

Attribution d'auteur : Application des méthodes de *qsums* au français

Judith Czellar

Université de Genève – FPSE – CH-1211 Genève 4 – Suisse

Abstract

The techniques of authorship attribution are well known in the literature. Among them is the use of the cumulative sum test (*qsum*, *cusum*). Although widely criticized, it was used in a legal context and for authorship identification. Until now, this technique has never been applied to the French language. The first goal of this paper, is to present the *qsum* technique with some statistical improvements called the “*weighted qsum*” (developed and based statistically by Goldsmith & Woodward 1964, Bissell, 1969, 1990, 1995a, 1995b). The measure of homogeneity of a text, between-text comparisons, as well as weighted *qsum* chart are presented. Secondly, we intend to apply it to the French language. Some characteristics of French will be described in contrast with English. We present the reasons for which indicators used in English, are not valid for the French. Conversely, some new indicators, which could be stable enough for a French-speaking author, will be clarified and tested. Six samples of two French authors were tested with fourteen indicators. We observe a good within-text and within-author homogeneity, for each of them. As for between-author comparisons, four indicators show a sufficiently good discrimination. We conclude that the *wqsum* test is applicable to French. However, we should continue to search for new indicators.

Résumé

Les méthodes d'attribution d'auteur sont largement connues dans la littérature. Parmi elles, celle qui utilise l'application des sommes cumulées (*qsum*, *cusum*), bien que largement controversée, a été utilisée dans un contexte légal et d'identification d'auteur. Jusqu'à présent cette méthode n'a jamais été appliquée au français. L'objectif du présent travail, en tenant compte des controverses, est, dans un premier temps, d'explicitier le test des *qsums* appelé *weighted qsum* développé et fondé statistiquement par Goldsmith et Woodward (1964), Bissell (1969, 1990, 1995a, 1995b). La mesure d'homogénéité d'un texte, les comparaisons intertextes ainsi qu'une représentation graphique adaptée seront présentées. Dans un deuxième temps, nous envisagerons son application au français. Il sera question de certaines caractéristiques de la langue française mises en contraste avec l'anglais. Nous exposerons les raisons pour lesquelles les indicateurs utilisés en anglais ne sont pas valables pour le français. Ainsi, quelques nouveaux indicateurs qui pourraient être suffisamment stables pour un auteur francophone seront explicités et testés. Six échantillons de textes français de deux auteurs ont été testés avec quatorze indicateurs. Nous observons une très bonne homogénéité pour tous les indicateurs, pour les comparaisons intra-textes et intra-auteurs. Quant à la comparaison inter-auteurs, quatre indicateurs se montrent suffisamment discriminatifs. Nous concluons que l'outil des *wqsums* est applicable pour le français. Cependant il faut continuer de rechercher de nouveaux indicateurs.

Mots-clés : *qsum*, *weighted qsum*, attribution d'auteur, français, indicateurs, habitudes langagières, mots-outils.

1. Introduction

La technique des sommes cumulées est traditionnellement utilisée pour les contrôles de qualité dans un contexte industriel. Le problème principal dans ce contexte est de s'assurer que la proportion des articles défectueux produits dans un laps de temps ne dépasse pas une limite prédéfinie. Son principe est de comparer les valeurs successives d'une variable avec une valeur de référence. La somme cumulée des déviations de cette valeur est représentée graphiquement. Le graphique des sommes cumulées représente les séquences des sommes des

différences successives entre les valeurs observées et des valeurs cibles. Si ce cumul dépasse la limite spécifique, ceci indique qu'un changement s'est produit sur la variable.

Goldsmith & al. (1964) soulignent qu'en dehors des contextes industriels, cette technique peut être appliquée à toutes données qui se produisent à intervalles réguliers.

Morton (1978) est le premier à l'adapter aux statistiques textuelles, plus précisément au domaine d'attribution d'auteur. Le principe est d'analyser des séquences linguistiques cumulées, l'unité de base étant la phrase. Plus précisément, il a utilisé la longueur des phrases et la fréquence de certains indices. Selon lui ces indices reflètent des patterns inconscients de l'auteur qui sont indépendants du moment et du sujet d'écriture. Morton souligne qu'il est important de montrer que le travail de l'auteur est statistiquement homogène et qu'il existe des variations homogènes à travers les différentes parties du texte. Ces comparaisons sont nécessaires afin de mettre en évidence que les différences entre les œuvres du même auteur ne sont pas plus grandes que les différences dues à l'échantillonnage, et que les œuvres peuvent donc être considérées comme issues d'une seule population. Ce qui est vrai pour un auteur doit être applicable à toutes les classes d'auteurs.

Le principe à la base de la technique des *qsums* est que chaque individu possède des habitudes langagières qui lui sont propres et qui se manifestent lors de ses communications verbales ou écrites. Ces habitudes sont quantifiables selon Morton (1978). Ce sont des composants particuliers des phrases : l'utilisation de mots courts, des mots outils, dont beaucoup, en anglais, commencent par une voyelle. On peut donc, pour l'anglais, faire l'économie de la statistique *lexicale*, pour ne regarder que la longueur des mots et la présence de voyelles ; ce qui rend l'utilisation de la technique très économique. Ainsi lorsque la proportion des occurrences de ces mots est constante au sein d'un segment de texte, les propos appartiennent probablement à une personne ; dans le cas contraire, le segment est l'expression de plusieurs auteurs. Par ailleurs, cette technique s'applique à des textes très courts, ce qui la rend performante dans des contextes où les statistiques habituelles ne sont pas applicables.

La méthode met donc en relation :

- la longueur de la phrase en nombre de mots, considérée comme unité syntaxique dans un texte ou séquence de texte. La *qsum* est la somme des déviations des longueurs des phrases de la longueur moyenne. Par la suite on cumule ces déviations formant ainsi la *distribution des longueurs des phrases* ;
- le nombre des occurrences d'un indicateur par phrase (« *habit* » *words* opérationnalisés indépendamment des formes). De nouveau, on prend la somme cumulée des déviations de la moyenne de ces formes.

La méthode mesure les proportions des habitudes langagières (indicateurs) relativement à la longueur des phrases. Par la combinaison de ces deux éléments, on obtient le graphique *qsum* (*qsum-chart*), qui est la superposition des *qsums* des longueurs de phrase et des *qsums* des indicateurs permettant ainsi une comparaison visuelle entre les deux. Si les deux courbes coïncident, cela veut dire que le texte provient bien d'un seul auteur (cf. Figure 1a), si elles sont séparées (b), la source peut être de deux ou de plusieurs auteurs (Farrington, 1996 : 60, 69).

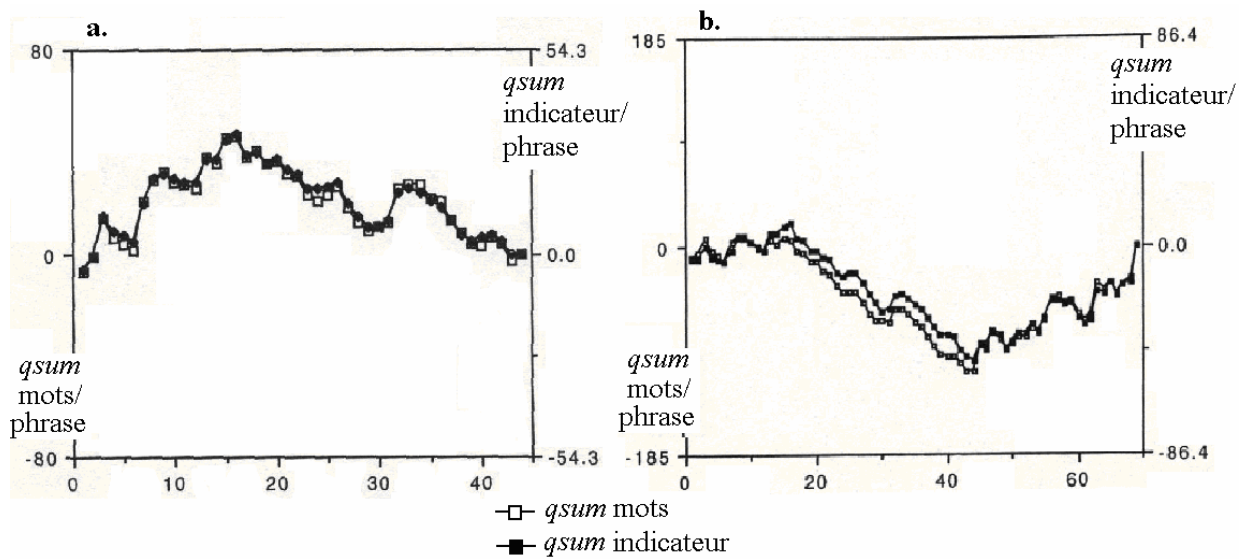


Figure 1 : Graphique de *qsum* pour le même auteur (a) et pour deux auteurs (b)

Vers 1990 déjà cette technique a été utilisée lors de procès et dans un contexte légal d'identification d'auteur en Angleterre, en Irlande et aux États-Unis. Jusqu'à aujourd'hui, elle a été appliquée à des textes en anglais, en latin, en grec et en allemand mais pas en français. Le but du présent travail est de l'appliquer au français.

L'étape clé lors de l'utilisation de cette méthode est de trouver des habitudes langagières (indicateurs) suffisamment stables pour un auteur présumé. En ce qui concerne l'anglais, le plus fréquemment on prend les mots de deux, trois lettres et les mots débutant par une voyelle mais les combinaisons dépendent de l'auteur. Par exemple, Farrington (1996) utilise des mots de deux, trois lettres et ceux débutant par une voyelle pour démontrer leur stabilité longitudinale dans des lettres de Helen Keller. Pour une comparaison des différents styles d'écriture de Muriel Spark, les mots de deux et de trois lettres se montrent efficaces. Ou encore, même l'utilisation des mots de deux, trois, quatre lettres et les mots débutant par une voyelle (indicateur très rare selon l'auteur) paraissent stables dans l'écriture d'un jeune délinquant. La Figure 1a illustre le cas d'un indicateur de deux, trois lettres + voyelles initiales et la Figure 1b, celui de deux et trois lettres.

En ce qui concerne l'anglais, le choix de ces indicateurs est justifié par le fait que parmi eux on trouve notamment les articles définis/indéfinis (*a, the*), les pronoms (*I, he, it*), les formes du verbe être (*to be, was, is*) et les prépositions les plus communes (*in, of, to*). Lorsqu'on enlève les mots à contenu d'un texte, les mots outils forment en quelque sorte le « squelette » du texte. Nous pouvons trouver quelques exemples d'application dans Bissell (1995a, 1995b).

La méthode des sommes cumulées dans le contexte d'attribution d'auteur a été largement controversée. Canter (1992), Hardcastle (1993, 1997) ont analysé un nombre important de corpora et arrivent à la conclusion que cette technique n'est pas fiable. Parmi les critiques émises, celles qui soulignent le manque de fidélité et la subjectivité de la méthode semblent être les plus pertinentes. Cette subjectivité est en grande partie attribuée au fait que la méthode se base sur l'appréciation visuelle de l'écart entre deux courbes. En conséquence, il serait opportun de trouver une mesure statistique plus solide, ce qui a été effectué par Bissell (Bissell, 1990, 1995a, 1995b) sous la forme des *weighted qsums* (sommes cumulées pondérées).

1.1. Sommes cumulées pondérées (*weighted qsums*)

Suite aux investigations statistiques ultérieures, les sommes cumulées pondérées (*weighted qsums*) ont d'une part permis de localiser et de donner une évaluation objective des anomalies des proportions de mots-outils (tests d'homogénéité) et ainsi qu'une représentation graphique plus adaptée. D'autre part, en adoptant un *quasi t-test* (t'), il est possible de faire des comparaisons entre les proportions moyennes de deux segments (Bissell, 1990, 1995a, 1995b).

Comme les *qsums* simples, les *qsums* pondérées mesurent également l'homogénéité des textes en prenant la phrase comme unité de base. Au lieu de cumuler les différences par rapport à la moyenne des longueurs de phrases et des habitudes linguistiques (indicateurs), on pondère le tout par le poids des proportions totales. De telle sorte, on cumule les différences entre les fréquences observées et les fréquences attendues des indicateurs.

Afin de voir s'il y a une différence entre deux textes, on peut estimer la variation moyenne de chaque texte, calculer ensuite un *quasi t-test* entre les textes, l'hypothèse nulle étant que deux textes A et B appartiennent au même auteur (pour les formules cf. Annexe). Plus la mesure t' est élevée, moins il y a de chances que les deux textes aient été écrits par la même personne (Sommers, 1998 ; Sommers & Tweedie, 2003).

Avec cette amélioration, la technique nous semble prometteuse et utile pour au moins trois raisons :

- elle s'applique à des textes courts ;
- elle ne porte pas sur les vocables, donc elle évite les problèmes d'étiquetage et de lemmatisation ;
- contrairement aux analyses en sac de mots, elle prend en compte l'ordre des phrases.

Afin de déterminer des régularités des éléments cibles, dans le contexte des *qsums*, le test du V_{\max} (*Span test*) est le plus couramment utilisé. Pour n'importe quel segment de la *qsum*, V_{\max} indique la distance verticale la plus grande, par rapport à la corde qui relie les extrémités du segment (voir Figure 2). Pour une variable distribuée normalement, V_{\max} est tracé en fonction de la longueur du segment pour pouvoir tester les probabilités de significativité (Goldsmith & Woodward, 1964).

Le graphique de la *wqsum* est semblable à celui de la *qsum* simple, à la différence que l'on tient compte de la longueur des phrases. Les sommes cumulées sont représentées non en fonction de chaque phrase considérée comme unité mais proportionnellement à sa longueur. Ceci signifie que pour les phrases très courtes, au lieu qu'elles déforment le graphique de manière exagérée, la somme des mots n'a que peu de poids. Elles contribuent à une petite distance horizontale. Inversement, les phrases longues contribuent davantage et sont représentées par un « pas » horizontal plus large. L'intégrité des proportions locales est maintenue.

Sur l'axe vertical de la Figure 2, nous avons les sommes cumulées standardisées, en abscisse le nombre de phrases. Nous constatons des « pas » irréguliers allant de gauche à droite en fonction des contributions des mots de chaque phrase.

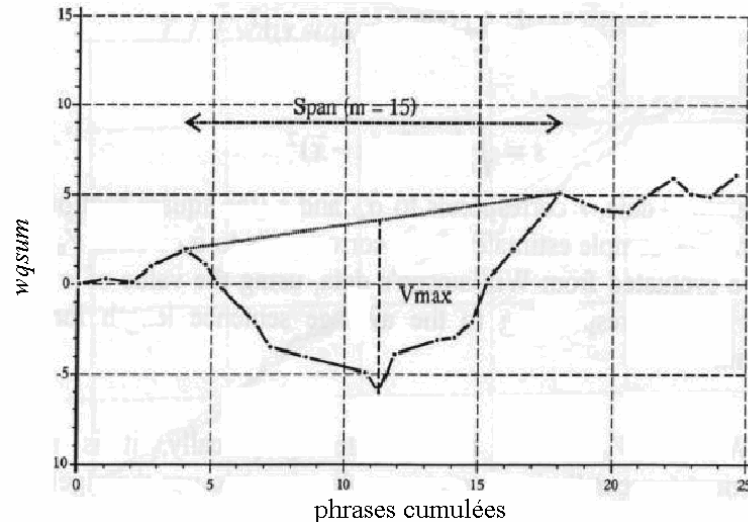


Figure 2 : Exemple de graphique de sommes cumulées pondérées (tirée de Bissell, 1995a : 36)

Il faut noter que l'application des *qsums* exige certains critères. Tout d'abord, il est important que le texte ne contienne aucun passage de dialogue et ne soit extrait ni du début ni de la fin d'une œuvre. Il faut également que l'indicateur testé soit homogène pour le texte et qu'il se répartisse de manière stable. En outre, la proportion globale des occurrences de l'indicateur ne doit être ni trop élevée ni trop basse. Ceci est important, pour que le « squelette » du texte soit suffisamment solide, mais pas trop. Farrington (1996) conseille une proportion des mots inclus dans l'indicateur comprise entre 45 et 55%.

1.2. Caractéristiques du français

Les habitudes langagières valables pour l'anglais ne le sont pas pour le français. L'essentiel de la tâche est de trouver des indicateurs suffisamment constants, qui soient également discriminatifs pour les auteurs francophones. Le présent travail va essentiellement examiner la possibilité d'appliquer la méthode des sommes cumulées au français.

Dans un corpus de textes (n=36879 mots) établi à partir de huit œuvres littéraires en français¹ nous observons que les mots les plus fréquents sont ceux d'une longueur de deux lettres, suivis de ceux de trois lettres (Tableau I). Parmi eux nous trouvons les mots-outils les plus courants, ce que montre le Tableau II. Étant donné que le calcul des *qsums* se base sur l'analyse de mots à non-contenu qui doivent former le « squelette » du texte, ceux qui dépassent une longueur de quatre lettres ne seront pas pris en compte en tant qu'indicateurs.

En ce qui concerne l'anglais il existe peu de mots courts qui ne soient des mots outils. Les mots débutant par une voyelle se justifient également, vu que ce sont souvent des mots qui commencent avec un préfixe latin (ex. : *ab-*, *ac-*, *al-*, *ex-*, *in-*, etc.). Ce n'est pas le cas pour le

¹ Ce sont quarante échantillons, tirés de : *Gros Câlin* (n=7), *Éducation européenne* (n=4), *Terre des Hommes* (n=4), *Vol de nuit* (n=4), *Du côté de chez Swann* (n=4), *À l'ombre des jeunes filles en fleurs* (n=6), *Bel Ami* (n=5), *Une vie* (n=6).

français. Nous avons choisi de prendre quelques groupes de mots spécifiques : les pronoms, les prépositions et les conjonctions qui représentent environ 10% des occurrences.

Longueurs en lettres	Fréquences	%
2	8612	23%
3	5121	14%
4	4539	12%
5	4169	11%
6	3221	9%
1	2882	8%
7	2762	7%
8	2105	6%
9	1497	4%
10	925	3%
11	480	1%
12	298	1%
13	146	<1%
14	68	<1%
15	34	<1%
16	17	<1%
17	1	<1%
18	1	<1%
19	1	<1%

Total 36879

Tableau I : Fréquences absolues et relatives des mots du corpus en fonction de leur longueur

Vu ce qui précède nous avons défini 14 indicateurs (Tableau III). Contrairement à Farrington pour l'anglais, nous n'avons pu nous contenter de définir des critères généraux sur les mots (longueur, présence de voyelles), mais nous avons établi une liste de formes graphiques. Pour l'instant, il n'a pas semblé nécessaire de passer à l'étape ultérieure de désambiguïsation. Les proportions pour les indicateurs B, E, G et M sont un peu faibles selon les critères de Farrington, et pour l'indicateur K, un peu élevée. Nous allons voir par l'exemple d'application qui suit que ceci n'affecte en rien la mesure de l'homogénéité ni celle de discriminabilité. Il est important de noter qu'un mot n'est compté qu'une fois. En conséquence, si un mot tombe dans la catégorie des conjonctions et également dans celle des « trois lettres », il n'est comptabilisé qu'une seule fois.

Mots employés	Fréquences	Longueurs
de	1610	2
et	1036	2
la	833	2
à	771	1
les	695	3
le	679	2
il	616	2
l	594	1
un	542	2
d	529	1
que	506	3
une	464	3
qui	417	3
en	393	2
dans	386	4
des	361	3
qu	357	2

Tableau II : Les 17 formes les plus fréquentes dans le corpus

2,3 lettres+préposition+conjonction+pronom	Indicateur A	50%
2,3+pronom	Indicateur B	41%
2,3 lettres+préposition+conjonction	Indicateur C	46%
1,2 lettres +préposition+conjonction+pronom	Indicateur D	44%
1,2 lettres+pronoms	Indicateur E	35%
1,2 lettres préposition+conjonction	Indicateur F	40%
1,2 lettres	Indicateur G	31%
1,2,3 lettres+ préposition+conjonction	Indicateur H	54%
1,2,3 lettres+pronom	Indicateur I	48%
1,2,3 lettres+préposition+conjonction	Indicateur J	54%
1,2,3,4 lettres	Indicateur K	57%
1,2,3 lettres	Indicateur L	45%
2,3,4 lettres	Indicateur M	50%
2,3 lettres	Indicateur N	37%

Tableau III : Liste des indicateurs utilisés avec la proportion d'occurrences qu'ils représentent dans le corpus

2. Essais d'application au français

2.1. Textes

Six textes ont été échantillonnés chez deux auteurs différents : deux échantillons de *Gros Câlin* (Ajar, 1974) : gc1, gc2 ; deux de *Bel Ami* (Maupassant, 1885) : ba1, ba2 ; et deux de *Une vie* (Maupassant, 1883) : vie1, vie2. Vingt-cinq phrases consécutives sans dialogues ont été sélectionnées vers le milieu des œuvres. Dans un premier temps, nous nous sommes intéressés uniquement aux caractéristiques des distributions de nos échantillons. Par la suite nous avons effectué des comparaisons intertextes.

2.2. Résultats

Pour tester l'homogénéité de nos échantillons, nous avons utilisé deux types de mesure : celle qui se base sur l'estimation de la déviation standard et celle du V_{\max} . Une fois l'homogénéité établie, les textes ont été comparés deux à deux par le biais du t' -test.

La Figure nous montre un exemple de représentation des sommes cumulées pondérées appliquées pour tester l'homogénéité de l'échantillon « vie1 » pour l'indicateur K. Nous pouvons constater une répartition relativement équitable des proportions des mots outils dans les phrases : mis à part les phrases 16 et 20, où on observe un « pas » plus long, les autres phrases se répartissent de manière homogène. L'appréciation visuelle du graphique permet de saisir intuitivement l'homogénéité mais le calcul seul permet de conclure. Ici l'échantillon s'avère être homogène (calculé entre les phrases 4 à 18, $V_{\max}\sigma_A = -4.01$, $p = ns$, $V_{\max}\sigma_D = -3.71$, $p = ns$).

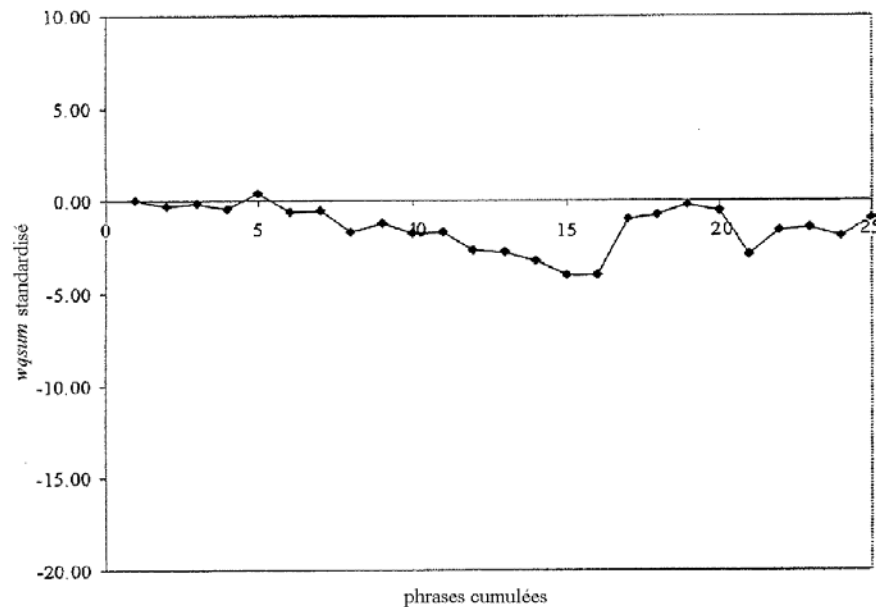


Figure 3 : Wqsum pour l'échantillon « vie1 » (n=25), indicateur K

Nous avons effectué le test d'homogénéité pour tous les indicateurs et pour tous les échantillons. Tous se montrent homogènes pour les textes sauf l'indicateur B pour gc2 ($u=2.1435$, $p<.05$).

Les comparaisons inter-textes indiquent d'abord une stabilité intra-auteur : la quasi-totalité des indicateurs utilisés ne mettent pas en évidence des différences significatives. Sur les 98 comparaisons, la seule exception concerne l'indicateur H pour les textes vie1-vie2 (voir Tableau IV, $t_{48}=-2.94$, $p<.05$).

Les comparaisons inter-auteurs (Tableau IV), montrent en revanche que les indicateurs n'ont pas le même pouvoir discriminant. Sur huit comparaisons inter-auteurs (gc1 et ba1-ba2-vie1-vie2 ; gc2 et ba1-ba2-vie1-vie2) nous constatons les faits suivants : l'indicateur A discrimine une fois ($p<.05$), le G trois fois ($p<.05$), le L quatre fois ($p<.05$), le J et le K cinq fois ($p<.05$) et le D, le F, le H et le I différencient six fois un auteur de l'autre ($p<.05$). Ainsi, se basant sur les résultats des comparaisons de ces échantillons, si nous souhaitons sélectionner les indicateurs qui seraient susceptibles de différencier Ajar et Maupassant, nous pourrions prendre ces quatre derniers. Cependant nous pourrions privilégier en particulier l'indicateur F : à part le fait qu'il discrimine six fois sur huit comparaisons (gc1 avec vie1-vie2 et gc2 avec ba1-ba2-vie1-vie2, $p<.05$), les deux autres ont des tendances significatives (gc1 avec ba1-ba2, $p<.10$).

Méthodes	gc2 (n=25)	ba1 (n=25)	ba2 (n=25)	vie1 (n=25)	vie2 (n=25)	
gc1 (n=25)	A	0.77	-0.82	-0.91	0.61	-2.40
	B	0.51	-0.45	-0.72	1.01	-1.07
	C	0.83	-0.46	-0.98	-0.19	-2.02
	D	1.25	-2.97	-2.05	-2.16	-3.89
	E	0.84	-2.48	-1.57	-1.18	-1.34
	F	0.76	-2.55	-2.67	-2.84	-3.79
	G	0.45	-2.22	-2.23	-1.88	-1.37
	H	-0.11	-3.37	-3.42	-2.16	-5.78
	I	-0.34	-2.91	-3.12	-1.82	-4.14
	J	-0.02	-2.77	-3.42	-2.71	-4.98
	K	-0.88	-4.45	-3.28	-2.79	-5.04
	L	-0.23	-2.45	-3.28	-2.47	-3.61
	M	-0.02	-1.47	-0.74	-0.18	-1.55
	N	0.60	-0.14	-0.83	0.19	-0.76
gc2 (n=25)	A		-1.40	-1.42	-0.15	-2.88
	B		-0.91	-1.14	0.48	-1.62
	C		-1.15	-1.58	-0.98	-2.71
	D		-3.76	-2.91	-3.02	-4.70
	E		-3.42	-2.34	-1.99	-2.23
	F		-3.12	-3.27	-3.43	-4.42
	G		-2.78	-2.77	-2.41	-1.94
	H		-3.09	-3.16	-1.92	-5.20
	I		-2.56	-2.77	-1.42	-3.65
	J		-2.78	-3.44	-2.72	-5.04
	K		-3.45	-2.45	-1.89	-3.91
	L		-2.29	-3.13	-2.29	-3.45
	M		-1.50	-0.74	-0.17	-1.58
	N		-0.67	-1.35	-0.47	-1.41
ba1 (n=25)	A			-0.19	-1.27	-1.22
	B			-0.30	-1.37	-0.49
	C			-0.51	-0.28	-1.26
	D			0.77	-0.72	-0.24
	E			0.63	-1.12	1.15
	F			0.09	0.07	-0.43
	G			-0.07	-0.31	1.02
	H			-0.16	-1.30	-1.27
	I			-0.25	-1.33	-0.51
	J			-0.49	-0.37	-1.29
	K			0.58	-1.41	-0.21
	L			-0.59	-0.40	-0.54
	M			0.49	-1.26	-0.13
	N			-0.64	-0.33	-0.53
ba2 (n=25)	A				-1.30	-0.85
	B				-1.55	-0.11
	C				-0.81	-0.57
	D				0.07	-1.12
	E				-0.42	0.40
	F				0.17	-0.58
	G				-0.36	1.05
	H				-1.42	-1.04
	I				-1.58	-0.21
	J				-0.93	-0.76
	K				-0.69	-0.80
	L				-1.12	0.12
	M				-0.58	-0.59
	N				-1.05	0.22
vie1 (n=25)	A					-2.73
	B					-2.22
	C					-1.76
	D					-1.06
	E					-0.05
	F					-0.38
	G					0.67
	H					-2.94
	I					-2.19
	J					-1.95
	K					-1.70
	L					-1.15
	M					-1.34
	N					-1.07

en **gras** : valeurs significatives, en **gras italique** : valeurs à tendances significatives
 en gris : comparaisons intratexte ; en encadré gras : comparaison intra-auteur

Tableau IV : Valeur t'_{dr} pour les comparaisons deux à deux de tous les textes

La Figure 4 illustre avec un graphique $Wqsum$ un aperçu plus détaillé des comparaisons des textes vie2 et gc2 pour l'indicateur F (une flèche indique l'endroit de la jonction). Les deux textes ont été accolés pour effectuer le test du V_{max} . Nous observons des phrases beaucoup plus régulières et courtes dans vie2 que dans gc2. L'échantillon gc2 a des phrases plus longues (phrase 29, 30, 39) en alternance avec des phrases plus courtes (phrases 32, 33, 41-50²). Cette tendance s'observe à partir de la phrase 26, l'endroit où débute gc2. Outre l'allure irrégulière du graphique, le t' -test nous donne une différence significative ($t_{48}=-4.42$, $p<.0001$), qui est également confirmée par V_{max} calculé sur l'empan de 50 phrases ($V_{max}\sigma = -13.08$, $p<.01$, $V_{max}\sigma_D = -5.66$, $p<.01$).

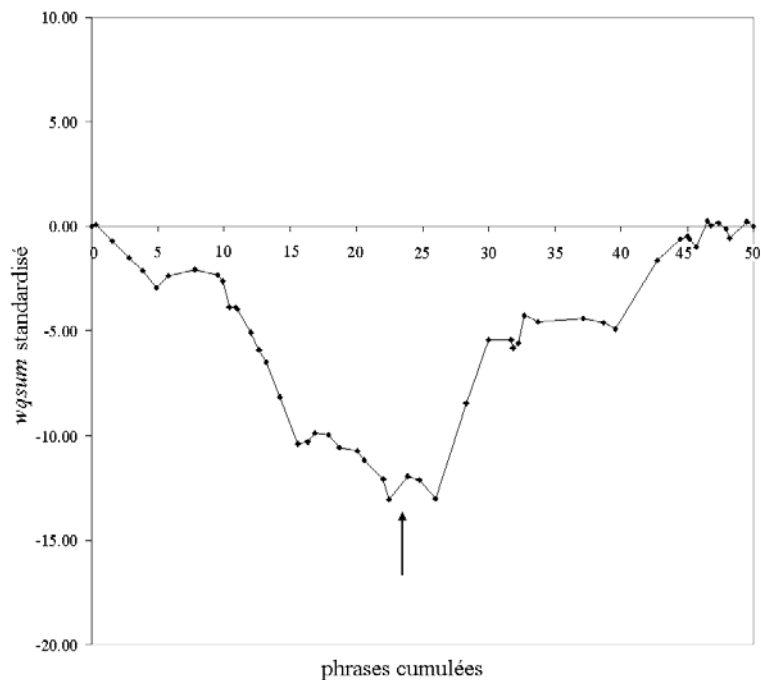


Figure 4 : $Wqsum$ pour « vie2 » ($n=25$) suivi de « gc2 » ($n=25$) pour l'indicateur F

3. Discussion

Ce travail a un but exploratoire, à savoir l'application du test des sommes cumulées pondérées au français. Au vu des résultats, qui semblent très prometteurs, nous constatons qu'il est possible d'utiliser cette méthode pour explorer des textes en français. Sur des comparaisons de six échantillons de textes très courts (25 phrases) de deux auteurs différents nous avons pu trouver grâce à cette méthode une bonne homogénéité intra-textes et une relativement grande hétérogénéité inter-textes. Outre ce fait, les extraits de deux œuvres différentes d'un même auteur se révèlent homogènes tout en étant différenciées des deux textes d'un autre auteur.

Nous soupçonnons que la quantité des occurrences retenues pour un indicateur, proportion dont Farrington souligne l'importance, sont secondaires dans l'analyse. Il se pourrait même qu'une proportion faible soit en fait suffisante pour démontrer la stabilité chez un auteur et le différencier éventuellement d'un autre.

Toutefois, il faut affiner cette technique quant à son utilisation dans un contexte d'identification d'auteur, notamment en ce qui concerne la discriminabilité des indicateurs.

² Les phrases doivent être repérées en comptant les points du graphique et non les unités de l'abscisse.

En effet, à présent nous pouvons seulement constater qu'il est possible de faire recours à ce type de calcul lorsqu'il s'agit de déterminer une stabilité des proportions d'un nombre limité de mots-outils dans un texte français. On ne peut décider pour le moment si c'est un moyen suffisamment fin pour différencier un auteur d'un autre. Le choix de ces éléments reste le sujet des investigations futures.

Il est également important de mentionner que vu que dans cette technique de nombreuses comparaisons sont effectuées, le risque d'augmenter l'erreur de type I est plus élevé. Nous choisissons donc un seuil de significativité beaucoup plus élevé : la valeur p pour un t' observé est multipliée par $2 \sqrt{N}$ (Bissell, 1995a).

Nous souhaitons donc poursuivre ces analyses de validation d'instrument qui peuvent ouvrir la voie à diverses applications possibles. Notamment, outre la simple recherche d'indicateurs d'« empreintes linguistiques », nous voulons étudier cette problématique d'un point de vue développemental en recherchant les moments d'apparition et d'acquisition de stabilité avec une méthode longitudinale en étudiant des textes rédigés par des enfants.

Références

- Bissell A. F. (1969). Cusum techniques for quality control. *Applied Statistics*, vol.(18) : 1-30.
- Bissell A. F. (1990). Weighted cusums : Method and applications. *Total Quality Management*, vol.(3) : 391-402.
- Bissell A. F. (1995a). *Statistical Methods for Text Analysis by Word-counts*. University of Wales, European School of Management.
- Bissell A. F. (1995b). Weighted cumulative sums for text analysis using word counts. *Journal of the Royal Statistical Society*, vol.(158) : 525-545.
- Canter D. (1992). An evaluation of the "CUSUM" stylistic analysis of confessions. *Expert Evidence*, vol.(1) : 93-99.
- Farrington J. M. (1996). *Analysing for Authorship : A Guide to the Cusum Technique*. University of Wales Press. [With contributions by A. Q. Morton, M. G. Farrington & M. D. Baker.]
- Goldsmith P. L. and Woodward R. H. (1964). *Cumulative Sum Techniques : Mathematical and Statistical Techniques for Industry*. Vol.(3). Oliver & Boyd.
- Hardcastle R. A. (1993). Forensic linguistics : An assessment of the Cusum method for the determination of authorship. *Journal of the Forensic Science Society*, vol.(33) : 95-106.
- Hardcastle R. A. (1997). Cusum : A credible method for the determination of authorship ? *Science and Justice*, vol.(37) : 129-138.
- Morton A. Q. (1978). *Literary Detection : How to Prove Authorship and Fraud in Literature and Documents*. Bowker.
- Somers H. (1998). An attempt to use weighted Cusum to identify sublanguages. In Powers D.M.W., editor, *New Methods in Language Processing and Computations in Natural Language Learning*, Proceedings of the NeMLaP3/CoNLL98 held at Macquarie University, January 11-17 : 131-139.
- Somers H. and Tweedie F. (2003). Authorship attribution and pastiche. *Computers and the Humanities*, vol.(37) : 407-429.

ANNEXE : Formules

Calcul des *qsums* simples :

$$\hat{w} = \frac{1}{n} \sum_{r=1}^n w_r$$

où \hat{w} signifie le nombre moyen de mots par phrase, w_r est le nombre de mots pour une phrase r .
Pour $i=1 \dots n$, nous avons c_i , le calcul des sommes cumulées :

$$c_i = \sum_{r=1}^i (w_r - \hat{w})$$

x_r est le nombre d'occurrence des indicateurs pour une phrase r , et :

$$\hat{x} = \frac{1}{n} \sum_{r=1}^n x_r$$

Et pour $i=1 \dots n$, nous avons h_i , le calcul des sommes cumulées des indicateurs:

$$h_i = \sum_{r=1}^i (x_r - \hat{x})$$

Calcul des *wqsums* :

$$C_i = \sum_{r=1}^i (X_r - \hat{\pi} w_r)$$

où :

$$\hat{\pi} = \frac{\sum_{i=1}^n X_i}{\sum_{i=1}^n w_i}$$

où w_i est le nombre de mots dans la phrase i , x_i correspond aux nombres d'identificateurs par phrase. Pour la séquence entière, la *qsum* devient :

$$C_n = \sum_{i=1}^n (X_i - \hat{\pi} w_i) = \sum_{i=1}^n X_i - \hat{\pi} \sum_{i=1}^n w_i$$

Le test de *qsum* de l'estimation de la variance $\sigma (w = \hat{w})$ est :

$$\hat{\sigma}^2 (w = \hat{w}) = \frac{\hat{w}}{n-1} \sum (X_i - \hat{\pi} w_i)^2 / w_i$$

puis l'estimateur (ANOVA analogue) :

$$\hat{\sigma}_A^2 = \frac{1}{n-1} \left\{ \sum \left(\frac{X_i^2}{w_i} \right) - \frac{(\sum X_i)^2}{\sum w_i} \right\}$$

et l'estimateur :

$$\hat{\sigma}_D^2 = \frac{1}{n-1} \sum_{i=1}^{n-1} \left\{ \left(\frac{X_i}{w_i} - \frac{X_{i+1}}{w_{i+1}} \right)^2 / \left(\frac{1}{w_i} + \frac{1}{w_{i+1}} \right) \right\}$$

Ainsi pour $n \geq 20$ la distribution de $v^2 = \hat{\sigma}_D^2 / \hat{\sigma}_A^2$ est pratiquement normale avec :

$$E(v^2) = 1, \quad \text{var}(v^2) \approx 1/(n+2)$$

et donc :

$$u = (v^2 - 1) \sqrt{(n+2)}$$

Comparaison de deux textes :

$$t' = \frac{\hat{\pi}_1 - \hat{\pi}_2}{\sqrt{(\hat{\sigma}_{A1}^2 / W_1) + \hat{\sigma}_{A2}^2 / W_2}}$$

où $\hat{\sigma}_{A1}^2$ et $\hat{\sigma}_{A2}^2$ sont les estimations pour $w=1$ et W_1 et W_2 sont les mots totaux dans les deux textes avec $v=n_A+n_B-2$ dl.

Calcul du V_{\max} :

Pour un segment allant de la phrase $(i+1)$ à j inclus et avec un écart maximal à la phrase r , nous avons :

$$V_{\max} = C_r - C_i - \frac{W_r - W_i}{W_j - W_i} (C_j - C_i)$$

La longueur moyenne de la phrase est :

$$\tilde{w} = (W_j - W_i) / (j - i)$$

$$V_{\max} \text{ standardisé} = V_{\max} / \hat{\sigma}$$

où :

$$\hat{\sigma} = \hat{\sigma}_{w=1} \sqrt{\tilde{w}}$$

