# Les techniques de *Q-sums* et de Multidimensional Scaling appliquées à l'étude de la stabilité de l'écriture au cours de la vie : le cas de George Sand

Judith Czellàr

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

## Abstract

Authorship attribution techniques, which rely on the psychological assumption that each person has unconscious linguistic habits as unique as fingerprints, can be used to test the stability of writing. We did show the applicability to the French of the technique of cumulative sums (*Q-sums*) whose main advantage is to be efficient with small samples (Czellàr, 2006). Our aim in this paper is to apply it to the correspondence of George Sand. In a first study we examine the stability of her writing over time by taking 35 letters written between the ages of 21 and 66. Fourteen style markers defined to operationalize linguistic fingerprint in French were used in 9'450 comparisons. The results show that letters written between 21 and 30 years are not homogeneous and differ from later letters which show similarities. To locate the break in style, ten other letters dating from this crucial decade were added (Study 2). This resulted in 3'850 new comparisons. We then analyse the proximities between letters by means of MDS: *t'*-values obtained from the weighted *Q-sums*, which were considered as distances between two texts. Our results show that the author might have had a period of transition at the beginning of her career: her linguistic habits show greater variety between 21 and 30 years and stabilize afterwards. Finally we proceed to the validation of our results by incorporating ten new letters into the model. We notice that the stylistic transition of the author is not abrupt, but a process that develops over several years. One change, between 26 and 28, is particularly conspicuous. This time period corresponds to Aurore Dupin's adoption of her literary name: George Sand. One may hypothesize that assuming a new identity between the ages of 26-27 involved a reorganization in writing that goes further than conscious expression.

## Résumé

Les techniques d'attribution de paternité littéraire, reposant sur l'hypothèse d'une identité linguistique pérenne (équivalant en cela à une empreinte digitale), peuvent de ce fait être utilisées pour tester la stabilité de l'écriture. Nous avons montré l'applicabilité au français de la technique des sommes cumulées (*Q-sums*) dont le principal avantage est d'être efficace avec des échantillons limités (Czellàr, 2006). L'objectif du présent travail est de l'appliquer à la correspondance de George Sand. Dans une première étude nous examinons la stabilité d'écriture à travers le temps chez cette auteure en prenant 35 lettres écrites entre 21 et 66 ans. Quatorze indicateurs ont été définis afin d'opérationnaliser l'empreinte linguistique, et utilisés dans 9'450 comparaisons. Les résultats montrent que le groupe des lettres écrites entre 21 et 30 ans n'est pas homogène, celles-ci se différenciant de la correspondance plus tardive où toutes portent la même empreinte. Pour localiser la rupture dans le style d'écriture nous avons introduit dix nouvelles missives datant de cette décennie cruciale (Etude 2). Ainsi 3'850 nouvelles comparaisons ont pu être effectuées. Nous examinons alors les proximités entre les lettres à l'aide du MDS : les valeurs *t'* issues des *Q-sums* pondérées étant prises comme mesure de distance entre deux textes. Nos résultats montrent que l'écrivaine aurait une période de transition stylistique au début de sa carrière : ses habitudes linguistiques sont plus variées entre 21 et 30 ans et se stabilisent par la suite. Dans un dernier temps nous procédons à la validation en incorporant dix nouvelles lettres dans le modèle. Nous constatons que la transition stylistique de l'auteure n'est pas abrupte, mais révèle une maturation de plusieurs années. Or celle-ci est plus marquée entre 26 et 28 ans, période qui coïncide avec l'adoption par Aurore Dupin de son nom de plume « George Sand ». On peut donc supposer que la nouvelle identité assumée vers 26-27 ans a impliqué une réorganisation de l'écriture qui va au-delà de l'expression consciente.

**Mots-clés :** *Q-sums*, MDS, empreinte linguistique, français, constance de style, George Sand.

# 1. Introduction

Parmi les nombreuses études sur le développement psycholinguistique seules peu d'entre elles portent sur l'évolution du langage écrit à l'âge adulte. Nous avons été amenée à nous intéresser à cette problématique en partant de la littérature sur l'attribution de paternité littéraire. Les techniques élaborées pour identifier des textes, nombreuses et variées, partent toutes d'un postulat de type psychologique qu'elles partagent en dépit de leur diversité : les textes produits par une personne donnée portent sa marque de fabrique, ils reflètent sa personnalité, et cette physionomie s'exprime indépendamment des contextes de production ou des intentions littéraires. Elle est constante dans le temps, et peut être décrite au moyen d'outils statistiques.

A partir de quand l'idiosyncrasie est-elle susceptible de se manifester dans un texte ? A partir de quand la personne qui écrit acquiert-elle des habitudes stables, des particularités dans son expression qui permettent de la distinguer des autres ? Si une certaine cohérence dans l'utilisation de la langue s'installe, n'y aurait-il pourtant pas des ruptures ou des changements au cours de la vie ? La ressemblance avec soi-même (qui permet de se différencier des autres) suppose-t-elle une écriture immuable, ou y a-t-il place pour une évolution ?

La technique des sommes cumulées (*Q-sums*) a été utilisée dans des contextes aussi bien scientifiques que légaux. Selon un de ses plus ardents promoteurs, elle permet de différencier très tôt les productions d'individus anglophones, et serait applicable même à des enfants (Farringdon, 1996, chap. 6). Czellàr (2006) a adapté cette technique pour le français. Notre but est de l'utiliser pour aborder la problématique génétique au moyen d'une étude de cas. Nous étudierons à cet effet la correspondance de George Sand.

# 2. Technique des *Q-sums* appliquée au français

Les techniques des sommes cumulées (*Q-sums*) ont été traditionnellement utilisées pour les contrôles de qualités dans les industries. Elles ont été adaptées pour des problèmes d'attribution d'auteur dans les années '70 pour des textes avant tout anglophones (Bee, 1971, 1972, Michaelson, Morton & Wake, 1978, Morton, 1978). Critiquée, essentiellement pour sa subjectivité, cette méthode a été améliorée et rendue statistiquement acceptable sous le nom de *weighted Q-sum* (*Q-sum* pondéré, *wQ-sum*) (Bissell, 1969, 1990, 1995a, 1995b, Goldsmith & Woodward, 1964). Récemment plusieurs études ont été effectuées pour tester l'applicabilité des *Q-sums* au français : elles se montrent efficaces pour mettre en évidence les différences de style de Gary-Ajar. Les résultats sont également très encourageants en ce qui concerne les problèmes d'attribution de paternité pour des textes d'origine douteuse (Czellàr 2006, Czellàr & Gillièron Paléologue, 2006). Une description détaillée de la technique se trouve dans Bissell, 1990, 1995a, 1995b.

Un des principaux avantages de la technique est qu'elle peut s'appliquer à des textes très courts (environ 25 phrases), ce qui la rend performante dans des contextes où les statistiques habituelles ne sont pas utilisables. Elle impose cependant des exigences. 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. Pour satisfaire ces exigences, les correspondances d'une personne représentent un texte idéal.

Les habitudes langagières (indicateurs) dégagées pour l'anglais ne s'appliquent pas au français. Quatorze indicateurs ont été définis à partir d'œuvres littéraires francophones[1] pour mettre en évidence l'empreinte linguistique pour le français (Tableau I).

Ce sont ces quatorze indicateurs que nous allons utiliser dans les deux études empiriques qui suivent.

| Indicateurs | | nb. | % |
|---|---|---|---|
| A | 2, 3 lettres+préposition+conjonction+pronom | 17 527 | 44% |
| B | 2, 3 lettres+pronom | 14 659 | 37% |
| C | 2, 3 lettres+préposition+conjonction | 16 604 | 41% |
| D | 1, 2 lettres+préposition+conjonction+pronom | 15 288 | 38% |
| E | 1, 2 lettres+pronoms | 12 420 | 31% |
| F | 1, 2 lettres+préposition+conjonction | 14 365 | 36% |
| G | 1, 2 lettres | 11 497 | 29% |
| H | 1, 2, 3 lettres+préposition+conjonction+pronom | 20 409 | 51% |
| I | 1, 2, 3 lettres+pronom | 17 541 | 44% |
| J | 1, 2, 3 lettres+préposition+conjonction | 19 486 | 49% |
| K | 1, 2, 3, 4 lettres | 21 157 | 53% |
| L | 1, 2, 3 lettres | 16 618 | 41% |
| M | 2, 3, 4 lettres | 18 275 | 46% |
| N | 2, 3 lettres | 13 736 | 34% |

*Tableau I. Liste des indicateurs utilisés, avec la proportion d'occurrences qu'ils représentent dans le corpus d'auteurs francophones.*

## 3. Première étude

### *3.1. Matériel*

La première édition d'importance de la correspondance de Sand (s'étendant de 1812 à 1876) a été publiée en six volumes chez Calmann Lévy entre 1882 et 1884 ; les cinq premiers volumes sont disponibles en mode texte sur le site du Project Gutenberg, et c'est de ces fichiers que nous sommes partie. Chaque lettre sélectionnée a été relue et vérifiée en comparaison avec la version imprimée, notamment en ce qui concerne les signes de ponctuation. Cette vérification n'a malheureusement pas été possible pour le Tome V, introuvable à Genève. Mais comme pour les quatre premiers volumes, nous n'avons détecté qu'un nombre infime de discordances, essentiellement dues à des confusions de caractères scannés, nous avons fait confiance à cette version électronique.

Nous avons réparti les 734 lettres en cinq groupes en fonction de l'âge de l'épistolière : de 21-30 ans, de 31-40 ans, de 41-50 ans, de 51-60 ans et de 60-66 ans. Cinq lettres de chaque groupe d'âge ont été sélectionnées aléatoirement, pour autant qu'elles répondent au critère

[1] Ce sont quarante échantillons de phrases consécutives dont la longueur varie entre 14 et 57 phrases (m = 33,6 ; é.t. = 9,18). Ils sont tirés d'Emile Ajar : (*Gros Câlin*, n = 7), Romain Gary (*Education européenne,* n = 4), Antoine de Saint-Exupéry (*Terre des Hommes*, n = 4 ; *Vol de nuit,* n = 4), Marcel Proust (*Du côté de chez Swann*, n = 4 ; *A l'ombre des jeunes filles en fleurs*, n = 6), et Guy de Maupassant (*Bel Ami*, n = 5 ; *Une vie*, n = 6). Ces chiffres sont définitifs et publiés dans Czellár, 2006, p. 68.

d'inclusion : contenir au minimum vingt-cinq phrases sans les phrases d'introduction et de conclusion. Pour faire l'analyse des *Q-sums*, les vingt-cinq phrases consécutives du milieu de chaque lettre ont été retenues (douze phrases avant et douze phrases après celle du milieu). La longueur moyenne des phrases par lettre (n = 25) variait entre 15.04 mots et 38.84 mots (m = 22.62 ; é.t. = 5.53), et l'écart-type entre 7.91 et 24.19 (m = 13.55 ; é.t. = 3.67). Tous les extraits ont été soigneusement nettoyés[2] et préparés pour l'analyse.

### *3.2. Résultats*

Les 25 lettres ont été comparées deux à deux pour les quatorze indicateurs, ce qui implique 4'200 comparaisons en tout. Nous n'avons retenu que les valeurs *t'* qui sont significativement trop élevées au seuil p' = .05, $p' = p \cdot 2n^{1/2}$ (ajustement proposé par Bissell, 1995a, p. 44, pour limiter l'erreur de type I). Le Tableau II donne pour chaque indicateur le nombre de cas où deux lettres diffèrent significativement, pour les comparaisons intra- et interpériodes. On voit que le corpus semble relativement homogène : les indicateurs ne différencient les échantillons que dans 10% des cas au maximum. Toutefois, nous pouvons attirer l'attention sur les indicateurs G, L et J avec des taux respectifs de 9%, 9% et 8%.

| | Intra-période | | | | | Interpériodes | Ensemble | |
|---|---|---|---|---|---|---|---|---|
| | I (n = 10) | II (n = 10) | III (n = 10) | IV (n = 10) | V (n = 10) | (n = 250) | (n = 300) | % |
| Indicateur A | — | — | — | — | — | 2 | 2 | 0.67 |
| B | — | — | — | — | — | 3 | 3 | 1.00 |
| C | — | — | — | 1 | — | 8 | 9 | 3.00 |
| D | — | — | — | 2 | 1 | 8 | 11 | 3.67 |
| E | — | — | — | 1 | 1 | 12 | 14 | 4.67 |
| F | — | — | — | 2 | — | 17 | 19 | 6.33 |
| G | — | — | — | 2 | — | 25 | 27 | 9.00 |
| H | — | — | — | 2 | — | 6 | 8 | 2.67 |
| I | — | — | — | 1 | — | 7 | 8 | 2.67 |
| J | — | — | — | 2 | — | 21 | 23 | 7.67 |
| K | — | — | — | 1 | — | 13 | 14 | 4.67 |
| L | — | — | — | 1 | — | 26 | 27 | 9.00 |
| M | 2 | — | — | — | — | 11 | 13 | 4.33 |
| N | — | — | — | 1 | 1 | 11 | 13 | 4.33 |

*Tableau II. Différences significatives selon le test t' pour les comparaisons deux à deux des lettres de Sand, par indicateur, pour les comparaisons intra- et interpériodes. Les signes — indiquent qu'il n'y a pas de différence significative.*

Il convient d'examiner si les différences qui apparaissent se concentrent sur certaines lettres, sur certaines périodes, ou si, au contraire elles se répartissent sur tout le corpus. Le premier cas peut faire craindre un problème au niveau des missives concernées, qui pour une raison ou une autre seraient atypiques. Le deuxième cas serait plus intéressant car il permettrait de voir une éventuelle évolution à travers le temps. A ce stade, nous ne pouvons encore nous prononcer ni sur la valeur discriminative des descripteurs, ni sur les caractéristiques spécifiques dues aux échantillons. Les comparaisons intrapériode montrent une bonne homogénéité, excepté pour la Période IV. A part les indicateurs A, B et M, on y trouve des

[2] Le nettoyage des textes concerne essentiellement la standardisation des noms propres, pour les détails voir Czellàr, 2006, pp. 69-75.

différences significatives dans près de 15% des cas. Y aurait-il une lettre très différente des autres qui contribue à cette inflation ? Dans ce cas, elle devrait également contribuer de manière exagérée aux différences interpériodes, ce qui invite à l'examiner de plus près (voir les premières lignes de chaque cellule du Tableau III). Pour chaque comparaison nous avons relevé le nombre de fois où l'on trouve des différences significatives, tous indicateurs confondus, et le/les indicateur/s qui sortent significativement dans quatre comparaisons ou plus (sur 25).

Deux principaux résultats sautent aux yeux : d'une part, le nombre de différences significatives qui concernent la Période IV, d'autre part le nombre relativement élevé de différences entre les Périodes I et V. Sur les 350 comparaisons, 34 différences sont significatives ; et à part l'indicateur D qui ne différencie jamais, tous sont discriminatifs au moins une fois. On peut donc soupçonner une évolution de l'écriture de Sand avec le temps.

Après examen des lettres de la Période IV, nous avons pu constater que les nombreuses différences interpériodes concernent essentiellement les lettres 444 et 470. Il semble alors probable que les résultats soient dus à un accident, toujours possible lorsque l'échantillon est petit (cinq textes !). Si on ne tient pas compte de ces deux lettres le paysage change considérablement :

- la Période IV montre désormais une très grande homogénéité *intra* : on ne trouve plus *aucune* différence significative, contrairement aux résultats du Tableau V qui en montrait 16 ;
- au lieu des 94 différences significatives entre les lettres de la Période IV et les autres, différences bien réparties entre tous les groupes, il n'en reste que 11, et celles-ci concernent uniquement les comparaisons avec la Période I (voir ligne 2 du Tableau III) ;
- pour ces onze différences aucun indicateur ne prime.

Nous pouvons donc raisonnablement dire que les différences trouvées qui impliquent la Période IV sont accidentelles et s'expliquent par la présence de deux lettres fortement atypiques qui compromettent les comparaisons intrapériode aussi bien qu'interpériodes.

| | II | III | IV | V |
|---|---|---|---|---|
| I | 6 / 350 1.7% — | 16 / 350 4.6% GM | 16 / 350 4.6% J<br>11 / 210 5.2% | 34 / 350 9.7% GEKLM |
| II | | 1 / 350 0.3% — | 21 / 350 6.0% JL<br>0 / 210 0.0% | 2 / 350 0.6% — |
| III | | | 26 / 350 7.4% FJL<br>1 / 210 0.5% | 7 / 350 2.0% — |
| IV | | | | 31 / 350 8.9% JKL<br>0 / 210 0.0% |

*Tableau III. Comparaison interpériodes pour les lettres de Sand. La première ligne de chaque cellule correspond aux comparaisons pour l'échantillon complet. On indique le nombre de différences significatives tous indicateurs confondus. Dans la partie droite des cellules figurent les indicateurs qui sont discriminatifs dans 4/25 cas ou plus. La deuxième ligne donne les résultats après suppression de deux lettres atypiques de la quatrième période (N°444 et N°470).*

Si l'on regarde de plus près la comparaison des Périodes I et V, il apparaît en revanche que seules deux lettres ne diffèrent jamais significativement des autres, les lettres 73 et 669. A

l'opposé, les lettres 6, 37 et 112, rédigées respectivement à 21, 26 et 30 ans, donnent systématiquement des différences avec la Période V. Pour ces trois lettres, nous trouvons 32 différences sur les 34 différences significatives.

Les différences significatives sont donc sensiblement plus nombreuses entre les deux périodes extrêmes, à savoir entre les Périodes I et V. Deux hypothèses sont possibles. Ou entre 21 et 30 ans le style de l'écrivain n'a pas encore atteint la maturité subséquente, ou les changements sont survenus après 60 ans. La première hypothèse est plus probable, car lorsqu'on ne tient pas compte de la Période I dans les comparaisons, il ne reste plus que 11 différences significatives sur 1680 comparaisons, soit 0.006%. Par ailleurs, les comparaisons de la Période I avec les autres montrent des différences qui croissent avec l'intervalle temporel qui les sépare, ce qui n'est pas le cas lorsqu'on regarde la Période V. Ceci nous amène à une première conclusion. Avec nos indicateurs, les *wQ-sums* permettent d'établir pour Georges Sand une stabilité d'écriture à travers le temps entre 31 et 66 ans. La période d'avant trente ans peut-elle être décomposée en deux sous-périodes, où la seconde marquerait la stabilisation du style de l'écrivain ? Peut-on localiser le moment où s'opère un changement dans « l'expression inconsciente » que mesurent les indicateurs ? C'est ce que nous voudrions établir dans la deuxième étude.

## 4. Deuxième étude

Si un changement s'était produit chez Georges Sand entre 21 et 30 ans, il devrait se traduire à la fois par un manque d'homogénéité à l'intérieur de cette tranche d'âge, et par un comportement différentiel dans les comparaisons interpériodes : les lettres écrites lorsqu'elle était plus jeune devraient se différencier des lettres ultérieures, au contraire de celles qui datent de la fin de l'épisode. Il devrait être possible de localiser une solution de continuité, pour autant que le nombre d'observations le permette. Nous avons donc constitué un échantillon additionnel pour pouvoir « zoomer » sur cette Période I.

### *4.1. Matériel*

Nous avons sélectionné de manière aléatoire dix nouvelles lettres écrites par Sand entre 21 et 30 ans, issues de la même édition que celles de la première étude. Les vingt-cinq phrases consécutives du milieu de chaque lettre ont été retenues et « nettoyées ». La longueur moyennes des phrases par lettre varie entre 14.64 mots et 26.96 mots (m = 17.86 ; é.t. = 4.04), avec un écart-type qui varie entre 6.95 mots et 16.08 mots (m = 10.96 ; é.t. = 3.96). Ces dix lettres ont des phrases plus courtes que les autres de la même période, mais lorsqu'on regarde l'ensemble des lettres, on ne trouve pas de corrélation entre l'âge et la longueur des phrases ($r$ = - .030, NS).

### *4.2. Procédure*

Nous avons inclus les nouvelles lettres dans l'échantillon de la Période I, qui comprend désormais quinze individus. Toutes les comparaisons *intra* et *inter* ont été effectuées, soit 1'470 comparaisons intrapériode et 3'780 comparaisons interpériodes, tous indicateurs confondus.

### *4.3. Résultats*

Le Tableau IV présente les fréquences absolues et relatives des différences significatives intra- (Période I) et interpériodes (Période I avec les Périodes II, III, IV et V). Dans la partie droite des cellules figurent les indicateurs les plus discriminatifs.

On constate que le nombre de différences significatives reste modeste, ce qui n'est pas étonnant vu le critère très conservateur adopté. Mais de manière plus intéressante, on peut remarquer que ces différences sont plus nombreuses dans les comparaisons *internes* à la Période I (5,44% *vs* 4,34% en inter). Ceci nous invite à tirer parti de toute l'information donnée par les valeurs $t'$. En effet, la statistique $t'$ constitue une véritable mesure de la différence dans la proportion d'un indicateur au sein des deux textes comparés (le signe indiquant l'excès ou le défaut de l'indicateur en question pour l'échantillon auquel le deuxième est comparé). Autrement dit, $|t'|$ peut être vu comme une mesure de distance ou de dissimilarité.

Nous nous intéressons à l'homogénéité de la Période I, avec l'hypothèse d'une fracture : deux sous-groupes devraient se dessiner, l'un, de lettres semblables à celles des autres périodes, l'autre, sur lequel se concentreraient les différences. Par conséquent, nous avons soumis les matrices des valeurs $|t'|$ des comparaisons *intra* (14 matrices de 105 comparaisons, une par indicateur) à une analyse multidimensionnelle (MDS, ALSCAL) ainsi qu'à une analyse hiérarchique (*clusters*, distance euclidienne, *average linkage*). Pour les indicateurs qui discriminent le plus souvent (L, G, J, N, C) les solutions du MDS en deux dimensions sont très satisfaisantes (stress toujours <0.04, RSQ toujours >0.99). Certaines lettres se regroupent de manière consistante, formant deux noyaux stables bien différenciés. Les indicateurs G, J et L, les plus discriminatifs en termes du nombre de différences significatives (Tableau V) permettent clairement d'opposer d'une part, les lettres 6, 24, 37, 40, 48 et 99, d'autre part les lettres 7, 27, 56, 73, 81, 88, 103, 112 et 118. Les numéros indiquant l'ordre chronologique, on voit que le premier groupe comprend principalement la correspondance de Sand dans une première phase, et le deuxième celle d'une Sand qui approche la trentaine (Wilcoxon-Mann-Whitney de 34, $p<.05$ selon Kanji, 1993, p. 191). Pour les autres indicateurs, il se peut en revanche qu'aucune limite claire ne puisse être tracée, ou alors que certaines lettres changent de groupe. Quatre lettres semblent ainsi avoir un comportement versatile : les lettres 7, 48, 103 et 112 ; sans constituer un type cohérent, elles peuvent être rassemblées du fait de ce caractère instable. Nous allons donc distinguer trois groupes : le Groupe Ia, comptant cinq lettres qui se ressemblent et qui devraient s'opposer, non seulement aux autres de la même période, mais également à celles des Périodes II, III, IV et V. Le Groupe Ib, de quatre lettres « capricieuses », qui tantôt ressemblent aux premières, tantôt à celles du groupe suivant, et que seul ce caractère variable permet de placer sous la même étiquette. Le Groupe Ic, de six lettres, cohérent et bien délimité. Dans l'hypothèse d'une stabilisation de l'écriture, ce troisième groupe ne devrait pas se différencier des lettres ultérieures, et c'est donc ici que les comparaisons interpériodes deviennent particulièrement informantes. Le Tableau V en résume l'essentiel.

Les trois groupes présentent une bonne homogénéité *intra*, ce qui va de soi puisqu'ils ont été formés à partir des dissimilarités que constituent les mesures $|t'|$. Une seule différence ressort sur l'ensemble des comparaisons, dans le Groupe Ib. De manière cohérente, et en accord avec

notre hypothèse, nous observons que dans les comparaisons avec les Périodes II à V, le Groupe Ia se démarque clairement des autres. Pour ces derniers, seule une différence importante semble émerger, entre le Groupe Ib et la Période V avec l'indicateur E. Pour les lettres du Groupe Ia, le contraste est saisissant. Tout comme dans la première étude, le nombre de différences significatives augmente avec l'écart d'âge, et l'on observe 5, 12, 10 et 15% de différences significatives. Non seulement les cas sont plus fréquents, mais les indicateurs qui permettent de les mettre en évidence deviennent plus variés. Huit indicateurs sont discriminatifs dans plus de 17% des comparaisons, lorsqu'il s'agit de la Période V. Or, cet ordre de grandeur correspond aux comparaisons internes à la Période I, lorsqu'on oppose le Groupe Ia au Groupe Ic. Qualitativement, les cinq indicateurs discriminatifs entre Ia et Ic (L, G, J, N et C) sont également discriminatifs entre Ia et la Période V. Les lettres du Groupe Ic, pourtant elles aussi écrites entre 21 et 30 ans, ressemblent davantage à celles qui sont écrites après 60 ans qu'avec celles du Groupe Ia, majoritaire avant 26 ans.

*Tableau IV. Nombre de différences significatives pour les comparaisons de la Période I, intra (n = 105) et inter (n = 270), pour chaque indicateur.*

| Comparaisons | intrapériode | | interpériodes | |
|---|---|---|---|---|
| Indicateur A | 2 | 1.9% | 1 | 0.4% |
| B | 2 | 1.9% | 2 | 0.7% |
| C | 7 | 6.7% | 9 | 3.3% |
| D | 0 | 0.0% | 1 | 0.4% |
| E | 5 | 4.8% | 9 | 3.3% |
| F | 4 | 3.8% | 18 | 6.7% |
| G | 14 | 13.3% | 23 | 8.5% |
| H | 0 | 0.0% | 3 | 1.1% |
| I | 0 | 0.0% | 4 | 1.5% |
| J | 9 | 8.6% | 23 | 8.5% |
| K | 5 | 4.8% | 11 | 4.1% |
| L | 19 | 18.1% | 35 | 13.0% |
| M | 6 | 5.7% | 13 | 4.8% |
| N | 7 | 6.7% | 12 | 4.4% |

*Tableau V. Comparaisons intra et inter pour les lettres de la Période I, regroupées en fonction des analyses de proximité (MDS et clusters). On indique le nombre de différences significatives tous indicateurs confondus. Dans la partie droite des cellules figurent les indicateurs qui sont discriminatifs dans 17% des cas ou plus.*

| | Période I | | | Périodes | | | |
|---|---|---|---|---|---|---|---|
| | Ia | Ib | Ic | II | III | IV | V |
| Ia | 0 / 140 — 0.0% | 4 / 280 — 1.4% | 69 / 420 LGJ NC 16.4% | 16 / 350 L 4.6% | 43 / 350 LJF GM 12.3% | 21 / 210 LJF 10.0% | 53 / 350 LGJF NCKM 15.1% |
| Ib | | 1 / 84 — 2.4% | 6 / 336 — 1.8% | 0 / 280 — 0.0% | 5 / 280 — 1.8% | 2 / 268 — 1.2% | 17 / 280 E 6.1% |
| Ic | | | 0 / 210 — 0.0% | 2 / 420 — 0.5% | 1 / 420 — 0.2% | 0 / 252 — 0.0% | 4 / 420 — 1.0% |

Du point de vue chronologique, on constate donc qu'il est difficile de définir une coupure nette à partir de laquelle s'opèrerait ce changement si sensible. S'il fallait absolument se prononcer, on serait certes tenté de le placer entre 26 et 27 ans, entre la lettre 48 et la lettre 56. Mais on doit se rappeler que la lettre 99, écrite à 29 ans, se trouve dans le Groupe Ia. Les lettres 7 et 48 (Sand avait alors 22, respectivement 26 ans) ne peuvent être rattachées de manière indiscutable ni au groupe « précoce » ni au groupe « tardif », pas plus que les lettres 103 et 112 (29 et 30 ans).

Si donc on peut avec certitude affirmer que des changements sont intervenus dans l'écriture de George Sand entre 21 et 30 ans, on se doit aussi de préciser qu'ils ne se font pas d'un coup. Des lettres plus tardives ressemblent aux lettres plus précoces, et l'inverse est également vrai. La transition n'est pas abrupte, et le processus se déroule sur plusieurs années, avec une évolution plus marquée entre 26 et 28 ans, qui se manifeste en particulier avec un indicateur discriminatif pour l'auteure, l'utilisation des mots de une, deux ou trois lettres (indicateur L).

Comment expliquer cette évolution, très tardive si l'on se réfère aux « habitudes de langage » discutées par Farringdon (1996) ? On pourrait avancer plusieurs hypothèses liées à l'exercice

de l'écriture. C'est en effet le moment où Sand se met à publier, d'abord des articles, puis des ouvrages. Une autre supposition n'est pas à négliger. En 1931 la baronne Dudevant adopte le pseudonyme de George Sand ; elle a alors 27 ans.

En étudiant les phénomènes liés au changement de nom, Lapierre (2006) souligne que le pseudonyme permet à l'écrivain non seulement d'avoir un nom de plume et d'être le créateur de ses personnages, mais aussi de devenir créateur de lui-même, s'engageant ainsi dans un processus d'autoengendrement où l'auteur est à la fois père (mère) et fils (fille) de son œuvre. Le nom met une empreinte sur le devenir de chacun. Nous pouvons donc émettre l'hypothèse que le pseudonyme de Sand a en quelque sorte facilité l'évolution du style de l'écrivaine.

En repérant les signatures finales des lettres recueillies dans l'édition Calmann Lévy, nous constatons que la majorité des lettres de l'écrivaine ne sont pas signées. Jusqu'en 1832, nous trouvons des signatures sous son nom de baptême, Aurore, ou comprenant ses prénom et nom de mariage (Aurore Dudevant ou Aurore D.), ou encore mentionnant un lien d'affiliation s'il s'agit de lettres écrites à sa mère (« votre fille Aurore », lettre 20) ou à son fils (« ta maman », lettre 51). Ce n'est qu'à partir de 1832 que nous trouvons des lettres signées par son pseudonyme (ex : « ton ami GEORGE », lettre 89 ; ou simplement « G.S. », lettre 105) mais en alternant encore avec son nom civil qui petit à petit va disparaître au profit du premier. A partir de 1834, les lettres qui sont signées le sont uniquement du pseudonyme.

Ainsi, il est possible de supposer que l'adoption d'un nouveau nom aux environs de 26-27 ans entraîne une réorganisation dans l'écriture qui va plus loin que l'expression consciente. On voit donc une atteinte des caractéristiques profondes du « *fingerprint* » linguistique.

## 5. Validation

Ces résultats nous amènent à penser que la transition stylistique de l'auteure n'est pas abrupte bien que plus marqué entre 26 et 28 ans. Si nous prenons les trois classes délimitées plus haut comme « modèle » de l'évolution de l'écriture de Sand, il serait possible d'y insérer de nouveaux individus (lettres). Si notre hypothèse se vérifie, ces nouveaux éléments devraient être dispersés dans l'espace du MDS et ne pas se concentrer dans une région. Pour « valider » nos résultats nous avons donc procédé à l'échantillonnage de nouvelles lettres de la période décisive, à savoir entre 25 et 30 ans en nous focalisant sur l'indicateur le plus caractéristique de l'auteure : l'indicateur L.

### *5.1. Matériel*

Dix lettres supplémentaires écrites entre 25 et 30 ans ont été échantillonnées selon les critères cités ci-dessus. La longueur moyenne des phrases par lettres varie entre 11.84 et 22.40 mots (moyenne=16.38, é.t.=3.10) et un écart-type qui varie entre 6.40 et 13.12 mots (moyenne=9.53, é.t.=2.34).

### *5.2. Procédure*

Avec les dix nouvelles lettres toutes les comparaisons inter et intra ont été effectuées avec le *wQ-sum*, à savoir 375 comparaisons deux à deux pour l'indicateur L. Il y a une bonne homogénéité intra : aucune différence significative n'est observée. Suite à cette étape nous avons conduit un MDS et une analyse hiérarchique sur les distances $|t'|$ de toutes les comparaisons deux à deux. Ces résultats sont représentés sur la Figure 1. Le modèle en deux

dimension est convenable (Stress=.04008, RSQ =.99449). Les lettres du nouvel échantillon sont soulignées sur le graphique.

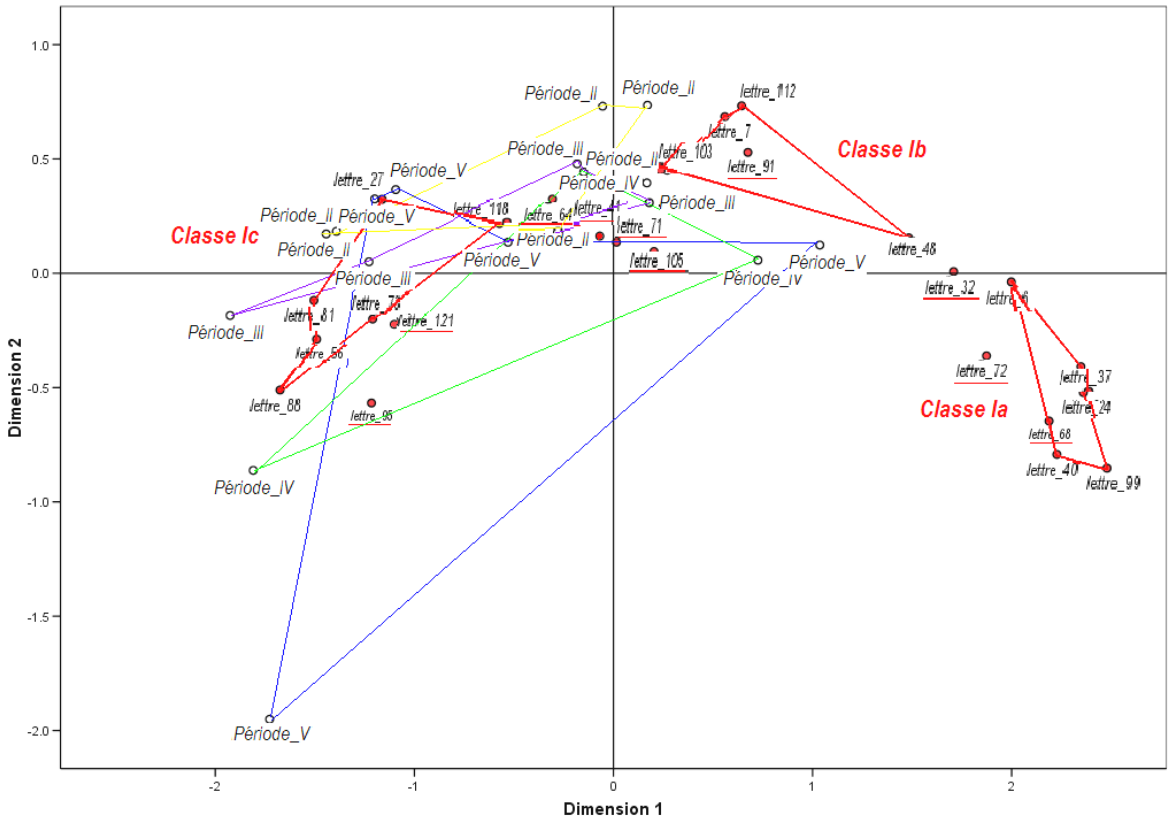


*Figure 1. MDS de 43 lettres de G.S. pour l'indicateur L. Les lettres d'une période sont reliées par des traits. Les classes Ia-Ic sont en rouge et les nouvelles lettres sont soulignées.*

### *5.3. Résultats*

Nous constatons que les lettres se distribuent en forme de U. Il est également possible de remarquer une tendance chronologique qui va de la droite vers la gauche. Il est clair que les distributions se recoupent et qu'il existe une grande dispersion, surtout des Périodes IV et V.

En ce qui concerne la Période I les trois sous-groupes ressortent, bien que l'analyse inclue les dix lettres supplémentaires. La classe Ia (lettres 6, 24, 37, 40 et 99) est la plus homogène, et se trouve à l'extrémité droite du graphique : ce sont les lettres portant l'empreinte de « jeunesse ». A elles se joignent des nouvelles lettres : la 32, la 68 et la 72. Ce groupe forme également un grand cluster sur l'analyse hiérarchique lorsqu'on coupe l'arbre après le troisième nœud. C'est un groupe « pur », aucune autre lettre des périodes ultérieures ne se joint à lui.

A l'autre extrémité, à gauche on retrouve la Période Ic (lettres 27, 56, 73, 81, 88 et 118) qui est entremêlée avec des lettres de toutes les périodes mais surtout celles plus tardives, les périodes IV et V. Ces écrits ressemblent davantage au style que l'auteure a adopté définitivement. L'analyse en cluster nous aide également à y classer deux nouvelles lettres : les lettres 121 et 95.

En ce qui concerne les échantillons de la classe Ib (7, 48, 103 et 112) ils se trouvent au centre du graphique, mais plutôt proches du groupe Ia. A ce groupe appartiennent quatre nouveaux textes : les lettres 41, 64, 71 et 105.

On constate donc que les nouveaux échantillons utilisés pour valider notre modèle se distribuent le long du graphique. En effet les lettres de la première décennie que nous avons appelée la Période I ont la plus grande dispersion parmi toutes les autres. Le sous-groupe Ia qui se différencie clairement du style subséquent de l'auteure (Période Ia). Ces lettres forment un sous-groupe homogène et se distinguent clairement des autres. Ce sont des lettres qui portent le style « jeune » d'Aurore Dupin. Par la suite, petit à petit, le style George Sand se dessine et laisse son empreinte sur les écrits. Comme présupposé, ce phénomène est un processus long. Il est intéressant de noter que parmi les lettres portant une signature, celles qui sont signées par le pseudonyme se retrouvent toutes dans le même cluster de la Période Ib, appelé « labile » (lettres 91 « George », 103 « Ton ami GEORGE SAND », 105 « G.S » et 112 « GS »).

## 6. Conclusion

La *wQ-sum* permet de mettre en évidence l'homogénéité d'un texte, et plus précisément de décider s'il est écrit par une ou plusieurs personnes. Nous ne nous sommes pas placée du point de vue de l'attribution de textes, mais plutôt de celui de la mesure de la stabilité des habitudes linguistiques. Notre choix s'est porté sur George Sand. Même si notre étude ne répond pas à la question du tout début des habitudes linguistiques dans le langage parlé, elle constitue un premier pas dans la direction des études génétiques. Nous avons procédé en trois temps.

Premièrement, en sélectionnant dans sa correspondance cinq lettres par périodes de dix ans entre 21 et 66 ans, nous avons effectué nos comparaisons avec quatorze indicateurs définis préalablement à partir d'un corpus de textes d'auteurs francophones. Les résultats obtenus montrent que l'hypothèse de stabilité et de cohérence intra-auteur est défendable. Nous avons cependant trouvé quelques différences entre la première et la dernière période. C'est sur ces différences que nous nous sommes focalisée. Dans un deuxième temps, afin de localiser ces changements, nous avons effectué des comparaisons avec dix nouvelles lettres de la Période I. Nous avons ainsi pu définir trois sous-groupes au sein de cette période en fonction de l'évolution stylistique de l'auteure. Dans un troisième temps ces résultats ont été validés en incorporant dix nouvelles lettres dans ce modèle. Les résultats suggèrent une phase de transition entre 26 et 27 ans, correspondant à celle où l'écrivaine a adopté son nom de plume.

Il s'agit certes d'une personne dont le métier est d'écrire, mais il s'agit également de lettres, et les correspondances constituent un matériel de choix pour tous ceux qui s'intéressent à l'identification de sources. Etant donné que Sand n'a pas été comparée à d'autres auteurs, ni même à elle-même avec d'autres types d'écrits, il n'était toutefois pas possible de nous prononcer sur les indicateurs qui la caractériseraient le mieux.

## Références

Bee R.E. (1971). A statistical study of the Sinai Pericope. *Journal of the Royal Statistical Society Series A,* 135, 406-421.

Bee R.E. (1972). Statistical methods in the study of the Masoretic Text of the Old Testament. *Journal of the Royal Statistical Society Series A,* 134, 611-622.

Bissell A. F. (1969). Cusum techniques for quality control. *Applied Statistics*, 18, 1-30.

Bissell A. F. (1990). Weighted cusums: Method and applications. *Total Quality Management*, 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*, 158, 525-545.

Czellàr J. (2006). *L'empreinte linguistique mesurée avec la technique des sommes cumulées (Q-sum) : Etude de validation en français*. Thèse de doctorat, Université de Genève.

Czellàr J. and Gillièron Paléologue Ch. (2006). Constance ou évolution dans l'écriture: l'exemple de George Sand analysé au moyen d'une technique d'identification linguistique. *Archives de Psychologie*, 72 (282-283), 211-242.

Farringdon J. M. (1996). *Analysing for authorship: A guide to the Cusum technique*. Cardiff: University of Wales Press. [With contributions by A. Q. Morton, M. G. Farringdon and M. D. Baker].

Goldsmith P. L. and Woodward R. H. (1964). *Cumulative sum techniques: Mathematical and statistical techniques for industry, Vol. 3*. Edinburgh: Oliver and Boyd.

Kanji G.K. (1993). *100 statistical tests*. London/Newbury Park: Sage.

Lapierre N. (2006). *Changer de nom*. Paris : Gallimard.

Michaelson S., Morton A.Q. and Wake W.C. (1978). Sentence length in Homer and hexameter verse. *Association for Literary and Linguistic Computing Bulletin, 6,* 254-267.

Morton A. Q. (1978). *Literary detection: How to prove authorship and fraud in literature and documents*. Epping: Bowker.

Sand G. (1883-1884). *Correspondance 1812-1876. [6 vols.]*. Paris : Calmann Lévy. (Les cinq premiers volumes sont disponibles à l'adresse http://www.gutenberg.org/).