**	**5 425**	**1 455**	**2 042**	**1 085**	**797**	**37 161**

Tableau 101

Au niveau des mots fonctionnels, la suprématie des particules, remarquée au niveau du corpus, est nettement confirmée dans chaque Nuit. La 2ème place est également gardée, au niveau de chaque Nuit, pour les noms-outils. Ce n'est qu'entre les verbes fonctionnalisés et les mots-outils composés qu'on observe quelques changements

de rangs au niveau de certaines Nuits. Dans les Nuits 1, 3, 5 et 10, les verbes fonctionnalisés cèdent leur 3ème place aux mots-outils composés.

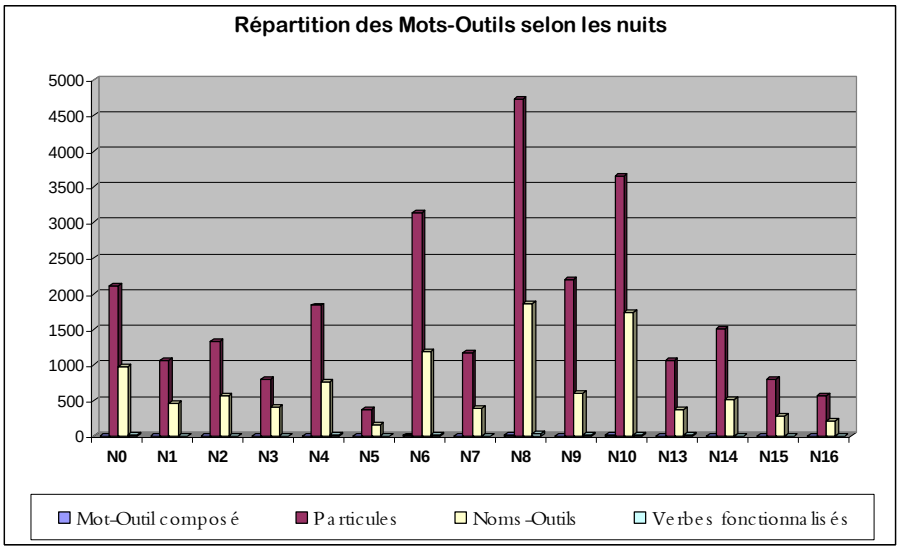


Figure 154

À l'intérieur des particules, il est intéressant de remarquer que la 1ère place que les coordonnants occupent au niveau du corpus, ne leur est pas due dans toutes les Nuits. En effet, les coordonnants ont pu garder cette 1ère place, du préambule jusqu'à la Nuit 8 ; la tendance est inversée à partir de la Nuit 9 pour voir ce 1er rang revenir à l'article défini. Les coordonnants se sont même vus relégués au 3ème rang, derrière l'article défini et les prépositions dans trois Nuits différentes, la Nuit 10, la Nuit 13 et la Nuit 15.

Répartition des mots-outils selon les nuits dans *al-ÞImtâÝ wa l-MuÞânasa* : Sous-catégories

	N0	N1	N2	N3	N4	N5	N6	N7	N8	N9	N10	N13	N14	N15	N16	
Mot-Outil composé	7	3	3	7	9	8	11	3	26	8	22	2	9	0	4	122
Article défini	566	286	388	190	498	86	861	372	1 317	777	1 229	343	477	248	187	7 825
Prépositions	605	299	361	221	544	111	852	281	1 350	516	1 034	284	408	234	157	7 257
Coordonnants	746	323	432	273	583	127	1111	373	1 477	691	991	277	462	196	159	8 221
Particules de négation	85	36	39	36	50	18	112	32	153	47	133	56	52	49	12	910
Partic. de comparaison	5	5	2	2	1	1	6	0	1	1	7	0	1	0	0	32
Partic. attente/certitude	14	21	19	7	30	4	23	8	56	20	16	8	9	7	4	246
Particules du futur	0	2	1	1	1	0	1	1	1	0	0	0	0	0	0	8
Particules de condition	14	11	19	19	25	10	27	24	69	46	77	13	20	7	9	390
Partic. de corroboration	37	46	36	23	58	12	81	40	157	55	91	43	35	30	29	773
P. espérance/apitoiement	4	1	0	0	0	0	0	1	3	1	0	0	0	0	0	10
Démonstratifs	71	64	55	41	60	14	155	55	236	81	122	42	78	38	30	1 142
Interrogatifs	2	2	7	7	12	2	8	10	27	0	1	0	9	5	8	100
Relatifs	93	54	68	43	54	13	121	42	215	72	113	31	73	34	24	1 050
Noms de verbes	2	1	1	0	1	0	2	0	2	0	0	0	0	0	0	9
Circonlocutions	3	4	1	3	1	0	1	1	8	3	2	0	4	1	0	32
Adverbes	86	41	55	33	59	11	114	35	173	70	243	38	58	36	12	1 064
Pronoms personnels	658	249	367	248	505	116	683	218	1 062	320	1 061	198	269	161	123	6 238
Annectifs	42	16	11	16	40	2	68	13	78	33	90	39	21	9	8	486
Verbe figé à l'accompli	13	1	9	3	16	2	15	5	28	15	15	20	10	2	3	157
Verbe figé à l'impératif	0	1	1	1	2	1	2	0	3	2	0	1	0	0	1	15
Particules de *"maÒdarité"*	12	10	11	12	21	4	30	19	67	22	35	17	16	11	1	288
Particules d'appel	5	8	2	2	6	0	3	2	14	0	0	0	2	1	1	46
Particules de réponse	2	1	3	0	0	2	2	1	4	0	1	6	1	0	0	23
Particules interrogatives	5	3	1	3	3	2	7	3	18	1	2	1	6	3	2	60
Particules d'exception	8	8	2	6	12	2	22	12	26	6	18	5	11	1	5	144
Particules d'ouverture	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
P. d'incitation/de remords	2	0	3	2	0	0	0	0	2	2	0	1	4	0	0	16
P. assimilées au verbe	2	1	7	0	3	0	2	1	6	9	12	5	0	3	2	53
Particules d'alternative	2	2	0	0	0	2	0	3	2	4	1	4	2	2	5	29
Nombres cardinaux	11	18	7	11	10	3	18	11	55	17	93	11	2	2	6	275
Nombres ordinaux	14	10	3	2	16	3	16	5	8	12	16	10	3	4	5	127
Nombres fractionnaires	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1
Particules d'explication	0	4	0	0	0	0	1	0	3	2	0	0	0	1	0	11
Mots-Outils	**3117**	**1531**	**1914**	**1212**	**2620**	**556**	**4355**	**1572**	**6647**	**2833**	**5425**	**1455**	**2042**	**1085**	**797**	**37 161**

Tableau 102

À l'intérieur des noms-outils, les pronoms personnels confirment bien leur prédominance dans toutes les Nuits. Dans la Nuit 10, leur effectif dépasse même celui de deux sous-catégories de la catégorie des particules qui est, rappelons-le, la catégorie la plus fréquente aussi bien au niveau du corpus qu'au niveau des Nuits. Les pronoms personnels enregistrent donc, dans la Nuit 10, 1 061 occurrences contre 991 pour les coordonnants et 1 034 pour les prépositions.

Entre démonstratifs et relatifs, le classement, au niveau du corpus où les premiers devancent les seconds, est quasiment le même exception faite des Nuits 0, 2 et 3 où l'effectif des relatifs est supérieur à celui des démonstratifs.

4. Catégories lexicales : effectifs théoriques

Après avoir présenté, globalement et analytiquement, les effectifs réellement observés dans le corpus, nous allons présenter ici les effectifs que nous avons calculés pour construire un modèle théorique de distribution des catégories lexicales. Une comparaison sera ensuite faite entre les effectifs observés et les effectifs théoriques. Le but de cette comparaison est de voir si les écarts entre effectifs observés et effectifs calculés sont significatifs ou aléatoires.

Comme les partitions lexicales relèvent d'un choix conscient chez un écrivain, les écarts, quand ils sont significatifs, peuvent nous renseigner sur les caractéristiques du style de celui-ci. Quand les écarts ne sont pas significatifs, c'est que l'écrivain ne s'éloigne pas davantage de la norme, des banalités de la langue.

Après le calcul des écarts, il faudra donc distinguer ceux qui sont significatifs de ceux qui ne le sont pas puisque, comme le souligne Philippe Thoiron, « parmi les écarts observés à partir d'un corpus, certains, dépendant du fonctionnement normal de la langue, sont aléatoires, donc non significatifs. D'autres, manifestations de l'exercice du choix chez le locuteur, sont susceptibles de relever du style »[320].

Pour évaluer les distances entre les effectifs concernant chaque catégorie lexicale et au niveau de chaque nuit, nous avons utilisé les écarts réduits qui sont calculés, rappelons-le, en divisant les écarts absolus par l'écart-type. Et pour avoir enfin, une vision globale des corrélations entre les effectifs observés et les effectifs théoriques, nous avons fait appel au test de corrélation de Pearson. Ce test de corrélation appliqué à chacune des catégories lexicales devra nous permettre d'évaluer le degré de similitude ou de dispersion entre les effectifs réels et les effectifs théoriques.

[320] Philippe Thoiron, *Dynamisme du texte et Stylostatistique*, 1980, p. 53.

Le modèle que nous allons construire part de l'hypothèse de la stabilité de la répartition des catégories lexicales. Partant du rapport de chaque catégorie lexicale au nombre total des occurrences du corpus, nous avons calculé les effectifs théoriques en considérant que ce rapport devrait rester le même dans toutes les Nuits du corpus.

Ce modèle théorique représente d'une certaine façon, des valeurs moyennes reflétant une norme par rapport à laquelle sont mesurés des écarts. Il n'est de ce fait pas censé refléter la réalité.

Calculons par exemple, les effectifs théoriques des noms dérivés dans la Nuit 2 : Étant donné qu'il y a dans tout le corpus 2 015 noms dérivés sur 61 177 occurrences, la probabilité d'apparition des noms dérivés est :

$$\rho = \frac{2015}{61177} = 0{,}03294$$

L'étendue de la Nuit 2 étant de N = 3 115, le modèle prévoit un effectif théorique des noms dérivés dans la Nuit 2 de :

$$3\ 115 \times 0{,}03294 = 102{,}60$$

Sachant que l'effectif réel est de 91 noms dérivés dans la Nuit 2, l'écart absolu est donc de :

$$91 - 102{,}60 = -\ 11{,}60$$

Pour avoir l'écart réduit, nous avons d'abord calculé l'écart-type de la distribution des noms dérivés : il est de 106,182. En divisant ensuite l'écart absolu par l'écart-type, nous obtenons l'écart réduit entre l'effectif réel et l'effectif théorique des noms dérivés dans la Nuit 2 :

$$\text{Écart réduit} = \frac{-11{,}60}{106{,}182} = -\ 0{,}1092$$

En appliquant ces calculs à toutes les catégories lexicales ainsi qu'à toutes les Nuits de notre corpus, nous obtenons l'ensemble des effectifs théoriques et des écarts, absolus et réduits, que nous présentons, en plus des effectifs réels, dans le tableau suivant :

Catégories lexicales : effectifs réels, effectifs théoriques et écarts entre les deux

	Verbes				Noms primitifs				Adjectifs				Noms dérivés				Mots-Outils			
	Réel	**Théor.**	**Ecart abs.**	**Ec. réduit**	**Réel**	**Théor.**	**Ecart abs.**	**Ec. réduit**	**Réel**	**Théor.**	**Ecart abs.**	**Ec. réduit**	**Réel**	**Théor.**	**Ecart abs.**	**Ec. réduit**	**Réel**	**Théor.**	**Ecart abs.**	**Ec. réduit**
N00	519	578,96	-59,96	-0,1637	842	873,77	-31,77	-0,0573	415	367,71	47,29	0,2253	169	166,73	2,27	0,0214	3117	3074,83	42,17	0,0239
N01	319	283,42	35,58	0,0971	384	427,74	-43,74	-0,0789	152	180,01	-28,01	-0,1334	92	81,62	10,38	0,0978	1531	1505,22	25,78	0,0146
N02	313	356,27	-43,27	-0,1181	532	537,69	-5,69	-0,0103	265	226,28	38,72	0,1844	91	102,60	-11,60	-0,1092	1914	1892,16	21,84	0,0124
N03	277	229,20	47,80	0,1305	314	345,92	-31,92	-0,0575	146	145,57	0,43	0,0020	55	66,01	-11,01	-0,1037	1212	1217,30	-5,30	-0,0030
N04	521	485,86	35,14	0,0959	625	733,26	-108,26	-0,1952	332	308,58	23,42	0,1116	150	139,92	10,08	0,0950	2620	2580,38	39,62	0,0225
N05	111	103,62	7,38	0,0201	122	156,39	-34,39	-0,0620	89	65,81	23,19	0,1104	28	29,84	-1,84	-0,0173	556	550,34	5,66	0,0032
N06	829	809,65	19,35	0,0528	1174	1221,93	-47,93	-0,0864	505	514,23	-9,23	-0,0440	216	233,16	-17,16	-0,1616	4355	4300,03	54,97	0,0312
N07	267	293,82	-26,82	-0,0732	425	443,45	-18,45	-0,0333	213	186,62	26,38	0,1257	92	84,62	7,38	0,0695	1572	1560,50	11,50	0,0065
N08	1224	1233,86	-9,86	-0,0269	1624	1862,16	-238,16	-0,4294	832	783,66	48,34	0,2303	461	355,33	105,67	0,9952	6647	6553,00	94,00	0,0533
N09	422	526,92	-104,92	-0,2864	876	795,23	80,77	0,1456	313	334,66	-21,66	-0,1032	163	151,74	11,26	0,1060	2833	2798,45	34,55	0,0196
N10	1287	1093,86	193,14	0,5272	2119	1650,88	468,12	0,8440	576	694,75	-118,75	-0,5656	157	315,01	-158,01	-1,4881	5425	5809,50	-384,50	-0,2182
N13	256	277,58	-21,58	-0,0589	490	418,93	71,07	0,1281	174	176,30	-2,30	-0,0110	52	79,94	-27,94	-0,2631	1455	1474,24	-19,24	-0,0109
N14	308	374,11	-66,11	-0,1805	517	564,62	-47,62	-0,0859	237	237,61	-0,61	-0,0029	167	107,74	59,26	0,5581	2042	1986,92	55,08	0,0313
N15	206	202,78	3,22	0,0088	301	306,04	-5,04	-0,0091	117	128,79	-11,79	-0,0562	64	58,40	5,60	0,0528	1085	1076,98	8,02	0,0046
N16	138	147,08	-9,08	-0,0248	215	221,98	-6,98	-0,0126	78	93,42	-15,42	-0,0734	58	42,36	15,64	0,1473	797	781,16	15,84	0,0090
	6997				**10560**				**4444**				**2015**				**37161**			

Tableau 103

4.1. Les verbes

Il est clair de la figure 155 que les deux courbes, celle des effectifs réels et celle des effectifs théoriques, ne sont confondues en aucun point. L'écart absolu entre les effectifs réels et les effectifs théoriques varie entre -104,92 (Nuit 9) et 193,14 (Nuit 10).

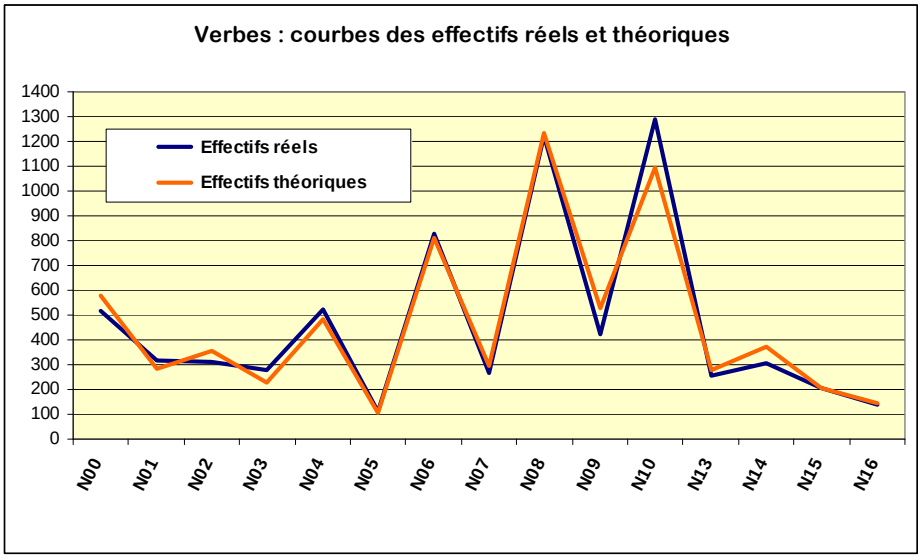


Figure 155

Les verbes enregistrent une sous-utilisation dans 8 Nuits sur 15, les Nuits 0, 2, 7, 8, 9, 13, 14 et 16. L'apogée de la sous-utilisation des verbes est enregistrée dans la Nuit 9 avec un écart réduit de -0,2864. Elle est suivie des Nuits 14 et 0 pour lesquelles l'écart est, respectivement, de -0,1805 et -0,1637.

La sur-utilisation des verbes, en revanche, est enregistrée dans 7 Nuits sur 15, les Nuits 1, 3, 4, 5, 6, 10 et 15. C'est dans la Nuit 10 où la sur-utilisation est la plus importante avec un écart réduit de 0,5272. Après la Nuit 10, les Nuits 3, 4 et 1 présentent une sur-utilisation des verbes relativement élevée et enregistrent des écarts réduits respectivement de 0,1305 , 0,0959 et 0,0971.

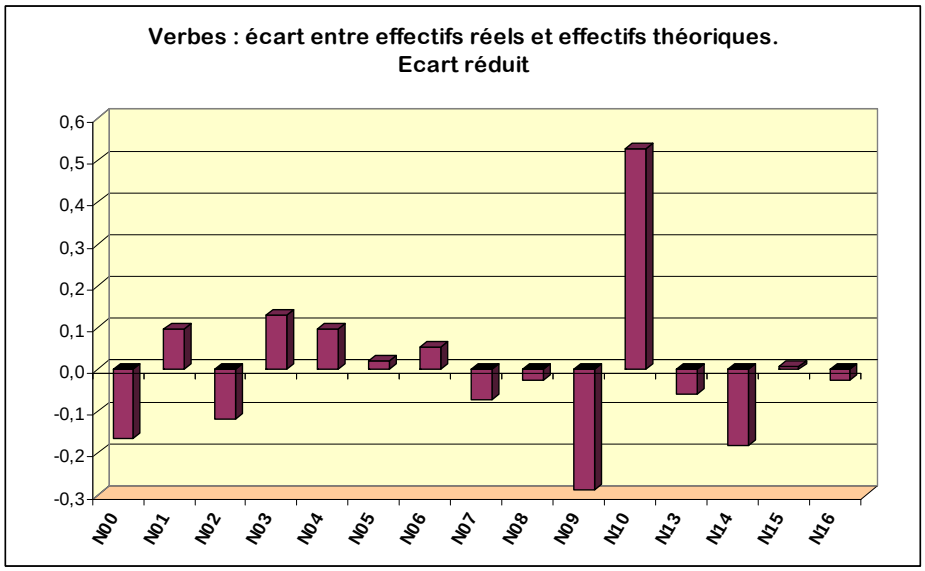


Figure 156

4.2. Les noms primitifs

Le décalage entre les deux courbes est encore plus net ici (figure 157) que pour les verbes. L'écart absolu varie entre -238,16 (Nuit 8) et 468,12 (Nuit 10).

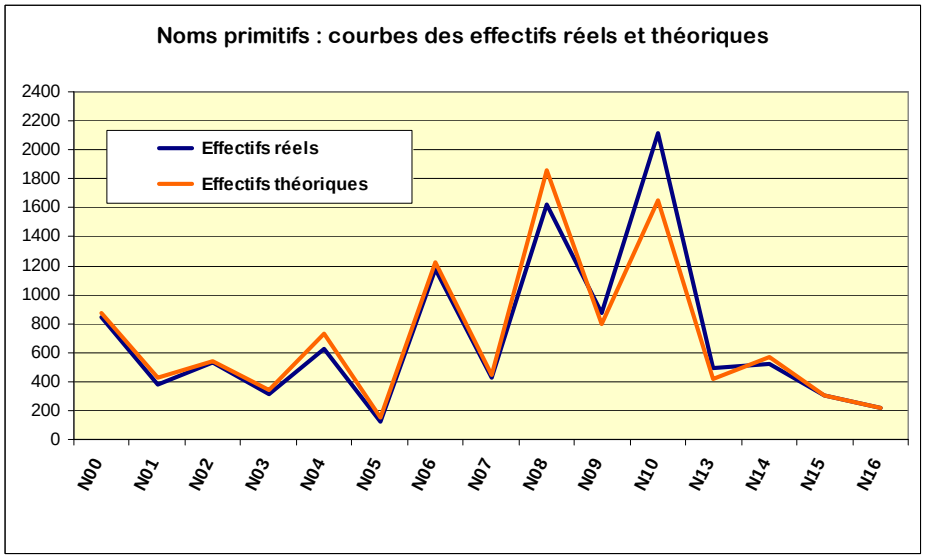


Figure 157

La sous-utilisation des noms primitifs par rapport aux valeurs théoriques est quasi générale ; sur les 15 Nuits, 12 accusent une sous-utilisation des noms primitifs. La Nuit 8 enregistre la plus haute sous-utilisation des noms primitifs avec un écart réduit de -0,4294, suivie de la Nuit 4 qui présente un écart réduit de -0,1952.

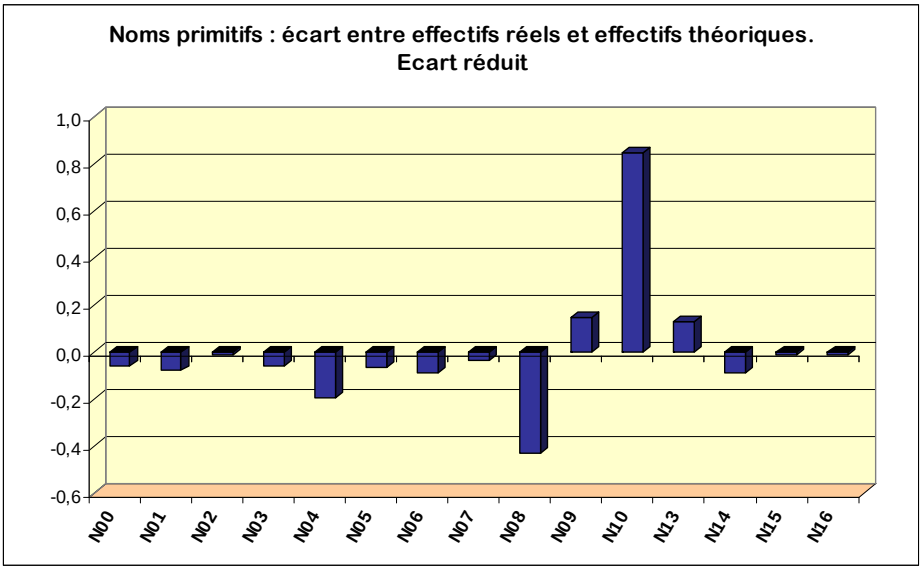


Figure 158

Les noms primitifs ne sont sur-utilisés que dans trois Nuits, la Nuit 9, la Nuit 10 et la Nuit 13. Comme pour les verbes, la Nuit 10 est marquée par la plus haute sur-utilisation des noms primitifs avec un écart réduit de 0,8440. Alors que la Nuit 9 a un écart de 0,1456 et la Nuit 13 de 0,1281.

4.3. Les adjectifs

Un décalage est aussi apparent sur la figure 159 entre la courbe des effectifs réels et celle des effectifs théoriques des adjectifs. L'écart absolu entre les deux effectifs varie de -118,75 (Nuit 10) à 48,34 (Nuit 8).

Les adjectifs sont sur-utilisés dans 7 Nuits parmi lesquelles, les Nuits 8, 0 et 2 présentent des écarts réduits rapprochés, ils ont respectivement la valeur 0,2303 , 0,2253

et 0,1844. Deux autres Nuits enregistrent également une sur-utilisation similaire des adjectifs avec un écart réduit de 0,1104 (Nuit 5) et 0,1116 (Nuit 4).

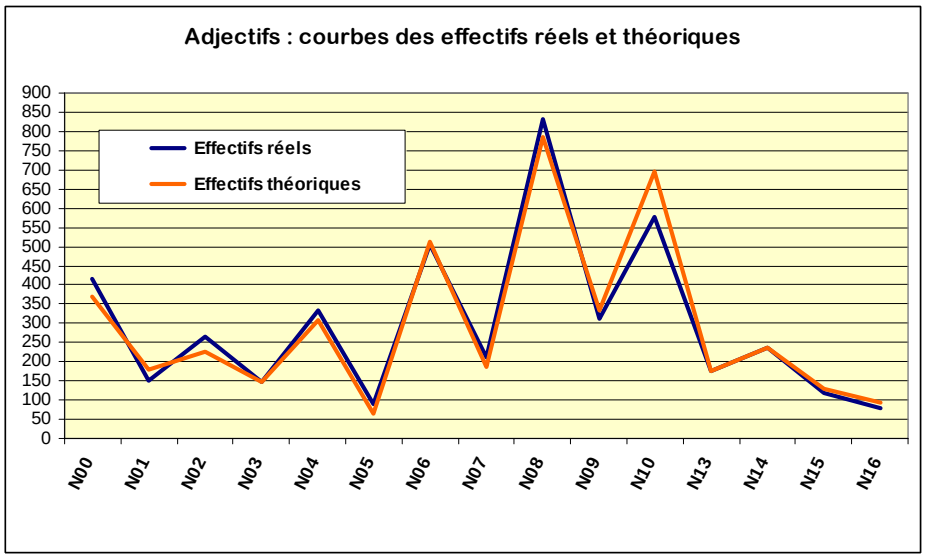


Figure 159

La forte sous-utilisation des adjectifs observée dans la Nuit 10 vient contrecarrer la sur-utilisation des verbes et des noms primitifs enregistrée dans cette même Nuit. À la suite de la Nuit 10, les deux Nuits qui enregistrent une sous-utilisation des adjectifs relativement importante sont la Nuit 1 avec un écart réduit de -0,1334 et la Nuit 9 avec -0,1032.

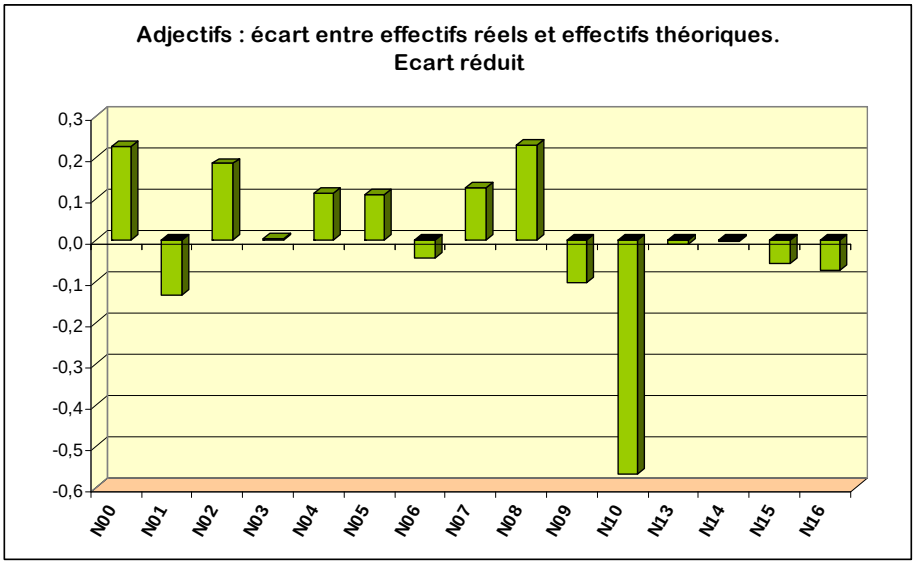


Figure 160

4.4. Les noms dérivés

Outre le décalage entre les deux courbes, ce qui caractérise le plus la figure 161, c'est le chevauchement nettement visible entre la courbe des effectifs réels et celle des effectifs théoriques des noms dérivés ; un chevauchement qui traduit une importante amplitude des variations des deux courbes.

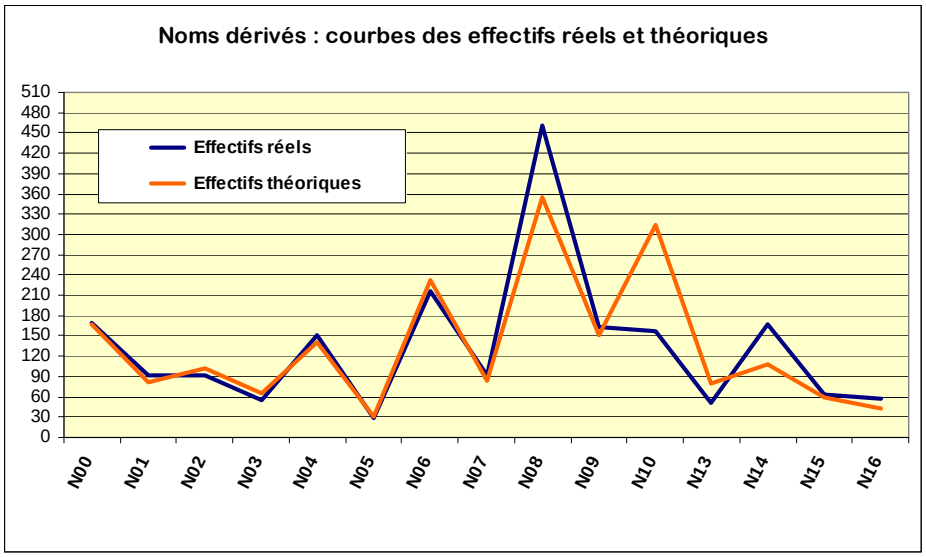


Figure 161

L'écart absolu entre les effectifs réels et les effectifs théoriques varie entre -158,01 (Nuit 10) et 105 (Nuit 8).

Neuf Nuits enregistrent une sur-utilisation des noms dérivés contre six accusant une sous-utilisation. Du côté de la sur-utilisation, la Nuit 8 est marquée par une valeur très élevée de l'écart réduit qui est de 0,9952. L'écart réduit enregistré pour la Nuit 14 en est de 0,5581.

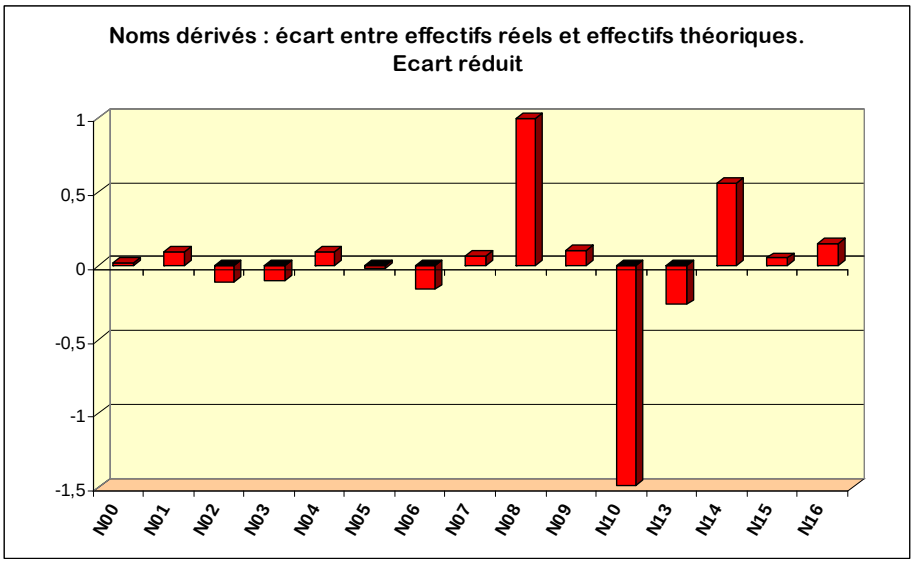


Figure 162

Du côté opposé, cinq des six Nuits où les noms dérivés sont sous-utilisés, présentent un écart réduit relativement faible ; alors que la Nuit 10 accuse une sous-utilisation anormalement forte des noms dérivés, ce qui se traduit par un écart réduit extrêmement élevé en valeur absolue, il est de -1,4881.

4.5. Les mots-outils

Au niveau des mots-outils, le décalage entre des deux courbes n'est pas notable. Il n'est visible qu'au point correspondant à la Nuit 10 et à peine visible aux points correspondant à la Nuit 8 et à la Nuit 14.

Au niveau de ces points, l'écart absolu est de -384,50 (Nuit 10), c'est sa valeur minimale, de 94 (Nuit 8), c'est sa valeur maximale, et de -19,24 (Nuit 13).

Seulement trois Nuits sont le lieu d'une sous-utilisation des mots-outils. Ces derniers sont particulièrement sous-utilisés dans la Nuit 10 avec un écart réduit de -0,2182. La Nuit 13 a un faible écart réduit de -0,0109 et la Nuit 3 un écart réduit extrêmement faible de -0,0030.

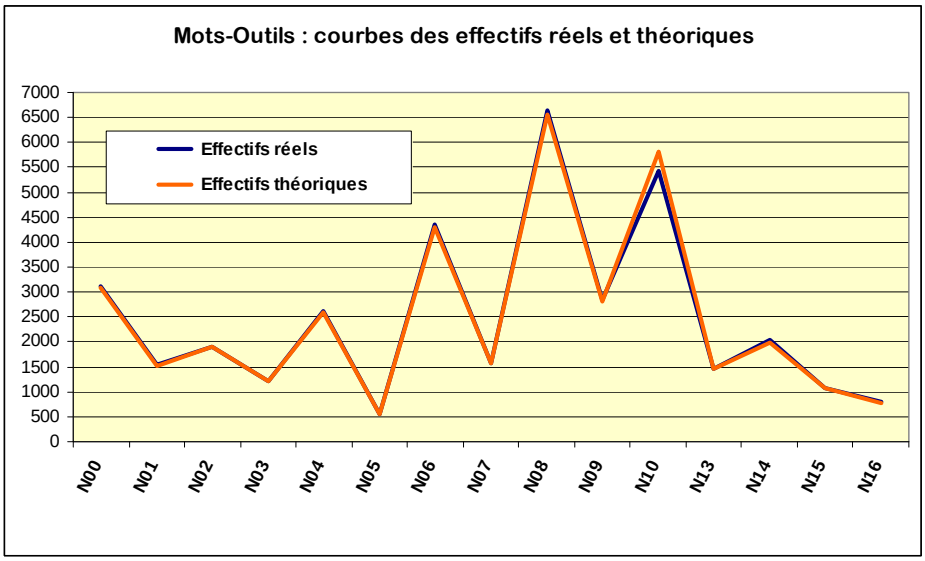


Figure 163

À l'opposé de cela, les mots-outils sont sur-utilisés dans 12 Nuits parmi lesquelles la Nuit 8 marquée par la sur utilisation la plus élevée avec un écart réduit de 0,0533. Après la Nuit 8, ce sont les Nuits 14 et 6 qui ont les plus grands écarts réduits : 0,0313 pour la première et 0,0312 pour la deuxième.

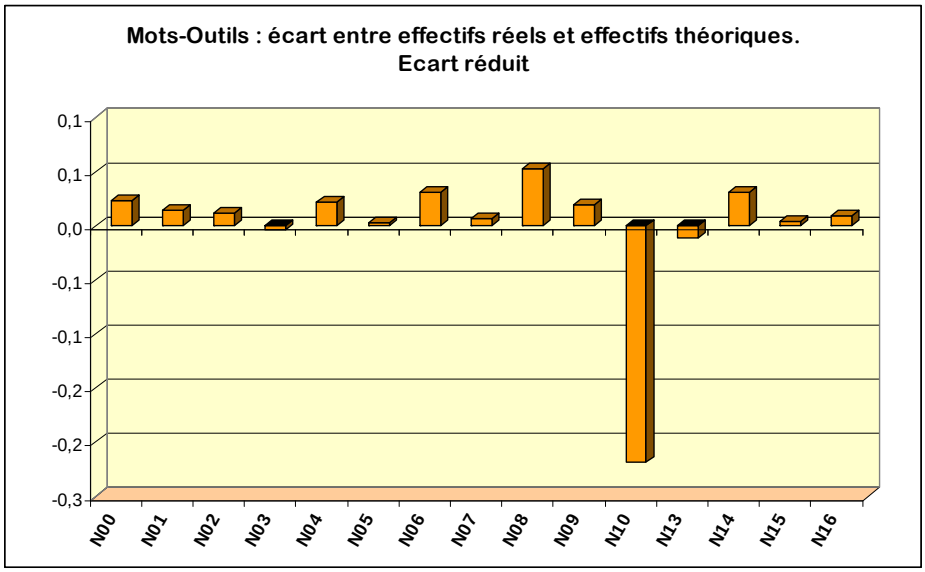


Figure 164

5. Nuits déficitaires et Nuits excédentaires

Nous avons pu évaluer les écarts entre les effectifs réels de chaque catégorie lexicale et ceux calculés selon le modèle théorique construit à partir de la distribution des catégories lexicales et leur rapport à l'étendue du corpus. Cette analyse dont le but était de distinguer les écarts significatifs de ceux qui ne l'étaient pas, a été menée en considérant une à une les catégories lexicales dans leur distribution selon les Nuits.

Une autre perspective sera considérée ici pour étudier la répartition des catégories lexicales et déceler les écarts significatifs : chaque Nuit sera examinée en fonction de l'excédent ou du déficit qu'elle peut avoir en telle ou telle catégorie lexicale. Autrement dit, les Nuits seront évaluées pour distinguer celles qui s'écartent significativement des effectifs théoriques positivement (excédent) ou négativement (déficit) et ce pour chacune des catégories lexicales.

Le Préambule (Nuit 0) se distingue particulièrement par l'excédent important en adjectifs et le déficit non moins élevé en verbes. Il est légèrement excédentaire en noms primitifs. La répartition des noms dérivés et des mots-outils n'y est pas significative.

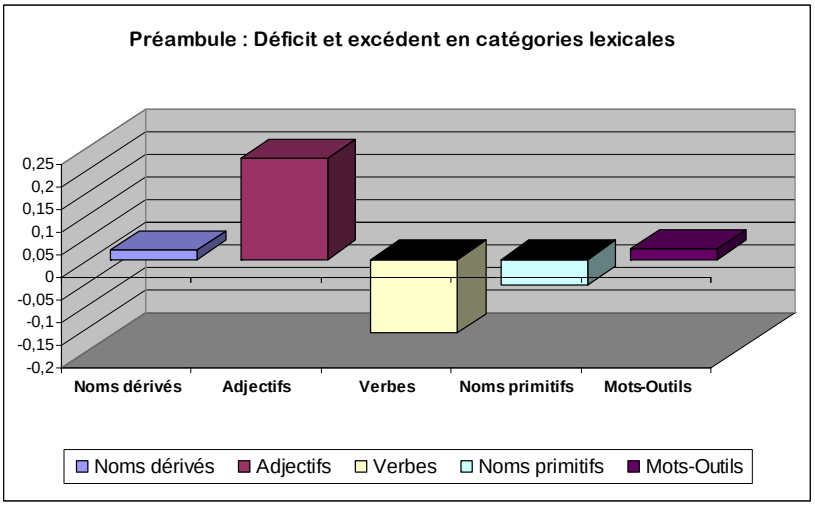


Figure 165

La Nuit 1 est caractérisée par quatre écarts significatifs. Au niveau des adjectifs et des verbes, elle se distingue, inversement à la Nuit 0, par le fait qu'elle soit significativement déficitaire en adjectifs et excédentaire en verbes. Elle se distingue également par un excédent en noms dérivés et un déficit moins important mais encore significatif en noms primitifs. Tout comme pour la Nuit 0, l'écart des mots-outils n'y est pas significatif.

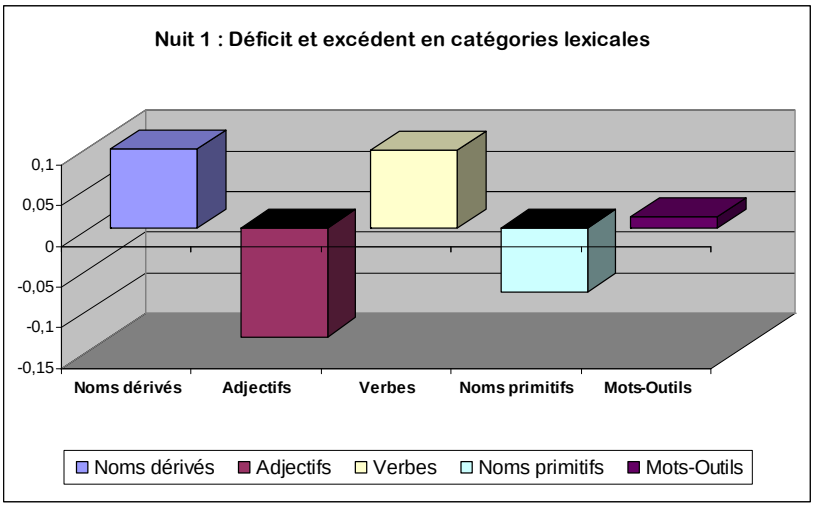


Figure 166

La Nuit 2 est particulièrement excédentaire en adjectifs. Elle est déficitaire en verbes et en noms dérivés. Les écarts au niveau des noms primitifs et des mots-outils ne sont pas significatifs.

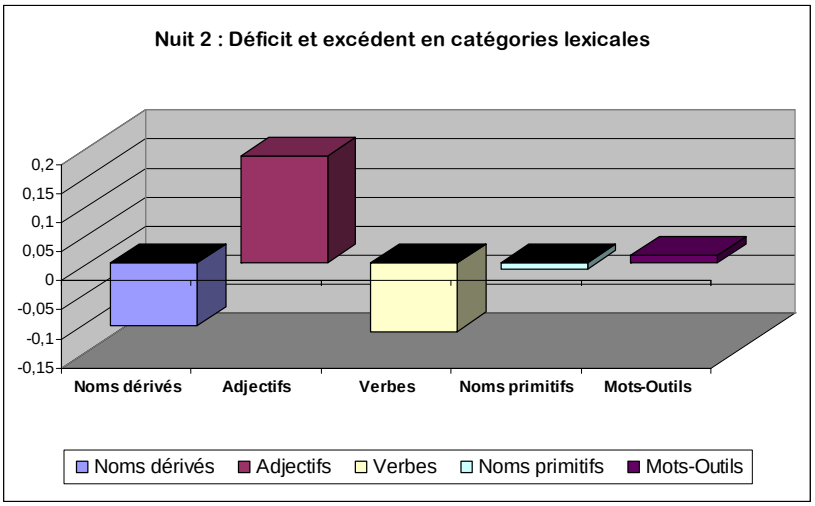


Figure 167

C'est principalement l'excédent en verbes qui caractérise la Nuit 3. Moins important mais aussi significatif que l'excédent en verbes, le déficit en noms dérivés en est une autre caractéristique. Cette Nuit est légèrement déficitaire en noms primitifs. Les écarts des adjectifs et des mots-outils ne sont pas significatifs.

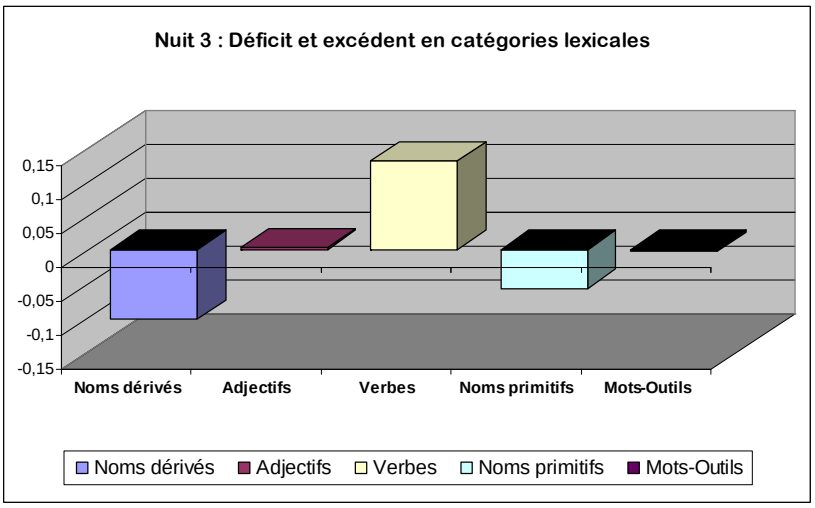


Figure 168

À l'image de la Nuit 1, la Nuit 4 est marquée par le nombre d'écarts significatifs. En effet, elle est excédentaire en adjectifs, en noms dérivés et en verbes. Elle est spécifiquement déficitaire en noms primitifs. Quant à l'écart des mots-outils, même si nous ne le retenons pas comme significatif, il n'est tout de même pas négligeable.

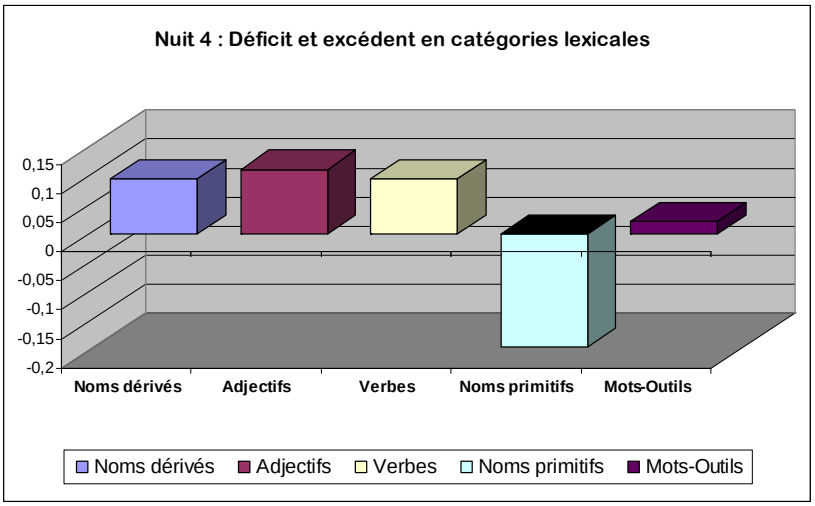


Figure 169

L'excédent en adjectifs et le déficit en noms primitifs sont ce qui caractérise le plus la Nuit 5. Celle-ci est aussi faiblement déficitaire en noms dérivés et faiblement excédentaire en verbes. L'écart des mots-outils n'est pas significatif.

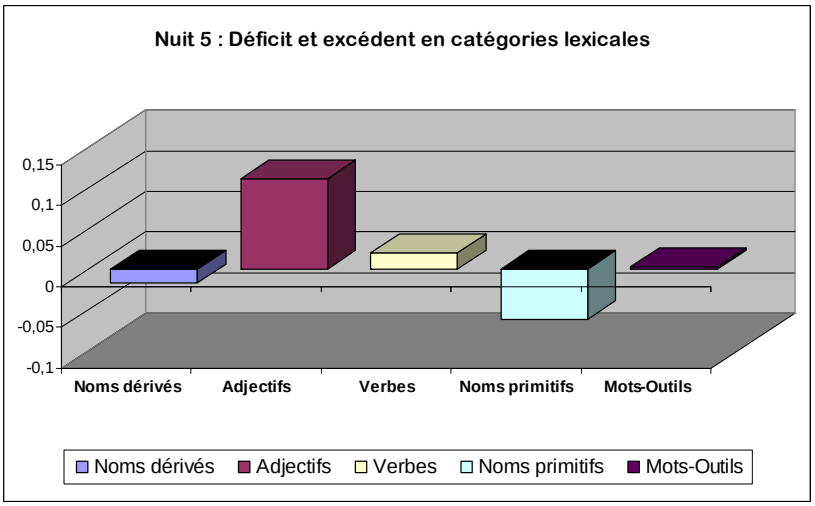


Figure 170

Les écarts significatifs battent leur plein dans la Nuit 6. Cette Nuit est, à des degrés différents, déficitaire en noms dérivés, en noms primitifs et en adjectifs. Elle est en même temps excédentaire en verbes et, à un degré moindre, en mots-outils.

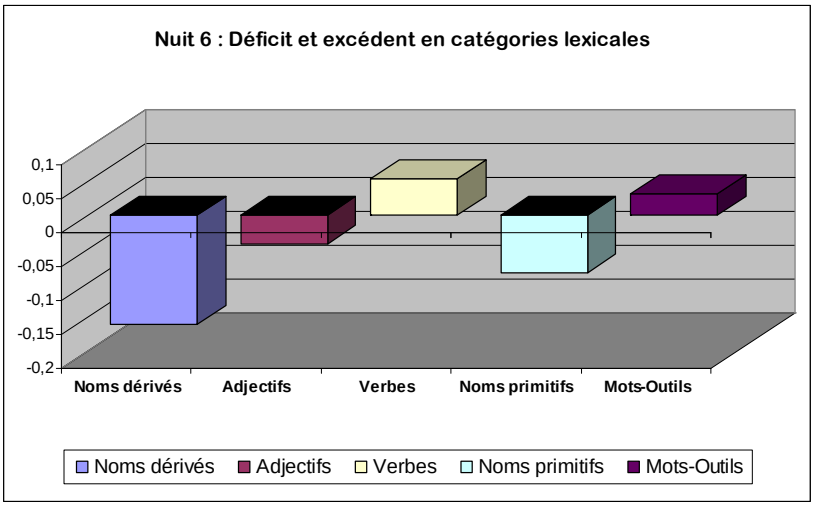


Figure 171

Quatre des cinq écarts sont significatifs dans la Nuit 7. Elle est particulièrement excédentaire en adjectifs et en en noms dérivés. Elle est en revanche déficitaire en verbes et modérément déficitaire en noms primitifs. L'écart des mots-outils n'est pas significatif.

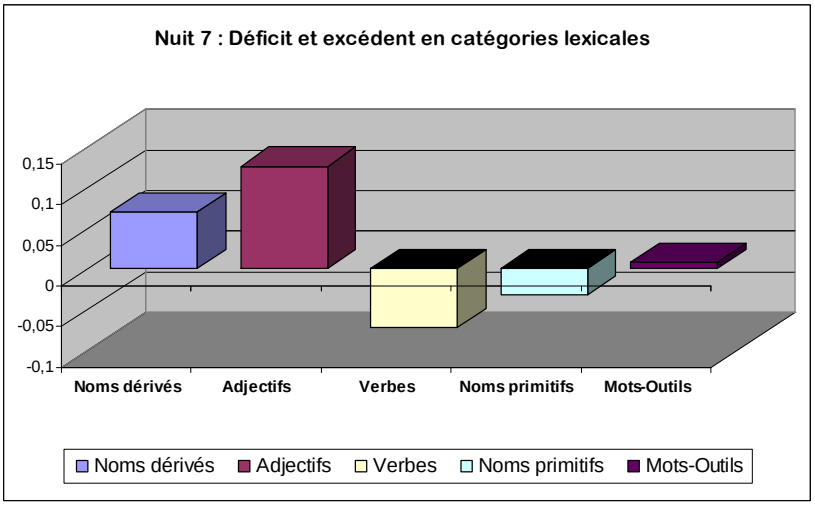


Figure 172

La Nuit 8 est extrêmement excédentaire en noms dérivés et moyennement excédentaire en adjectifs. Elle est en revanche déficitaire en noms primitifs. Les écarts des verbes et des mots-outils ne sont pas significatifs.

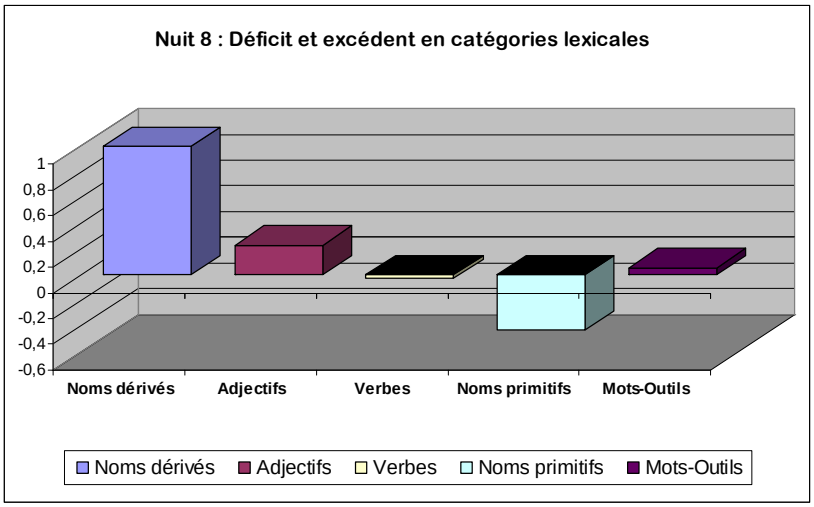


Figure 173

Deux excédents et deux déficits caractérisent la Nuit 9. Cette Nuit est déficitaire en verbes et en adjectifs. Elle est excédentaire en noms primitifs et en noms dérivés. L'écart des mots-outils n'est pas significatif.

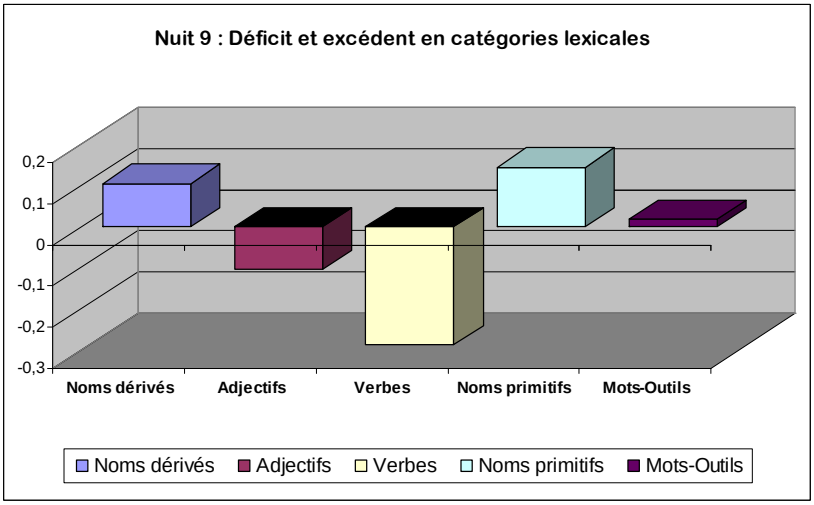


Figure 174

Comme la Nuit 6 mais à des degrés plus importants, la Nuit 10 est marquée par le fait que tous les écarts soient significatifs. Elle est extrêmement déficitaire en noms dérivés, déficitaire en adjectifs et modérément déficitaire en mots-outils. Elle est par ailleurs, excédentaire en noms primitifs et en verbes.

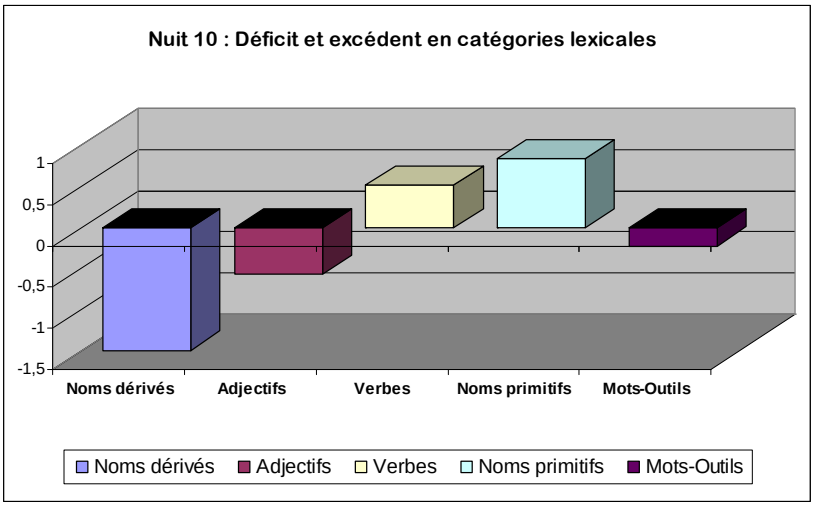


Figure 175

La Nuit 13 est déficitaire en noms dérivés et légèrement déficitaire en verbes. Elle est excédentaire en noms primitifs. Les écarts des adjectifs et des mots-outils ne sont pas significatifs.

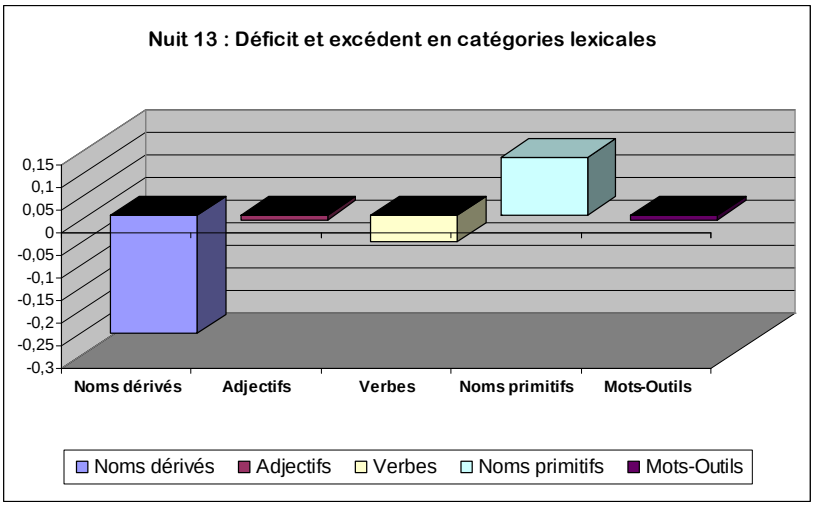


Figure 176

La Nuit 14 est pleinement excédentaire en noms dérivés, déficitaire en noms primitifs. Les écarts des adjectifs et des mots-outils ne sont pas significatifs.

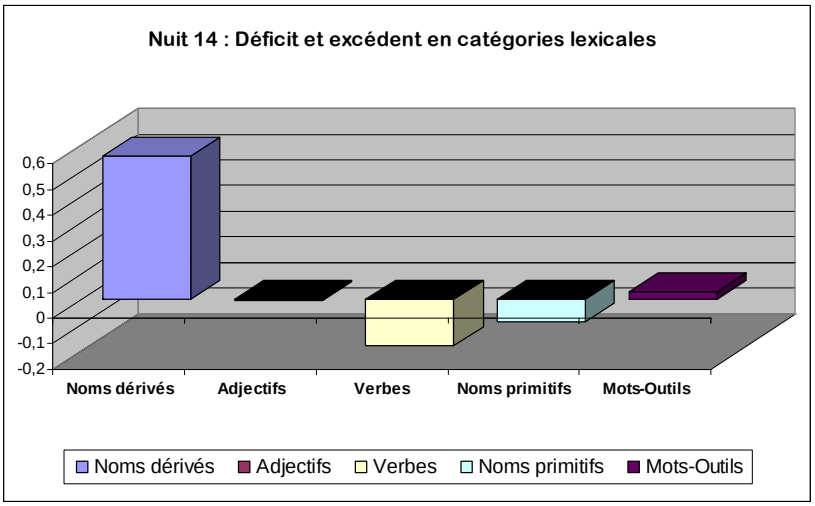


Figure 177

La Nuit 15 est excédentaire en noms dérivés et déficitaire en adjectifs. Les écarts des verbes et des noms primitifs sont très faibles pour être considérés comme significatifs, ils rejoignent ainsi les mots-outils.

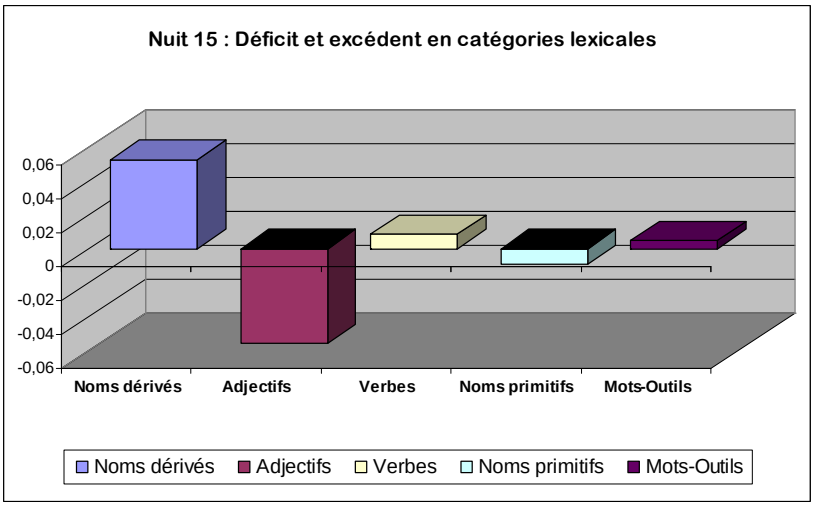


Figure 178

Enfin, la Nuit 16 est excédentaire en noms dérivés, déficitaire adjectifs et légèrement déficitaire en verbes. Les écarts des noms primitifs et des mots-outils ne sont pas significatifs.

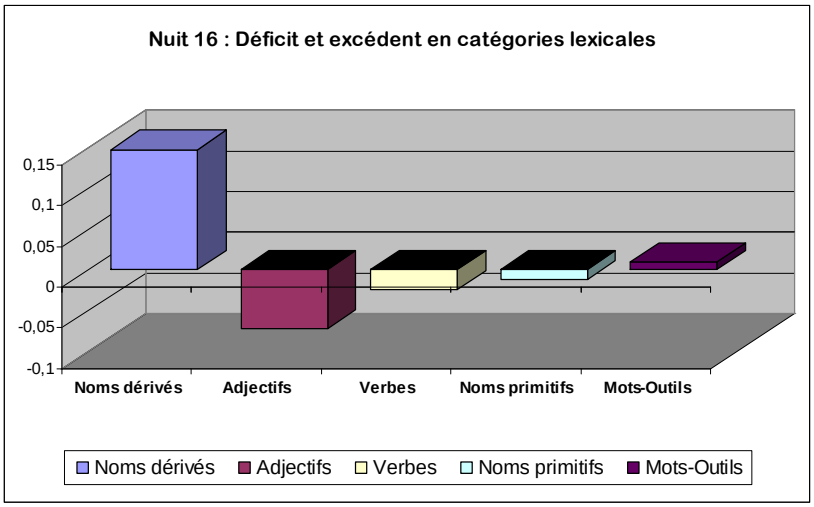


Figure 179

Quelques observations générales se dégagent de cette série d'évaluations des écarts significatifs ayant mis en exergue les excédents et les déficits en telle ou telle catégorie lexicale :

↳ La répartition des mots-outils est quasiment banale tout le long du corpus. L'écart des mots-outils reste non significatif dans pratiquement toutes les Nuits du corpus, exception faite de la Nuit 6 où il est légèrement excédentaire et de la Nuit 10 où il est légèrement déficitaire. Ce qui met en avant le profil moyen de cette catégorie lexicale.

↳ Huit Nuits enregistrent un petit nombre (≤ 3) d'écarts significatifs ; ce sont les Nuits 0, 2, 3, 8, 13, 14, 15 et 16. Elles se rapprochent plus ou moins d'une répartition moyenne des catégories lexicales.

↳ Sept Nuits enregistrent un grand nombre (4 ou 5) d'écarts significatifs ; ce sont les Nuits 1, 4, 5, 6, 7, 9 et 10. Elles se distinguent par leur originalité quant à la répartition des catégories lexicales.

↳ Mais cette opposition entre les deux groupes sur la base du nombre d'écarts significatifs ne doit pas masquer d'autres oppositions possibles, binaires ou de

groupes, pouvant être faites sur d'autres critères. L'opposition, par exemple, entre la Nuit 8 et la Nuit 10 est explicite et où l'on a une opposition au niveau des noms dérivés (Nuit 8:excédent, Nuit 10:déficit), au niveau des adjectifs (Nuit 8:excédent, Nuit 10:déficit) et au niveau des noms primitifs (Nuit 8:déficit, Nuit 10:excédent). L'opposition entre le Préambule et la Nuit 1 est aussi notable, etc.

↳ Le nombre de déficits et d'excédents constitue également un critère valable de regroupement et d'opposition.

↳ Pour chaque catégorie lexicale, l'on peut former des couples de Nuits ou regrouper celles-ci sur la base d'opposition ou de similitude vis-à-vis de la catégorie lexicale considérée. Dans l'analyse factorielle des correspondances que nous ferons plus loin, l'on verra mieux la disposition des Nuits par rapport aux différentes catégories lexicales.

6. Corrélations

Dans le but d'étudier le système d'opposition/similitude qui règle le jeu des catégories lexicales dans notre corpus, nous présentons dans cette section les résultats d'un certain nombre de tests de corrélation de Pearson que nous avons appliqués, d'un côté aux effectifs réels et aux effectifs théoriques des catégories lexicales, et de l'autre côté aux écarts réduits entre effectifs réels et effectifs théoriques comparant ainsi, d'abord les classes lexicales entre elles, et ensuite les Nuits entre elles.

6.1. Corrélation entre effectifs réels et effectifs théoriques

Pour chaque catégorie lexicale, le test de corrélation de Pearson est utilisé pour évaluer le degré de similitude ou de dispersion entre les effectifs réellement observés dans le corpus et les effectifs calculés selon le modèle théorique.

↳ **Les verbes :** Le test de corrélation de Pearson appliqué aux verbes nous révèle un coefficient de corrélation élevé de 0,985. Ce qui nous permet de conclure qu'au seuil de signification $\alpha = 0{,}050$, on peut rejeter l'hypothèse nulle d'absence de corrélation. Autrement dit, la corrélation est significative.

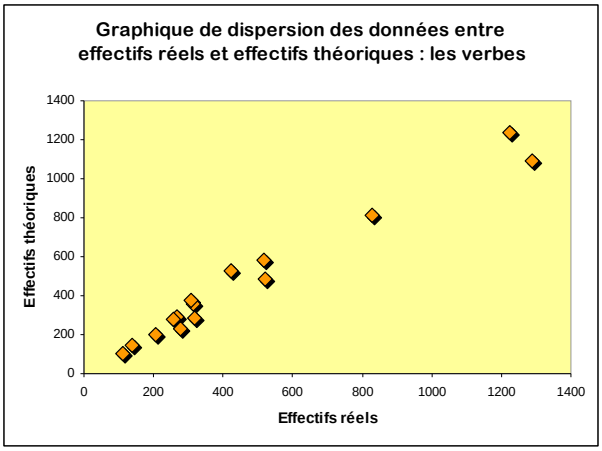


Figure 180

En outre, le graphique de dispersion de la figure 180 montre bien un nuage de points presque linéaire où les points sont quasiment alignés et orientés du coin bas gauche au coin haut droit ; c'est la disposition d'une forte corrélation.

↳ **Les noms primitifs :** Le test de corrélation nous livre ici un coefficient de 0,965 qui est moins élevé que celui de la corrélation des verbes, mais qui reste quand même important. Avec cette valeur du coefficient de corrélation et au seuil de signification $\alpha = 0{,}050$, on peut rejeter l'hypothèse nulle d'absence de corrélation. Autrement dit, la corrélation est significative. et le graphique de la figure 181 le confirme.

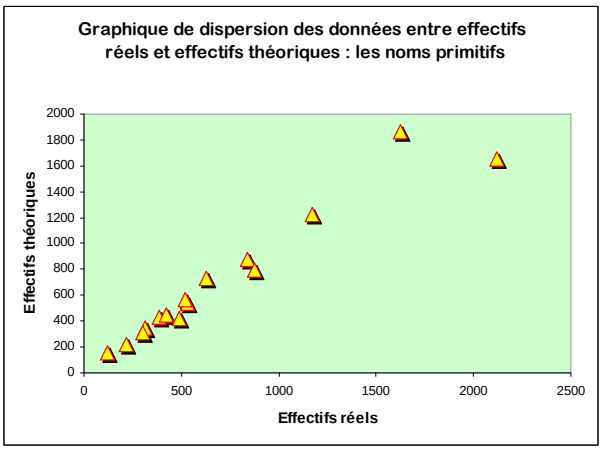


Figure 181

↳ **Les adjectifs** : Situé entre le coefficient de corrélation des verbes et celui, moins important, des noms primitifs, le coefficient de corrélation des adjectifs est de 0,981. Un coefficient dont la valeur traduit une corrélation significative et pour laquelle, au seuil de signification $\alpha = 0{,}050$, l'hypothèse nulle d'absence de corrélation peut être rejetée.

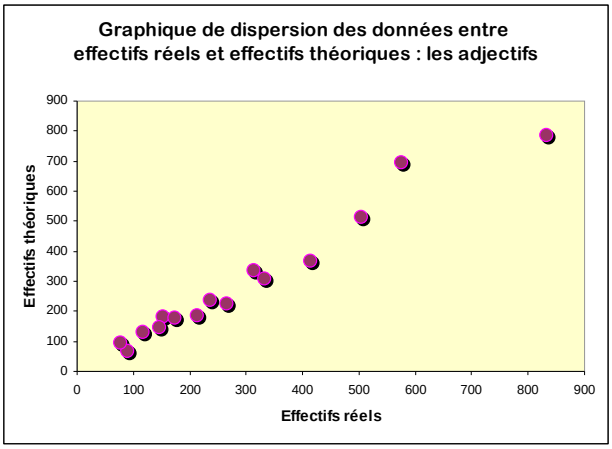


Figure 182

↳ **Les noms dérivés** : Avec un coefficient de corrélation de 0,860 , même si la corrélation des noms dérivés est déclarée significative, elle reste tout de même la moins significative de toutes les autres catégories lexicales, c'est-à-dire que les effectifs réels et les effectifs théoriques pour les noms dérivés, sont les moins corrélés entre eux que ceux de toutes les autres catégories lexicales. Ce qui explique et confirme l'importante amplitude des variations observée plus haut dans le graphique de la figure 161 représentant les deux courbes, celle des effectifs réels et celle des effectifs théoriques.

En dépit de cela, on peut rejeter l'hypothèse nulle d'absence de corrélation au seuil de signification $\alpha = 0{,}050$.

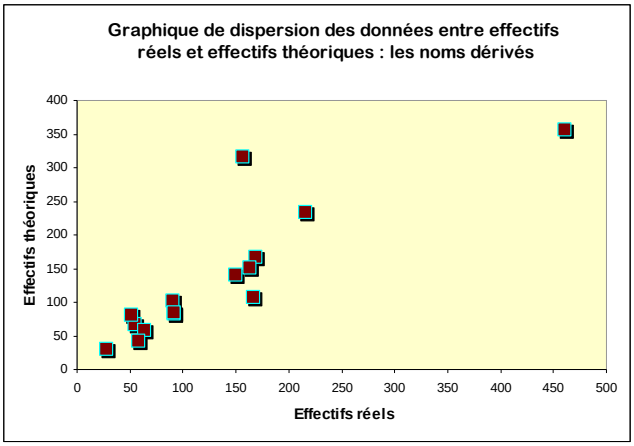


Figure 183

↳ **Les mots-outils :** Avec un coefficient de corrélation (0,998) frôlant la valeur maximale 1, les mots-outils enregistrent une corrélation presque parfaite entre effectifs réels et effectifs théoriques. Parmi celles de toutes les catégories lexicales, la corrélation des mots-outils est celle qui met en évidence les grandes similitudes entre les effectifs réels et les effectifs théoriques. Cela confirme le profil moyen des mots-outils constaté plus haut au niveau de l'excédent et du déficit des Nuits en mots-outils.

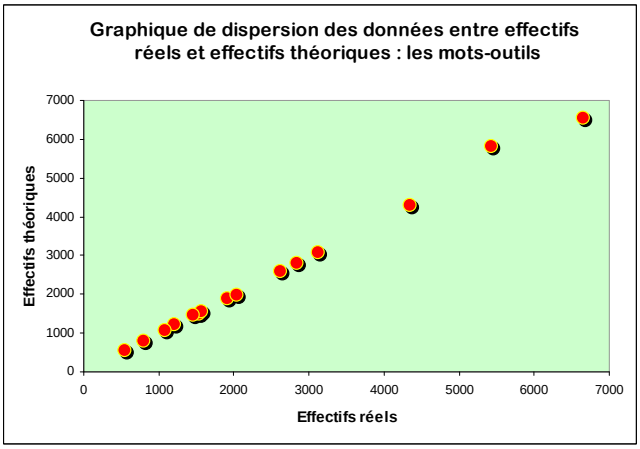


Figure 184

6.2. Corrélation des catégories lexicales sur la base des écarts réduits

Le test de corrélation de Pearson est appliqué ici aux écarts réduits entre effectifs réels et effectifs théoriques de chaque catégorie lexicale, pour évaluer le degré de corrélation ou d'opposition entre les catégories lexicales prises deux à deux.

Tous les coefficients de corrélation du tableau 17 suivant montrent que les corrélations sont toutes significatives, quelles soient directes ou inverses.

Matrice de corrélation de Pearson entre les catégories lexicales sur la base des écarts réduits

	Mots-Outils	Noms primitifs	Verbes	Adjectifs
Noms dérivés	0,892	- 0,913	- 0,696	0,741
Adjectifs	0,823	- 0,850	- 0,672	
Verbes	- 0,791	0,569		
Noms primitifs	- 0,933			

Tableau 104

Au niveau des corrélations directes, nous enregistrons une forte corrélation entre noms dérivés et mots-outils (0,892), et entre adjectifs et mots-outils (0,823). Une corrélation significative mais faible existe entre verbes et noms primitifs (0,569). Entre ces deux degrés de forte et de faible corrélation directe, une corrélation moyenne entre adjectifs et noms dérivés (0,741) est enregistrée.

Quant aux corrélations inverses, une très forte opposition est enregistrée entre noms primitifs et mots-outils (- 0,933) et entre noms primitifs et noms dérivés (- 0,913). Les noms primitifs et les adjectifs sont également opposés (- 0,850). Un coefficient de - 0,791 oppose aussi les verbes aux mots-outils. Enfin, deux oppositions sont également enregistrées mais à un degré moindre, la première oppose les verbes aux noms dérivés (- 0,696), l'autre les verbes aux adjectifs(- 0,672).

6.3. Corrélation des Nuits sur la base des écarts réduits

Ici également, le test de corrélation de Pearson est appliqué aux écarts réduits entre effectifs réels et effectifs théoriques, mais le but est d'évaluer le degré de corrélation ou d'opposition entre les Nuits prises deux à deux.

Matrice de corrélation de Pearson entre les Nuits sur la base des écarts réduits

	N16	N15	N14	N13	N10	N09	N08	N07	N06	N05	N04	N03	N02	N01
N00	-0,187	-0,553	0,264	-0,049	-0,551	0,192	0,319	**0,945**	-0,285	0,670	0,299	-0,391	0,858	-0,657
N01	0,665	**0,879**	0,374	-0,686	-0,232	-0,205	0,455	-0,433	0,065	-0,360	0,366	0,185	**-0,900**	
N02	-0,643	**-0,889**	-0,258	0,440	-0,050	0,012	-0,200	0,653	0,067	0,675	0,035	-0,080		
N03	-0,606	-0,285	-0,736	0,218	0,518	**-0,938**	-0,415	-0,501	**0,881**	0,383	0,330			
N04	0,115	0,072	0,301	-0,709	-0,633	-0,599	0,674	0,435	0,173	0,716				
N05	-0,514	-0,641	-0,160	-0,096	-0,295	-0,599	0,191	0,605	0,303					
N06	-0,641	-0,338	-0,773	0,394	0,545	-0,728	-0,543	-0,488						
N07	0,092	-0,299	0,530	-0,346	-0,767	0,232	0,595							
N08	0,741	0,529	**0,904**	**-0,955**	**-0,958**	0,129								
N09	0,511	0,280	0,529	0,035	-0,237									
N10	-0,637	-0,350	**-0,889**	0,839										
N13	-0,772	-0,676	-0,820											
N14	**0,893**	0,641												
N15	**0,911**													

Tableau 105

Dans la matrice de corrélation de Pearson présentée dans le tableau 18, l'on relève des coefficients de corrélation significatifs, positifs et négatifs, et des coefficients de corrélation non significatifs pour lesquelles l'hypothèse nulle d'absence de corrélation ne peut être rejetée. Nous avons mis en gras seulement les coefficients correspondants aux corrélations significatives, directes et inverses.

Les Nuits les plus fortement corrélées entre elles sont, dans l'ordre décroissant, les Nuits 0 et 7 (0,945), les Nuits 15 et 16 (0,911), les Nuits 8 et 14 (0,904), les Nuits 14 et 16 (0,893) et les Nuits 3 et 6 (0,881).

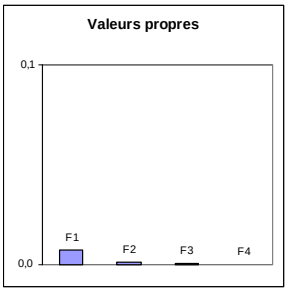
Celles qui sont les plus fortement opposées sont, dans l'ordre décroissant, les Nuits 8 et 10 (- 0,958), les Nuits 8 et 13 (- 0,955), les Nuits 3 et 9 (- 0,938), les Nuits 2 et 15 (- 0,889) et les Nuits 10 et 14 (- 0,889).

7. Analyse factorielle des correspondances

Partant du principe que la représentation graphique est la plus riche en information et qu'elle peut synthétiser un grand tableau de chiffres, le but principal de l'analyse factorielle des correspondance (AFC) est de représenter dans le même espace les lignes et les colonnes du tableau à synthétiser - les Nuits et les catégories lexicales - de telle façon que chaque Nuit (ou chaque catégorie lexicale) ait une seule localisation résumant au mieux la distance variable entre elle et chacune des catégories lexicales (ou chacune des Nuits). La représentation graphique que l'on obtiendra sur la base de ces considérations sera, pour notre corpus, beaucoup plus claire et plus synthétique, à la fois du tableau de répartition des catégories lexicales selon les Nuits (tableau 96) et de tous les graphiques que nous avons présentés plus haut concernant les courbes de distribution des effectifs des catégories lexicales. Cette représentation graphique est présentée dans la figure 185.

Valeurs propres et pourcentage de variance

	F1	F2	F3	F4
Valeur propre	0,007	0,001	0,001	0,000
% variance	74,145	15,740	8,668	1,447
% cumulé	74,145	89,885	98,553	100,000



Ce graphique est un graphique asymétrique puisque les points-lignes et les points-colonnes sont représentés dans deux échelles différentes. les deux axes retenus F1 et F2 ont des valeurs propres respectivement, de 0,007 et 0,001 et des pourcentages d'inertie de 74,14 % et 15,74 %. Les points-Nuits ont deux dispositions combinées ; un nuage de six points-Nuits ramassés autour du centroïde moyen et le reste des points-Nuits sont dispersés.

Contributions des points-lignes (%)

	F1	F2	F3	F4
Verbes	7,712	70,757	4,555	5,539
Noms primitifs	45,705	25,844	1,494	9,696
Noms dérivés	33,585	2,965	34,832	25,325
Adjectifs	6,274	0,052	59,031	27,378
Mots-Outils	6,725	0,382	0,087	32,063

Cosinus carrés des points-lignes

	F1	F2	F3	F4
Verbes	0,330	0,643	0,023	0,005
Noms primitifs	0,887	0,106	0,003	0,004
Noms dérivés	0,866	0,016	0,105	0,013
Adjectifs	0,457	0,001	0,503	0,039
Mots-Outils	0,904	0,011	0,001	0,084

Tableau 106

Contributions des points-colonnes (%)

	F1	F2	F3	F4
N0	1,263	2,691	12,150	0,009
N1	0,311	7,185	11,940	13,114
N2	0,193	2,718	20,131	0,200
N3	0,066	15,518	0,254	1,613
N4	2,507	11,235	0,951	0,406
N5	1,163	4,885	13,222	3,008
N6	0,009	2,143	0,314	27,427
N7	0,974	1,054	4,222	4,396
N8	14,060	2,696	2,915	19,524
N9	0,016	33,308	1,282	6,038
N10	66,985	0,693	2,024	11,417
N13	3,654	5,719	4,749	4,318
N14	8,126	9,282	11,307	6,226
N15	0,030	0,032	3,031	1,452
N16	0,642	0,840	11,508	0,853

Cosinus carrés des points-colonnes

	F1	F2	F3	F4
N0	0,388	0,175	0,436	0,000
N1	0,089	0,437	0,400	0,073
N2	0,062	0,185	0,752	0,001
N3	0,019	0,963	0,009	0,009
N4	0,500	0,476	0,022	0,002
N5	0,306	0,273	0,406	0,015
N6	0,009	0,439	0,035	0,516
N7	0,548	0,126	0,278	0,048
N8	0,916	0,037	0,022	0,025
N9	0,002	0,961	0,020	0,016
N10	0,991	0,002	0,004	0,003
N13	0,663	0,220	0,101	0,015
N14	0,704	0,171	0,115	0,011
N15	0,072	0,016	0,845	0,068
N16	0,294	0,082	0,616	0,008

Tableau 107

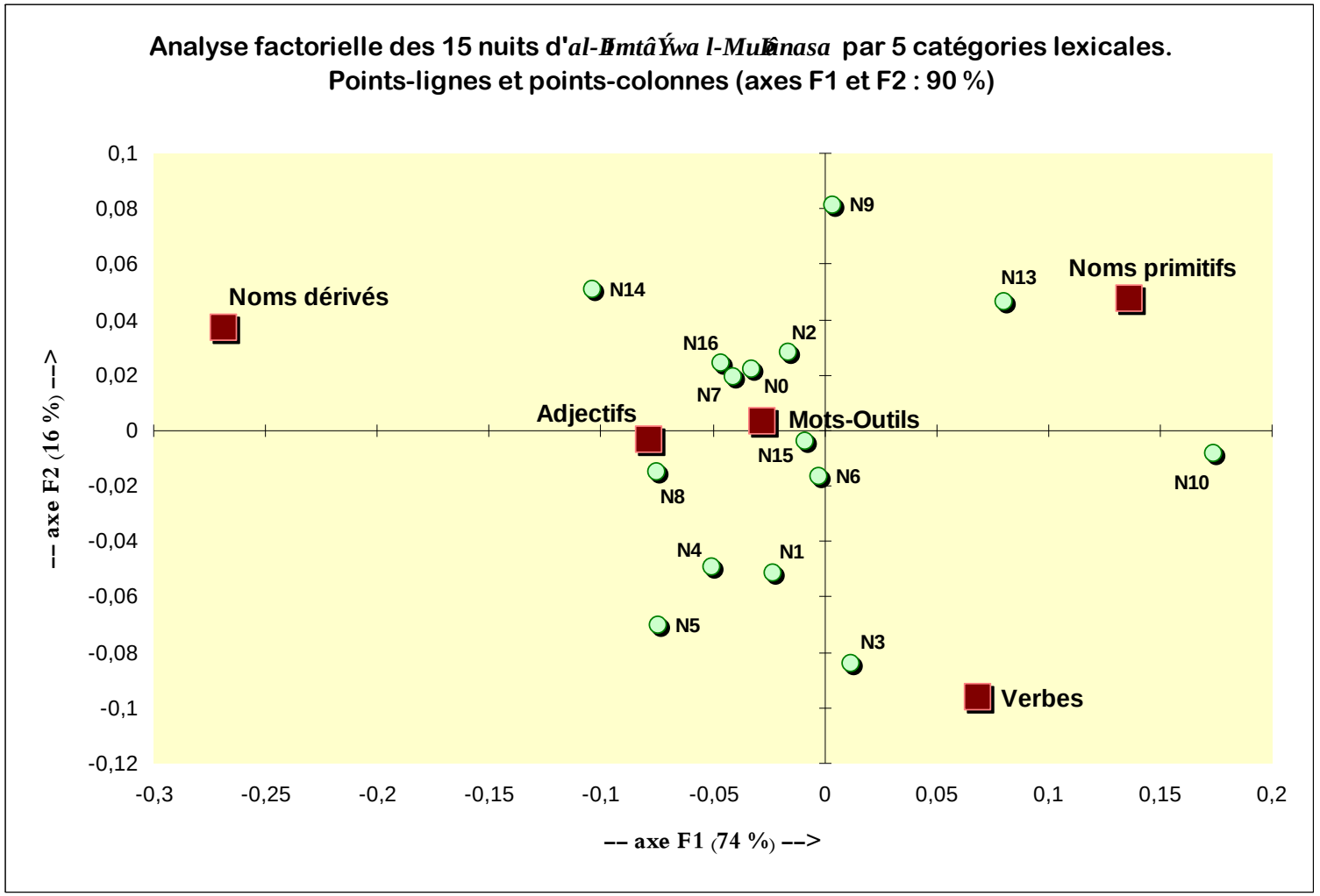

Analyse factorielle des 15 nuits d'al-Imtâ'wa l-Mu'ânasa par 5 catégories lexicales.
Points-lignes et points-colonnes (axes F1 et F2 : 90 %)
0,1
0,08
0,06
0,04
0,02
0
-0,02
-0,04
-0,06
-0,08
-0,1
-0,12
-0,3
-0,25
-0,2
-0,15
-0,1
-0,05
0
0,05
0,1
0,15
0,2
-- axe F2 (16 %) -->
-- axe F1 (74 %) -->
N9
N14
N13
Noms primitifs
Noms dérivés
N16
N2
N0
N7
Adjectifs
Mots-Outils
N15
N6
N10
N8
N4
N1
N5
N3
Verbes

Figure 185

L'axe F1, qui use de 74 % de l'inertie, oppose nettement, au niveau des catégories lexicales, les noms primitifs et les verbes, à droite du graphique, aux noms dérivés, adjectifs et mots-outils, à gauche du graphique, les mots-outils se rapprochant de l'intersection des axes, c'est-à-dire du centroïde moyen. Cette disposition autour de l'axe F1 s'explique principalement par le système d'opposition/similitude étudié plus haut concernant les corrélations. Dans ce système l'on a détecté une très forte opposition entre noms primitifs et mots-outils, et entre noms primitifs et noms dérivés ; une forte opposition entre noms primitifs et adjectifs, et entre verbes et mots-outils ; enfin une opposition moins importante entre verbes et adjectifs, et entre verbes et noms dérivés. Cependant, ce jeu d'oppositions détectées entre les catégories lexicales, doit être combiné à un jeu de similitudes décelées entre ces mêmes catégories lexicales et où l'on a une forte corrélation entre noms dérivés et mots-outils, et entre adjectifs et mots-outils ; une corrélation moins forte entre adjectifs et noms dérivés, et entre verbes et noms primitifs est aussi révélée.

Au niveau des nuits, l'axe F1 met en opposition les Nuits excédentaires en noms primitifs, les Nuits 9, 10 et 13, positionnées à droite de l'axe F2, aux Nuits déficitaires en noms primitifs, toutes les autres Nuits, positionnées à gauche de l'axe F2. Plus les Nuits sont placées à droite, plus elles sont excédentaires en noms primitifs ; inversement, plus les Nuits sont placées à gauche, plus elles sont déficitaires en noms primitifs.

L'axe F2 qui n'utilise que 16 % de l'inertie, est interprétable quant à lui, comme un axe opposant les Nuits excédentaires en verbes, les nuits 1, 3, 4, 5, 6, 10 et 15, placées au-dessous de l'axe F1, aux Nuits déficitaires en verbes, le reste des Nuits, placées au-dessus de l'axe F1. Plus les Nuits sont placées en haut, plus elles sont déficitaires en verbes ; inversement, plus les Nuits sont placées en bas, plus elles sont excédentaires en verbes.

Les Nuits 0, 2 7 et 16 sont très rapprochées les unes des autres auxquelles se joignent les Nuits 15 et 6 pour former un nuage ramassé autour des mots-outils du fait de leurs profils similaires en ce sens où elles sont très faiblement excédentaires en mots-outils.

Les Nuits dispersées loin du centroïde moyen comme les Nuits 10, 13, 9, 14, 5 et 3, elles le sont pour des raisons diverses se rapportant au système d'opposition qui règle le jeu des catégories lexicales :

- La Nuit 10, en effet, occupe cette position à cause à la fois de son grand excédent en noms primitifs et en verbes, et de son énorme déficit en noms dérivés ainsi que de son déficit plus ou moins important en adjectifs et en mots-outils.

- La Nuit 9 fuit les verbes en se plaçant à leur opposé tout en restant proche à la fois des noms primitifs et des noms dérivés à cause de son considérable déficit en verbes et de son excédent en noms primitifs et en noms dérivés.

- La Nuit 3 quant à elle, étant donné qu'elle est très excédentaire en verbes, elle se rapproche d'eux tout en s'éloignant des noms dérivés qui lui font cruellement défaut.

- Du fait de son immense déficit en noms dérivés, de son léger déficit en verbes et de son assez important excédent en noms primitifs, la Nuit 13 s'est trouvée cette place proche des noms primitifs, moyennement éloignée des verbes et totalement à l'opposé des noms dérivés.

- Extrêmement excédentaire en noms dérivés et déficitaire en noms primitifs et, à un degré plus important, en verbes, la Nuits 14 qui a un profil inverse de celui de la Nuit 13, s'est rapprochée autant que faire se peut des noms dérivés en s'éloignant et des noms primitifs et des verbes.

- Enfin, étant excédentaire en adjectifs et en verbes, et déficitaire à la fois en noms dérivés et, surtout, en noms primitifs, la Nuit 5 devait se trouver une position qui soit à la fois proche des adjectifs et des verbes, et loin des noms primitifs et des noms dérivés ; elle l'a trouvée précisément à cet endroit du quart inférieur gauche du graphique.

Chapitre 13

La connexion lexicale

Née de la notion de "parenté lexicale", la connexion lexicale, élaborée et définie par Charles Muller (Muller, 1967 et 1977), permet de savoir dans quelle mesure deux textes partagent le même contenu lexical ou présentent des similitudes au niveau de la structure lexicale. De ce fait, elle peut être utile pour caractériser une œuvre, le style d'un auteur ou le vocabulaire d'une époque ou d'un genre. La mesure de connexion lexicale a même constitué une piste de recherche par une partie des travaux lexicométriques sur l'attribution d'auteur.

La connexion lexicale paraît certes comme une question relative au contenu, mais nous la traitons ici, dans cette partie consacrée aux faits de la structure lexicale, parce que justement « la connexion lexicale n'est pas seulement affaire de contenu, mais aussi de structure »[321]. Elle peut, en effet, relever du contenu lexical, non dans le sens où les mots composant le vocabulaire sont considérés individuellement, sous leur aspect thématique, mais parce qu'ils sont vus sous l'angle de leur parenté ou de leur indépendance en considérant et en manipulant leurs effectifs.

Pour ce faire, les parties du corpus sont comparées deux à deux pour évaluer l'effectif des mots qu'elles ont en commun et celui des mots qu'elles ont en propre. Le nombre des parties de notre corpus étant de 15 Nuits, le nombre des comparaisons binaires s'élève à : $\frac{15 \times 14}{2} = 105$. Pour chacune de ces comparaisons binaires, nous avons calculé :

- l'étendue du vocabulaire de la Nuit A et celle de la Nuit B
- l'étendue du vocabulaire des deux Nuits réunies dans le même ensemble
- la part du vocabulaire commun aux deux Nuits (Vocabulaire commun)
- la part privative de chacune des deux Nuits, c'est-à-dire le nombre des vocables appartenant à A ou à B mais pas aux deux en même temps (Vocabulaire exclusif)
- le coefficient de connexion lexicale, qui est défini comme le rapport du vocabulaire commun aux deux Nuits sur le vocabulaire exclusif.

En plus de ces calculs concernant les vocables, nous avons aussi fait les mêmes calculs pour les occurrences.

[321] Ch. Muller, *Principes et méthodes de statistique lexicale*, 1992, p. 146

Soient donc une Nuit A et une Nuit B, nous désignons par :

V_A : Nombre des vocables de A

V_B : Nombre des vocables de B

V_{AB} : Nombre des vocables communs à A et à B (Vocabulaire commun)

V_{A+B} (= $V_A + V_B - V_{AB}$) : Nombre des vocables appartenant à A ou à B mais pas aux deux en même temps (Vocabulaire exclusif)

CV = V_{AB} / V_{A+B} : Coefficient de connexion lexicale entre A et B qui est égal au rapport du vocabulaire commun sur le vocabulaire exclusif

V_{AB} V_{A+B} ($=V_A+V_B-V_{AB}$)	$CV = \frac{V_{AB}}{V_{A+B}}$

Pour les occurrences, l'on a :

N_A : Nombre des occurrences de A

N_B : Nombre des occurrences de B

N_{AB} : Nombre des occurrences communes à A et à B (Vocabulaire commun)

N_{A+B} (= $N_A+N_B-N_{AB}$) : Nombre des occurrences appartenant à A ou à B mais pas aux deux en même temps (Vocabulaire exclusif)

CN = N_{AB} / N_{A+B} : Le coefficient de connexion lexicale qui est égal au rapport du vocabulaire commun sur le vocabulaire exclusif

N_{AB} N_{A+B} ($=N_A+N_B-N_{AB}$)	$CN = \frac{N_{AB}}{N_{A+B}}$

Exemple :
Calculons la connexion lexicale, au niveau des vocables, entre la Nuit 0 et la Nuit 1 :

V_{N0}	1 584
V_{N1}	803
$V_{N0} + V_{N1}$	2 387
$V_{N0 \times N1}$	240
$V_{N0+N1} = V_{N0} + V_{N1} - V_{N0.N1}$	2 147

Le coefficient de connexion lexicale entre la Nuit 0 et la Nuit 1 est donc :

$$\mathbf{CV} = \frac{V_{N0\times N1}}{V_{N0+N1}} = \frac{240}{2147} = \mathbf{0,112}$$

De la même manière, nous avons donc calculé les coefficients des 105 comparaisons binaires au niveau des vocables et ceux des 105 au niveau des occurrences. Les données sont présentées dans le tableau 1 pour les premiers et dans le tableau 2 pour les seconds.

Par ailleurs, il faut savoir que la connexion lexicale entre deux textes est d'autant plus forte que :

- le vocabulaire commun à ces deux textes est plus important
- le vocabulaire exclusif est plus faible ou comprenant des vocables de faibles fréquences.

Inversement, la connexion lexicale entre deux textes est d'autant plus faible que :

- le vocabulaire commun à ces deux textes est plus faible et qu'il ne comporte surtout que des vocables de fréquence élevée.
- le vocabulaire exclusif est plus important ou comprenant des vocables de fréquence moyenne ou élevée.

Connexion lexicale des vocables dans *al-ÞImtâÝ wa l-MuÞânasa*

	Nuit 0	Nuit 1	Nuit 2	Nuit 3	Nuit 4	Nuit 5	Nuit 6	Nuit 7	Nuit 8	Nuit 9	Nuit 10	Nuit 13	Nuit 14	Nuit 15
Nuit 1	0,112													
Nuit 2	0,072	0,135												
Nuit 3	0,117	0,116	0,123											
Nuit 4	0,099	0,111	0,123	0,107										
Nuit 5	0,122	0,106	0,098	0,108	0,090									
Nuit 6	0,123	0,099	0,112	0,089	0,116	0,064								
Nuit 7	0,107	0,119	0,117	0,107	0,111	0,107	0,099							
Nuit 8	0,130	0,111	0,120	0,091	0,135	0,064	0,138	0,099						
Nuit 9	0,105	0,121	0,121	0,100	0,118	0,083	0,120	0,102	0,121					
Nuit 10	0,077	0,070	0,079	0,068	0,082	0,044	0,098	0,052	0,097	0,093				
Nuit 13	0,098	0,128	0,124	0,115	0,101	0,105	0,094	0,110	0,095	0,127	0,072			
Nuit 14	0,097	0,103	0,120	0,096	0,098	0,094	0,112	0,097	0,119	0,135	0,072	0,129		
Nuit 15	0,087	0,118	0,114	0,100	0,093	0,112	0,089	0,108	0,087	0,115	0,064	0,138	0,128	
Nuit 16	0,084	0,123	0,107	0,116	0,097	0,128	0,080	0,104	0,073	0,095	0,053	0,129	0,108	0,125

Tableau 108

Connexion lexicale des occurrences dans *al-ÞImtâÝ wa l-MuÞânasa*

	Nuit 0	Nuit 1	Nuit 2	Nuit 3	Nuit 4	Nuit 5	Nuit 6	Nuit 7	Nuit 8	Nuit 9	Nuit 10	Nuit 13	Nuit 14	Nuit 15
Nuit 1	2,399													
Nuit 2	1,615	2,463												
Nuit 3	2,519	2,229	2,236											
Nuit 4	2,231	2,382	2,486	2,306										
Nuit 5	2,456	1,848	1,901	1,916	1,932									
Nuit 6	2,574	2,336	2,452	2,154	2,597	1,920								
Nuit 7	2,335	2,233	2,291	2,128	2,355	1,832	2,349							
Nuit 8	2,827	2,677	2,749	2,354	2,975	1,857	3,192	2,489						
Nuit 9	2,364	2,478	2,497	2,037	2,454	1,807	2,631	2,143	2,689					
Nuit 10	1,750	1,628	1,689	1,583	1,827	1,249	2,072	1,532	2,155	1,974				
Nuit 13	1,911	2,270	2,347	2,045	2,113	1,851	2,287	2,030	2,240	2,462	1,681			
Nuit 14	2,220	2,252	2,321	2,039	2,234	1,838	2,471	2,106	2,709	2,662	1,670	2,322		
Nuit 15	2,047	2,216	2,184	1,901	2,134	1,899	2,164	2,011	2,239	2,275	1,563	2,440	2,338	
Nuit 16	1,984	2,236	2,067	2,024	2,130	2,061	2,054	1,979	2,107	2,130	1,415	2,149	2,087	2,118

Tableau 109

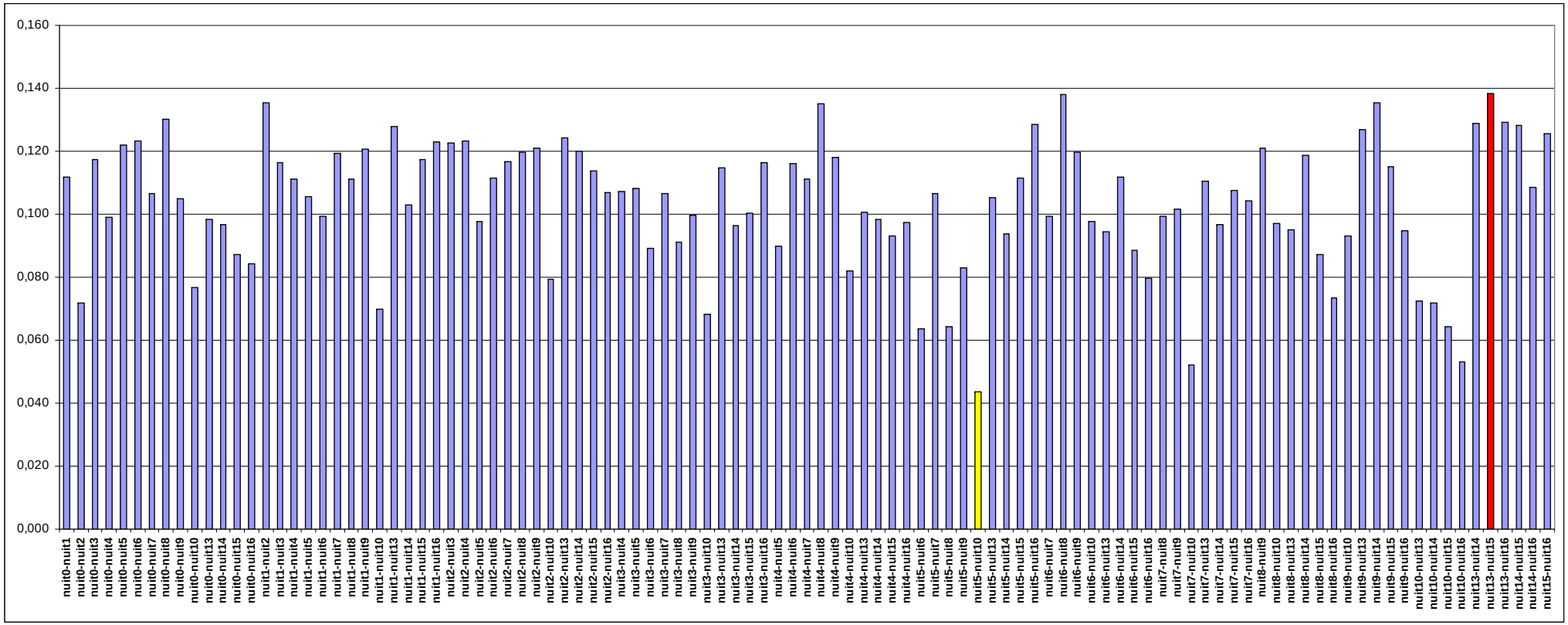


Figure 186
Connexion lexicale des vocables (en rouge, la connexion la plus forte et en jaune, la connexion la plus faible)

À la fin de cette série de comparaisons binaires, nous avons relevé deux pelotons d'une dizaine de couples chacun, le peloton de tête comprenant les Nuits ayant la connexion lexicale la plus forte et le peloton de queue composé des Nuits ayant la connexion lexicale la plus faible.

Connexion lexicale forte (vocables)		
1	Nuit 13 – Nuit 15	0,1385
2	Nuit 6 – Nuit 8	0,1379
3	Nuit 9 – Nuit 14	0,1354
4	Nuit 1 – Nuit 2	0,1354
5	Nuit 4 – Nuit 8	0,1349
6	Nuit 0 – Nuit 8	0,1302
7	Nuit 13 – Nuit 16	0,1292
8	Nuit 13 – Nuit 14	0,1290
9	Nuit 5 – Nuit 16	0,1285
10	Nuit 14 – Nuit 15	0,1281

Il est à noter que les Nuits qui sont les plus proches thématiquement sont celles où la connexion lexicale est la plus forte ; inversement celles dont les thèmes sont les plus éloignés sont celles où la connexion lexicale est la plus faible. C'est dans cette dernière perspective qu'il faut comprendre la présence, dans le peloton de queue, de la Nuit 10 dans sept des dix couples ayant la connexion lexicale la plus faible. En effet, dans la Nuit 10 TawÎîdî traite du thème des étrangetés chez les animaux, thème qu'il n'a traité nulle part ailleurs. Ce qui a pour conséquence d'avoir un vocabulaire commun à la Nuit 10 et à n'importe quelle autre Nuit du corpus, plus faible que n'importe quel autre vocabulaire commun.

La connexion lexicale la plus forte de tout le corpus est celle qui lie la Nuit 13 à la Nuit 15 avec un coefficient de 0,1385. Les deux Nuits étant très proches thématiquement l'une de l'autre puisque TawÎîdî traite dans la Nuit 13 de deux thèmes : l'essence de l'âme et les types d'esprit ; et dans la Nuit 15, il traite principalement de l'esprit et de sa relation au concret.

Le coefficient de connexion lexicale le plus faible de tout le corpus est de 0,0437 correspondant à la connexion lexicale entre la Nuit 5 et la Nuit 10 ; ce qui fait de ces deux Nuits, les Nuits les plus indépendantes lexicalement.

Connexion lexicale faible (vocables)		
96	Nuit 10 – Nuit 14	0,0718
97	Nuit 0 – Nuit 2	0,0717
98	Nuit 1 – Nuit 10	0,0700
99	Nuit 3 – Nuit 10	0,0682
100	Nuit 10 – Nuit 15	0,0642
101	Nuit 5 – Nuit 8	0,0641
102	Nuit 5 – Nuit 6	0,0637
103	Nuit 10 – Nuit 16	0,0531
104	Nuit 7 – Nuit 10	0,0521
105	Nuit 5 – Nuit 10	0,0437

Au niveau de la connexion des occurrences, la toute première place est occupée par le couple Nuit 6 – Nuit 8 avec un coefficient de 3,192 (il occupe la deuxième place au niveau des vocables), alors que la toute dernière place est confirmée pour le couple Nuit 5 – Nuit 10 avec un coefficient de 1,249.