Segmentation automatique des corpus
Voyages de l'autre côté de J.-M. Le Clézio*

Pierre Hubert[1], Cyril Labbé [2], Dominique Labbé[3]

1 Ecole des Mines de Paris – 35 rue Saint Honoré – F 77305 Fontainebleau – France – iahs@ensmp.fr

2 Université Grenoble I – France – Cyril.labbe@imag.fr

3 Université Grenoble II – France – dominique.labbe@iep.upmf-grenoble.fr

Abstract

We present an original segmentation method applied to textual data series. The vocabulary growth and variations of its diversity are calculated. Then a segmentation algorithm, associated with a validity test, gives the optimal successive stages. This method is applied to a novel by Jean-Marie Le Clezio : Voyages de l'autre côté.

Résumé

Méthode originale pour segmenter un corpus en sous-parties homogènes. On calcule l'accroissement du vocabulaire et les variations de sa diversité. Un algorithme de segmentation associé à un test de validité donne le découpage optimal des deux séries. Application à un roman de Jean-Marie Le Clezio : Voyages de l'autre côté.

Mots-clés : Segmentation - œuvre littéraire - diversité du vocabulaire- test de validité - Le Clezio

1. Introduction

Le découpage des vastes corpus demeure une opération empirique. Le choix des césures reste généralement intuitif : dates charnières de la vie de l'auteur, genres littéraires, oeuvres que l'on estime capitales… Sans doute, faute de procédure automatique, une longue familiarité avec le corpus peut-elle justifier l'opération. Il n'en reste pas moins que les découpages sont effectués à partir de critères largement extérieurs aux textes. Dès lors, qui peut garantir que les résultats des traitements lexicométriques effectués sur cette partition ne sont pas influencés par les ciseaux de l'analyste ?

Nous voudrions proposer ici une méthode de segmentation automatique des corpus, méthode qui repose sur l'observation de l'accroissement du vocabulaire et des variations de sa diversité. Après avoir rappelé la manière dont sont calculées les valeurs de la variable observée, nous

* Les auteurs remercient les relecteurs anonymes pour leurs remarques qui ont permis d'améliorer sensiblement ce texte, Margareta Kastberg qui a mis à notre disposition ses fichiers et sa connaissance de l'œuvre de JM Le Clezio et le prof. Jan de Leeuw, de l'UCLA Department of Statistics, pour son aide dans la programmation de nos tests.

présentons l'algorithme de segmentation. A titre d'illustration, nous utiliserons une œuvre de Jean-Marie Le Clezio qui a été mise à notre disposition par M. Kastberg (Kastberg-Brunet, 2000 et Kastberg, 2002). Les graphies en ont été normalisées puis lemmatisées (Labbé, 1990). Les calculs présentés ci-dessous ont été effectués sur ces données lemmatisées (le roman comporte au total 98 862 mots dont 8 809 formes graphiques normalisées différentes et 5 372 vocables).

2. La diversité du vocabulaire

La diversité du vocabulaire est l'une des quatre notions issues de la *richesse du vocabulaire* (Hubert-Labbé, 1994, Alvarez et Al, 2000). Dans un texte donné, elle mesure l'étendue du *lexique* mobilisé pour la réalisation de ce texte. Une faible diversité du vocabulaire sera l'indice d'un texte pauvre stylistiquement parlant (comme c'est le cas pour les textes à dominante scientifique ou pédagogique). A l'inverse, un indice élevé signale généralement une meilleure élaboration et une visée polémique ou poétique. Naturellement, la diversité du vocabulaire est plus élevée à l'écrit qu'à l'oral, c'est même un indice certain qui permet de reconnaître les paroles improvisées des déclarations préparées (voir l'analyse des interventions télévisées de F. Mitterrand et de C. de Gaulle dans Labbé 1990b et Labbé 1998).

Pour neutraliser l'effet de la *longueur* et utiliser un étalon commun entre les textes comparés, on a eu l'idée de prélever aléatoirement dans ces textes des extraits de même longueur où l'on comptait le nombre de *vocables* différents (Cossette, 1994).

La taille de ces tranches de mots peut varier suivant la nature des textes étudiés : pour un texte oral, il faut choisir une taille adaptée à celle de la mémoire immédiate du locuteur mais l'augmenter nettement dans le cas d'une œuvre littéraire comme le roman de J.-M. Le Clézio. Une taille de 1.000 mots semble s'être imposée dans ce cas essentiellement pour des raisons de normalisation des résultats et de commodité dans l'interprétation.

On en tire la définition de l'*indice* de diversité du vocabulaire d'un texte littéraire :

Soit un texte de *N* mots (N > 1000), la diversité du vocabulaire sera le nombre moyen de vocables différents observés dans les (N-1) tranches de 1000 mots contigus que l'on peut extraire du corpus.

Ce nombre est calculé de la manière suivante.

Soit un corpus, de taille N mots, comportant V vocables différents dont V_i de fréquence i (i variant de 1 à f) Pour un fragment de taille N' ($N'<N$), l'espérance mathématique du nombre de vocables différents contenus dans le fragment est, selon la formule de Muller (Muller 1977 et 1979) :

$$VCM(u) = V - \sum_{1}^{f} V_i Q_i(u) \quad avec \ u = \frac{N'}{N} \quad et \ Q_i(u) = (1 - u)^i$$

et selon le modèle de partition (Hubert-Labbé, 1988) :

$$VPA(u) = p.u.V + q \sum_{1}^{f} V_i Q_i(u) \quad avec \ q = 1 - p$$

p étant le paramètre de partition (partition du vocabulaire en deux ensembles : vocabulaire général et vocabulaires spécialisés). Pour estimer les valeurs de ces différentes variables, on se déplace le long du texte en observant, à espace régulier, le nombre de vocables différents

(V'^*_k) apparus depuis le début du texte. Le paramètre de partition est calculé en minimisant les écarts entre les valeurs observées et les valeurs calculées grâce à la formule de Muller.

Ce calcul indique que, dans *Voyages de l'autre côt*é, on peut s'attendre à trouver, en moyenne, 335 vocables différents dans n'importe quelle tranche de 1.000 mots contigus. Au passage, on notera qu'il s'agit d'une diversité faible, plus proche de l'oral que de l'écrit. En effet, habituellement, dans les œuvres littéraires, la diversité du vocabulaire tourne autour de 400 vocables différents pour 1000 mots. C'est d'ailleurs une caractéristique constante dans toute l'œuvre de cet auteur que cette sobriété du style qui frise la pauvreté de l'expression (sur l'ensemble de l'œuvre, l'indice de la diversité est de 355 ‰).

On associe à cette valeur, un écart-type, racine carrée de la variance :

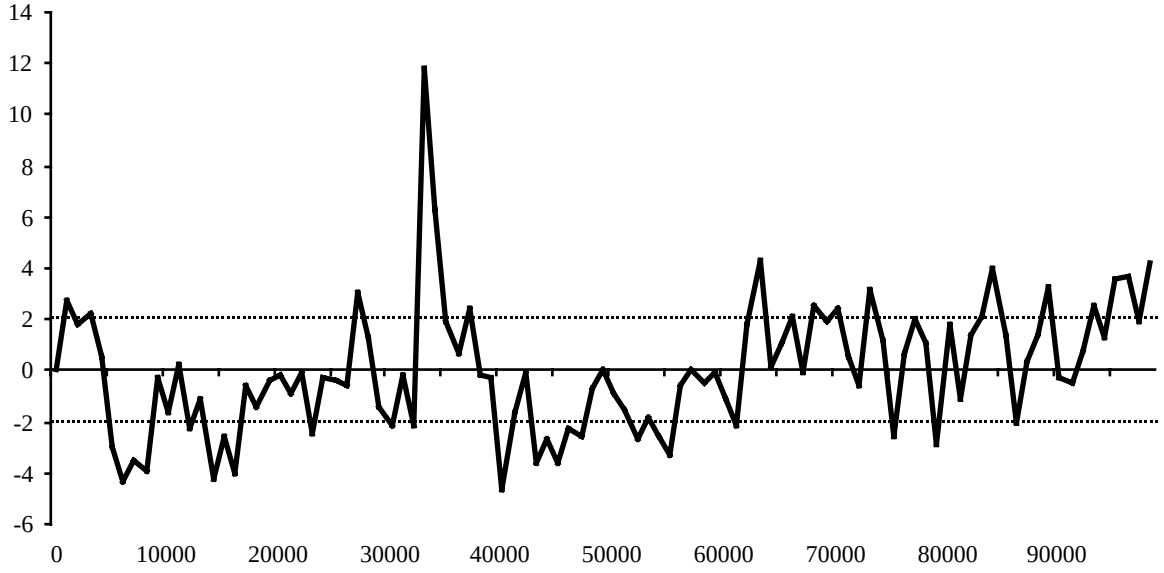
$$Var(V'u) = \sum_{1}^{f} V_i.Q_i(u) * (1 - Q_i(u)) \quad avec \quad u = \frac{1000}{N}$$

Une fois connus la diversité moyenne et l'écart type, l'algorithme se déplace au long du texte, selon un pas choisi par l'opérateur (au maximum égal à la dimension choisie pour mesurer la diversité), et compte le nombre de vocables différents dans les 1000 derniers mots qu'il vient de lire (procédé dit de la "fenêtre glissante"). Ces valeurs observées (V'*) sont centrées sur la valeur théorique calculée (V') et réduites. Sur les graphiques, la diversité moyenne du vocabulaire est figurée par l'axe des abscisses. Pour le *k*ième fragment, l'ordonnée sera :

$$Y_k = \frac{V'_{*k} - V'_{(1000)}}{\sqrt{Var(V'_{(1000)})}}$$

Le graphique ci-dessous représente les résultats obtenus sur le roman de J.-M. Le Clezio découpé en tranches égales de 1.000 mots.

Graphique 1. Diversité du vocabulaire dans Voyages de l'autre côté
(variable centrée et réduite)



L'axe des abscisses représente la tendance théorique et la courbe grasse les valeurs observées (centrées sur la moyenne et réduites). Si l'accroissement du vocabulaire avait été régulier tout au long de l'œuvre, les valeurs observées auraient été, sinon confondues avec l'axe, du moins comprises dans un intervalle de fluctuation "normal", ici symbolisé par les deux traits pointillés situés à ± 2 écarts-types.

On peut donc examiner la position de chaque point (diversité du vocabulaire au point considéré) ou l'orientation de segments plus ou moins longs : une portion de courbe, dont la pente est supérieure à l'unité, signale un passage plus divers, donc un changement stylistique certain et un renouvellement thématique probable. Ainsi on peut noter un pic à la fin du premier tiers du livre qui correspond exactement à un afflux brutal de noms propres (ce passage particulier est cité dans la communication de M. Kastberg). A l'inverse, une portion de courbe orientée vers le bas signale un appauvrissement de l'expression. Enfin, le grand nombre de points, situés hors de l'*intervalle de fluctuation normale* signale un texte hétérogène qui peut être légitimement découpé en plusieurs parties. De ce point de vue, le meilleur *découpage* sera celui qui permettra de faire entrer, dans l'*intervalle de fluctuation normale*, le maximum de points des courbes centrées et réduites calculées sur les sous-corpus découpés comme nous l'expliquons plus bas.

3. Procédure de segmentation et test de validité

On remarquera d'abord que la brutalité des évolutions retracées dans le graphique 1 et le grand nombre de points situés hors de l'intervalle, où devraient s'inscrire des fluctuations "normales", laissent supposer l'existence de ruptures difficiles à intégrer dans un modèle stationnaire. Cette caractéristique nous a suggéré le recours à des procédures développées pour le traitement de séries discontinues notamment en économie ou dans l'étude du climat (Hubert et al 1989).

Considérons la série des n valeurs V_i de la diversité du vocabulaire découpée en m segments. Nous notons, v_k la moyenne du $k$ième segment et d_k sa variance :

$$d_k = \sum_{i_{k-1}+1}^{i_k} (v_i - \overline{v_k})^2$$

Si l'on découpe la série en m segments, D_m, la somme des d_k — pour k variant de 1 à m — sera l'écart quadratique entre la série et la segmentation considérée. Cet écart permet d'apprécier la proximité entre la série et la segmentation qui lui est appliquée. Il varie entre 0 (pour $m = n$) et σ^2, variance de la série initiale, pour $m = 1$.

Nous recherchons un algorithme capable de déterminer, pour toute segmentation comprise entre 1 et n, un découpage optimal de la série. Naturellement, il est hors de question de rechercher cette segmentation par examen exhaustif de toutes les solutions possibles puisque leur nombre croît exponentiellement avec celui des données initiales et du nombre des segments considérés. Il est donc nécessaire d'une part de définir un algorithme d'optimisation efficace et, d'autre part, de lui adjoindre des tests de validité permettant de ne retenir que les segmentations pertinentes.

Une segmentation d'ordre m d'une série de longueur n peut être représentée comme les rameaux d'une arborescence. Pour éviter l'explosion combinatoire de celle-ci, on utilise un algorithme d'optimisation de type "branch and bound" (Minoux, 1983) abrégé chaque fois que cela est possible. La procédure peut être résumée ainsi. On recherche d'abord pour les ordres

inférieurs à *m*, en débutant par l'ordre 1, les segmentations optimales des sous-séries obtenues, à partir de la série initiale, en considérant successivement les 1,2,…, *n* derniers termes. On dispose ainsi des valeurs de *D*, pour un niveau donné de l'arborescence, ce qui permet d'abréger l'exploration des rameaux au niveau inférieur : l'écart relatif à la meilleure segmentation déjà obtenue est comparé à l'écart relatif à la meilleure segmentation totale incluant la segmentation partielle en cours. Si la seconde est supérieure à la première, il est évidemment inutile de poursuivre plus profondément l'exploration du rameau correspondant dans l'arborescence.

Pour un nombre donné de segments, le meilleur découpage est celui qui minimise D. On trouvera en annexe à ce texte, les résultats obtenus sur le roman de Le Clezio. Si l'on cherche simplement à constituer un certain nombre de sous-parties dans le corpus, on peut se satisfaire de ce résultat en se plaçant au niveau de finesse souhaité.

En revanche, si l'on recherche les coupures les plus significatives, il ne faudra accepter une segmentation que si les moyennes de deux segments contigus sont significativement différentes. Ceci pourra être testé grâce à la notion de contraste introduite par Scheffé (1959) dont on peut trouver un exposé simplifié dans Dagnelie (1970).

Soit une segmentation d'ordre *m*, obtenue par l'algorithme qui vient d'être décrit. Pour être considéré comme non significatif, le contraste φ_k entre la moyenne du segment *k* et celle du segment suivant :

$$\varphi_k = \overline{V_k} - \overline{V_{k+1}}$$

doit vérifier, avec une probabilité 1 - α , l'inégalité :

$$\varphi_k - S\sigma < 0 < \varphi_k + S\sigma$$

avec : $\sigma^2 = \dfrac{D_m}{n - m}$

n_k n_{k+1} les tailles respectives du *k*ième segment et de celui qui le suit,

et : $S^2 = (m - 1)F_{m-1,n-m}(\alpha) \sum_{1}^{m} \dfrac{\varphi_k^2}{n_k + n_{k+1}}$

où $F_{m-1,n-m}(\alpha)$ est la valeur de la variable de Fischer à *m-1* et *n-m* degrés de liberté.

Si les valeurs de φ_k - Sσ et de φ_k + Sσ sont respectivement négative et positive, alors on pourra considérer que les valeurs des segments *k* et *k+1* appartiennent à une série stationnaire, ou encore qu'il ne se produit pas de rupture significative en passant du segment *k* au segment *k+1*.

Une segmentation d'ordre *m* sera réputée acceptable si tous les contrastes entre les moyennes des segments voisins sont différents de zéro au niveau de signification α.

Au cours de l'exploration de l'arborescence des segmentations, un segment dont l'écart quadratrique à la série est inférieur au plus faible écart déjà obtenu n'est retenu que si l'hypothèse nulle du test de Sheffé est rejetée. Ce test permet donc d'interrompre la segmentation au niveau optimal.

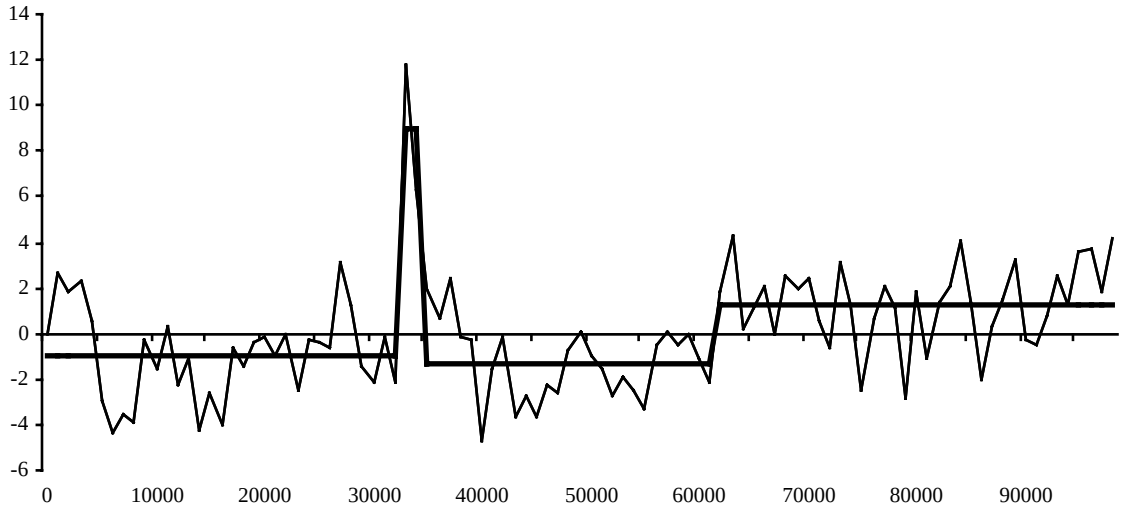
La procédure de segmentation que nous venons de présenter peut être regardée comme un test de stationnarité, l'affirmation selon laquelle "la série est stationnaire" constituant l'hypothèse nulle qui sera acceptée si la procédure ne produit pas de segmentation acceptable d'ordre au

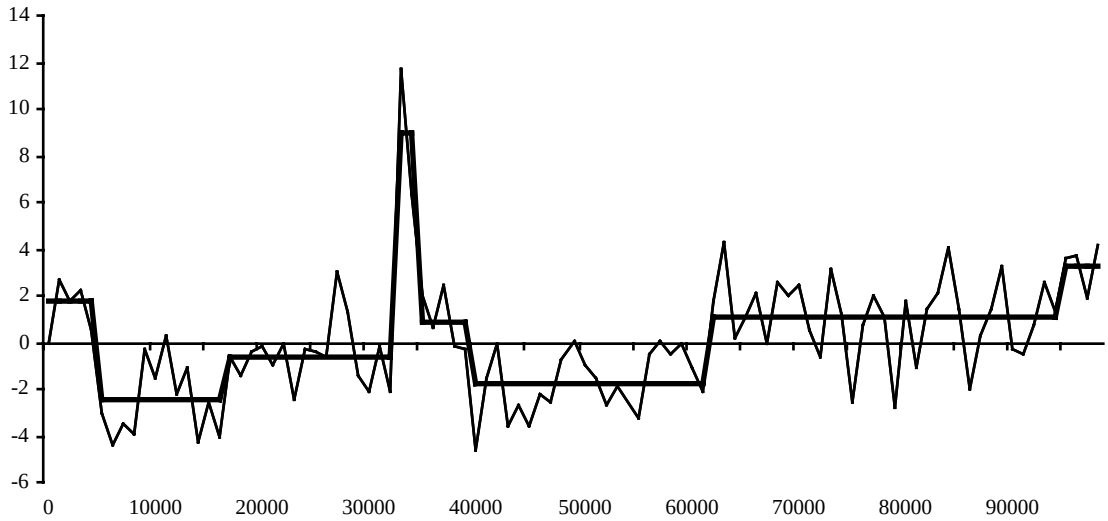
moins égal à deux, et rejetée dans le cas inverse. Naturellement, la décision est soumise à un risque de première espèce : rejeter l'hypothèse nulle alors qu'elle est vraie. Pour évaluer ce risque, à différentes valeurs de α, les auteurs de la publication citée ont appliqué ce test à un ensemble de 100 séries stationnaires, comportant chacune 50 valeurs aléatoires distribuées normalement. Pour $\alpha = .01$, le test de Scheffé conduit à segmenter, à tort, 11 de ces séries, ce qui peut être considéré comme un niveau de "bruit" acceptable pour une application, comme celle présentée ici, où l'on souhaite naturellement ne pas "trop" couper les œuvres littéraires... C'est ce qui risquerait de se produire avec une valeur trop élevée de α qui augmenterait le risque de première espèce. Cependant, cette simulation ne nous apprend rien sur le risque de deuxième espèce (rejeter la segmentation alors que la série n'est pas stationnaire).

4. Application : segmentation des *Voyages de l'autre côté* de J.-M. Le Clézio

La première partie de l'algorithme de segmentation appliqué au roman *Voyages de l'autre côté* aboutit à définir un découpage optimal pour un nombre de segments choisi. Nous donnons, en annexe à cette communication, les résultats des itérations successives du programme et nous reproduisons ci-dessous la segmentation du roman en 4 puis 8 parties.

Graphique 2. Découpage en 4 puis 8 segments des Voyages de l'autre côté (en fonction de la diversité du vocabulaire, variables centrées et réduites)





L'opération présente plusieurs intérêts. En premier lieu, elle offre des découpages plus ou moins fins, tout en indiquant précisément les moyennes, les débuts et les fins de segments. Elle permet également de repérer les coupures "stables" qui représentent les principales césures du texte… Par exemple, contrairement à ce que pouvait laisser attendre un regard superficiel, ou une technique classique, de type "moyenne mobile", la principale césure ne se situe probablement pas aux alentours des 33.000 mots mais autour de 62.000.

En se reportant au texte, il est possible de localiser avec précision cette coupure — qui correspond d'ailleurs à un point d'inflexion de la courbe d'accroissement du vocabulaire (reproduite dans la communication de M. Kastberg) — à la page 196 du roman, page qui ne comporte que ces quelques mots :

"Pourquoi ne pas aller là où il n'y a pas d'hiver ni d'été ? Où se trouve cet endroit ? Quand vient l'hiver vous frissonnez. Quand vient l'été vous transpirez."

La page suivante, déplace le lieu du récit — si tant est qu'on puisse employer cette expression pour ce texte typique du "nouveau roman" — et introduit de nouveaux personnages qui occuperont plus ou moins la fin du livre.

La série des segmentations donne également une indication concernant l'homogénéité relative d'un segment sans avoir besoin de se reporter aux valeurs de l'écart-type pour le segment considéré. Il suffit que ce segment réapparaisse sans changement lors des segmentations plus fines. Ainsi, du point de vue de la diversité du vocabulaire, le dernier tiers du texte — qui commence après la page qui vient d'être citée — "résiste" jusqu'à la septième segmentation (l'accroissement du vocabulaire vérifie ce même phénomène). On peut donc en conclure que la dernière partie du livre est la plus homogène thématiquement et stylistiquement.

Enfin, l'opération permet de repérer les passages singuliers. Par exemple, les segmentations fines placent nettement à part trois brefs passages : le début du roman — jusqu'à 5.000 mots — c'est-à-dire avant le chapitre intitulé "Naja-Naja" qui introduit l'héroïne principale du roman — au milieu, le bref passage déjà cité entre 33 et 35.000 mots et, enfin, les 5.000 derniers mots environ qui semblent presque aussi singuliers que le début de l'oeuvre. Elles soulignent également l'homogénéité relative du troisième tiers du livre (moins les derniers mots).

Après avoir localisé, avec précision, les principaux "épisodes" contenus dans le livre, on constituera autant de sous-corpus que de segments retenus, et on analysera leur vocabulaire spécifique grâce aux techniques classiques.

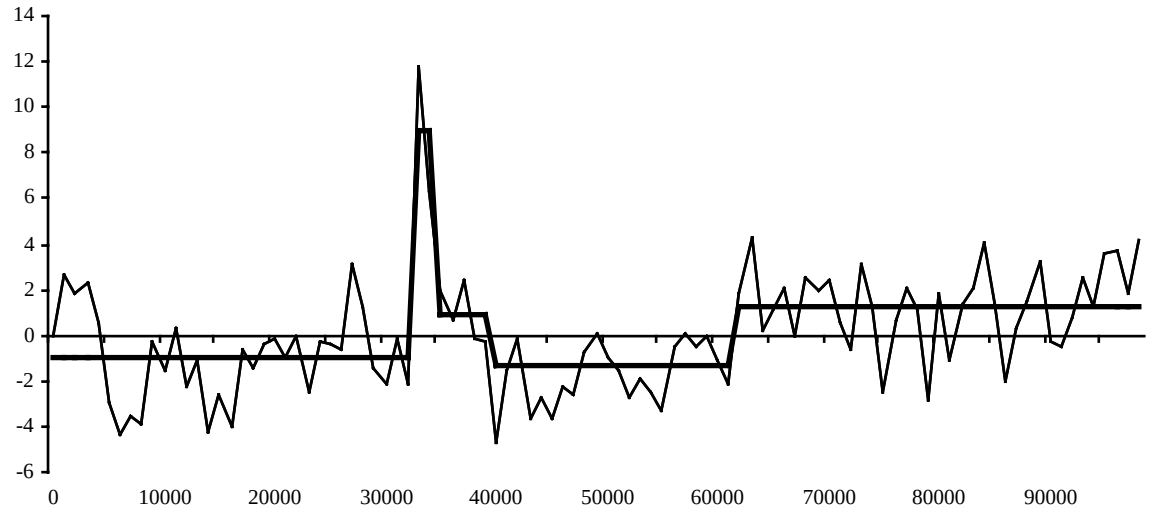
Toutefois, si l'on centre l'analyse, non plus sur les périodes, mais sur les principales ruptures dans le texte, il faut recourir au test statistique décrit ci-dessus. Au sein des *kième* et *(k+1)ième* segments voisins, le phénomène "diversité du vocabulaire" est considéré comme stationnaire autour des moyennes locales v_k et v_{k+1}, si l'écart entre ces deux moyennes locales est trop faible, selon le test de validité. Autrement dit, on considérera qu'il n'y a pas une variation suffisante dans la diversité pour conclure à un changement stylistique significatif. La segmentation sera rejetée.

Ce test a été intégré dans l'algorithme de segmentation. Quand il est activé, le programme s'interrompt à l'ultime segmentation — pour laquelle tous les contrastes entre segments sont non nuls — qui peut être considérée comme le découpage optimal (au sens des moindres carrés).

Appliquée au roman de JM Le Clezio, cette procédure localise 4 ruptures significatives, autour des passages analysés ci-dessus, et découpe le roman en 5 "épisodes" (graphique 3). Trois de ces épisodes présentent une taille suffisante pour réaliser une étude contrastive de leur vocabulaire par la technique des spécificités, ou par toute autre technique d'analyse stylistique comme celles présentées par ailleurs (Monière et Labbé, 2002).

Graphique 3. Segmentation optimale de Voyages de l'autre côté

(variable centrée et réduite)



En conclusion, nous voudrions d'abord souligner que l'outil est puissant et semble bien adapté aux corpus d'assez grandes dimensions, telles les œuvres d'un écrivain, une collection d'articles de journaux, etc. Nous pensons que cette méthode permettra de supprimer l'un des points "aveugles" de la statistique textuelle : le découpage sans biais des corpus en autant de sous-parties qu'ils en contiennent effectivement et non plus selon les a priori de l'observateur…

Naturellement, ces techniques sont "exploratoires". Elles assistent le lecteur sans du tout se substituer à lui ; elles n'épargnent pas le "retour au texte" mais le facilitent. L'observateur saura où il doit chercher et prendra les décisions ultimes : nombre des césures, localisation exacte de celles-ci dans le texte…

Cependant, l'outil est actuellement en cours de développement. Il n'est pas possible d'entrer ici dans le détail des discussions autour de l'analyse des séries chronologiques dont nous tirons cet algorithme. Rappelons que ce type d'analyses rencontre un certain nombre de problèmes classiques :

— la robustesse (c'est-à-dire le degré de sensibilité à la modification marginale de certaines données ou à l'introduction d'une donnée aberrante…) ;

— la limitation relative de la puissance — à cause de l'explosion combinatoire qui n'est pas totalement maîtrisée pour certains profils de séries très "accidentées" ;

— la difficile appréciation des risques de première et de deuxième espèce : faut-il mieux passer à côté d'une rupture réelle dans une œuvre où découper celle-ci à tort ?

Rappelons enfin que, par construction, le résultat dépend de l'ordre des données. A l'heure actuelle, en effet, l'algorithme ne s'applique qu'aux séries "ordonnées" — comme les données chronologiques ou l'évolution du vocabulaire dans une œuvre. Certes, il semble possible d'appliquer un algorithme semblable aux séries "sans ordre" (où le classement des données est arbitraire) mais les problèmes posés par cette transposition dépassent le cadre de cette communication.

Références

Alvarez R., Becue M. et Lanero J.-J. (2000). Vocabulary Diversity and its Variability : A Tool for the Analysis of Discoursive Strategies. Application to the Investiture Speeches of the Spanish Democracy. In Rajman M. et Chappelier J.-C. (eds). *Actes des 5e journées internationales d'analyse des données textuelles*. Lausanne, Ecole polytechnique fédérale, 2000, vol 1, p 111-118.

Cossette A. (1994). *La richesse lexicale et sa mesure*. Paris, Genève, Slatkine-Champion.

Dagnelie P. (1970). *Théories et méthodes statistiques*. Tome 2, Gembloux, Duculot.

Hubert P. et Labbé D. (1988). Note sur l'approximation de la loi hypergéométrique par la formule de Muller. In Labbé D., Serant D. et Thoiron P. *Etudes sur la richesse et la structures lexicales*. Paris-Genève, Slatkine-Champion, 1988, p 77-91.

Hubert P., Carbonnel J.-P. et Chaouche A. (1989). Segmentation des séries hydrométéorologiques - Application à des séries de précipitations et de débits de l'Afrique de l'Ouest. *Journal of hydrology*, 110, p 349-367.

Hubert P. et Labbé D. (1988). Un modèle de partition du vocabulaire. In Labbé D., Serant D. et Thoiron P. *Etudes sur la richesse et la structures lexicales*. Paris-Genève, Slatkine-Champion, 1988, p 93-114.

Hubert P. et Labbé D. (1994). "La richesse du vocabulaire". *Communication au congrès de l'ALLC-ACH*, Paris, La Sorbonne (Reproduit dans *Lexicometrica*, 0, 1997).

Kastberg-Sjoblom M. et Brunet E. (2000). La thématique. Essai de repérage automatique dans l'œuvre d'un écrivain, RAJMAN M. et CHAPPELIER J.-C. (eds), *Actes des 5e journées internationales d'analyse des données textuelles*, Lausanne, Ecole polytechnique fédérale, 2000, vol 2, p 457-466.

Kastberg-Sjoblom M. (2002). Le choix de la lemmatisation. Différentes méthodes appliquées à un même corpus. *Communication aux VIe Journées d'Analyse des Données Textuelles*, Saint-Malo, mars 2002.

Labbé D. (1990a), *Normes de saisie et de dépouillement des textes politiques*, Grenoble, Cahier du CERAT.

Labbé D. (1990b), *Le vocabulaire de F. Mitterrand*, Paris, Presses de la FNSP.

Labbé D. (1998), La richesse du vocabulaire politique : de Gaulle et Mitterrand, Mellet S. et Vuillaume M. (eds). *Mots chiffrés et déchiffrés: mélanges offerts à Étienne Brunet*. Paris, Champion, p. 173-186.

Lebart L. et Salem A. (1994), *Statistique textuelle*, Paris, Dunod.

Minoux M. (1989). *Programmation mathématique : Théorie et Algorithmes*. tome 2, Paris, Dunod.

Monière D. et Labbé D. (2002). Essai de stylistique quantitative. *Communication aux VIe Journées d'Analyse des Données Textuelles*. Saint-Malo, mars 2002.

Muller C. (1977). "Calcul des probabilités et calcul d'un vocabulaire". *Langue française et linguistique quantitative.* Genève-Paris, Slatkine-Champion, 1979, pp. 167-176.

Muller C. (1979). *Principes et méthodes de statistique lexicale*. Paris, Hachette, 1977.

Scheffe M. (1959). *The Analysis of Variance*. New York, Wiley.

Annexe

Segmentations successives de *Voyages de l'autre côté* en fonction de la diversité du vocabulaire

Nombre de segments	D	Début et fin (en milliers)	Moyenne
2	526	0 - 61	-0,79
		62 - 98	1,33
3	430	0 - 32	-1.00
		33 - 34	9,04
		35 - 98	-0.23
4	325	0 - 32	-1,00
		33 - 34	9,04
		35 - 61	-1,27
		62 - 98	1,33
5	288	0 - 4	1,83
		5 - 32	-1,41
		33 - 34	9,04
		35 - 61	-1,27
		62 - 98	1,33
6	259	0 - 4	1,83
		5 - 32	-1,41
		33 - 34	9,04
		35 - 39	0,92
		40 - 61	-1,77
		62 - 98	1,33
7	231	0 - 4	1,83
		5 - 16	-2,54
		17 - 32	-0,56
		33 - 34	9,04
		35 - 39	0,92
		40 - 61	-1,77
		62 - 98	1,33

Nombre de segments	D	Début et fin (en milliers)	Moyenne
8	213	0 - 4	1,83
		5 - 16	-2,54
		17 - 32	-0,56
		33 - 34	9,04
		35 - 39	0,92
		40 - 61	-1,77
		62 - 94	1,08
		95-98	3,33
9	204	0 - 4	1,83
		5 - 16	-2,54
		17 - 32	-0,56
		33 - 34	9,04
		35 - 39	0,92
		40 - 47	-2,65
		48 - 61	-1,27
		62 - 94	1,08
		95-98	3,33
10	193	0 - 4	1,83
		5 - 9	-3,70
		9 - 16	-1,9
		17 - 32	-0,56
		33 - 34	9,04
		35 - 39	0,92
		40 - 47	-2,65
		48 - 61	-1,27
		62 - 94	1,08
		95-98	3,33