# Segmentation des corpus chronologiques : 143 ans de discours gouvernemental au Québec

Denis Monière [1], Dominique Labbé [2]

[1] Université de Montréal

[2] Institut d'Etudes Politiques de Grenoble

## Résumé

Méthode originale pour segmenter un corpus chronologique en périodes homogènes. On calcule l'accroissement du vocabulaire et son ajustement par une tendance. Un algorithme de segmentation associé à des tests de validité donne le découpage optimal du corpus. Une série d'indicateurs mesure l'ampleur des mouvements de vocabulaire caractérisant chacune des périodes. Application aux déclarations du gouvernement québécois à l'ouverture de chaque session du parlement provincial de 1867 à 2009.

**Mots-clés :** discours politique, accroissement du vocabulaire, segmentation des corpus, corpus chronologique, Québec

## Abstract

A method for segmenting large textual corpora in uniform periods. Firstly, vocabulary growth is adjusted by calculating the trend. Then a segmentation algorithm, associated with validity tests, indicates the optimal succession in distinct periods. This method is applied to the "Queen's speeches" which have been given by the Quebec government at the beginning of each parliamentary session since 1867 and up until 2009.

**Keywords :** political discourse, vocabulary increase, corpus segmentation, chronological corpus, Quebec

## 1. Introduction

Intuitivement on peut supposer que le temps influence le vocabulaire politique soit parce qu'il transforme le lexique, soit parce que les problématiques et enjeux traités par les décideurs politiques changent selon les époques. Comment mesurer l'effet du temps sur le vocabulaire politique ? Comment identifier de façon objective les périodes de stabilité et les périodes de changement politique ? L'arrivée d'un nouveau parti au pouvoir provoque-t-elle des changements dans le discours gouvernemental ?

Le découpage des vastes corpus – dont la production s'étale sur une longue période - demeure une opération empirique. Le choix des césures reste généralement intuitif : changements de dirigeants, de majorités parlementaires, années civiles, textes que l'on estime capitaux… Sans doute, une longue familiarité avec l'objet d'étude peut-elle justifier ces choix. Il n'en reste pas moins qu'ils sont effectués à partir de critères extérieurs aux textes. Dès lors, qui peut garantir que les résultats des analyses de données textuelles portant sur ces découpages ne sont pas influencés par les ciseaux de l'analyste ?

Nous présentons une méthode de segmentation automatique des corpus qui devrait supprimer cet arbitraire. Pour cela il fallait un corpus de discours s'étendant sur une longue période, mais avec des conditions d'énonciation semblables. Les discours du trône des premiers ministres québécois, depuis 1867, répondent à cette contrainte : dans un cadre constitutionnel stable, chaque session parlementaire annuelle a été ouverte par un discours exposant les principales orientations de la politique gouvernementale. On les a appelés discours du trône parce que dans la tradition parlementaire britannique, c'est le monarque ou son représentant qui lisait ce discours préparé par le premier ministre. En 1867, l'acte de l'Amérique du Nord britannique a institué quatre provinces autonomes – dont le Québec – qui étaient dotées d'un parlement et d'un exécutif responsable devant la chambre basse. La fonction de premier ministre n'était pas formellement reconnue, mais elle existait par convention, le chef du parti qui détenait la majorité des sièges à l'Assemblée nationale étant appelé à former le gouvernement.

Tous ces discours ont pu être retrouvés. Ils se prêtent bien à une analyse lexicométrique longitudinale en raison de leur caractère formalisé : ils sont soumis à un processus d'élaboration rigoureux afin de refléter les objectifs du gouvernement et sont toujours prononcés dans le même cadre. Dès lors, l'effet du temps peut être analysé sans «parasitage» dus à des changements dans les conditions d'énonciation. Les 128 discours dont nous disposons comportent 309.237 mots et 8.264 vocables différents.

La segmentation de ce vaste corpus s'effectuera en deux temps : détermination d'une tendance séculaire puis délimitation de périodes homogènes.

## 2. Détermination de la tendance séculaire

L'observation de l'accroissement du vocabulaire, en fonction de la longueur des textes et des corpus est un thème classique de l'analyse des données textuelles (pour une synthèse de la question : Wimmer and Altmann, 1999). Nous présentons ici une méthode pour déterminer la tendance à l'œuvre dans cet accroissement (Hubert and Labbé, 1988b ; Labbé et al., 2004).

### *2.1. Accroissement du vocabulaire dans un corpus chronologique*

Dans un corpus comme celui des discours du trône, il se produit un afflux important de vocabulaire au début puis cet accroissement ralentit lentement sans jamais devenir nul quelle que soit la longueur du corpus (Müller, 2002). Soit N la longueur totale du corpus (ici : 309.237 mots) et V son vocabulaire (ici : 8.264 vocables différents). Fig. 1 représente cet accroissement (trait gras) en mesurant, selon un pas $t$ de mille mots, le nombre de vocables différents ($V_t$) apparus depuis le début.

La courbe des valeurs observées (trait gras) est irrégulière et suggère qu'il existe des accidents. Pour les localiser, on utilise une procédure en deux temps : détermination de la tendance puis mise en valeur des variations «conjoncturelles» autour de cette tendance.

### *2.2. Calcul de la tendance chronologique*

Premièrement, la courbe est ajustée selon une procédure décrite dans Hubert and Labbé (1988a), Hubert and Labbé (2002), Labbé et al. (2004).

Le corpus est découpé en $T$ segments successifs selon un pas régulier (ici de 1.000 mots).

Les $V$ vocables du corpus total sont rangés par ordre croissant de fréquence dans $n$ classes de fréquences. Soit $V_i$ le nombre de vocables qui apparaissent $i$ fois.

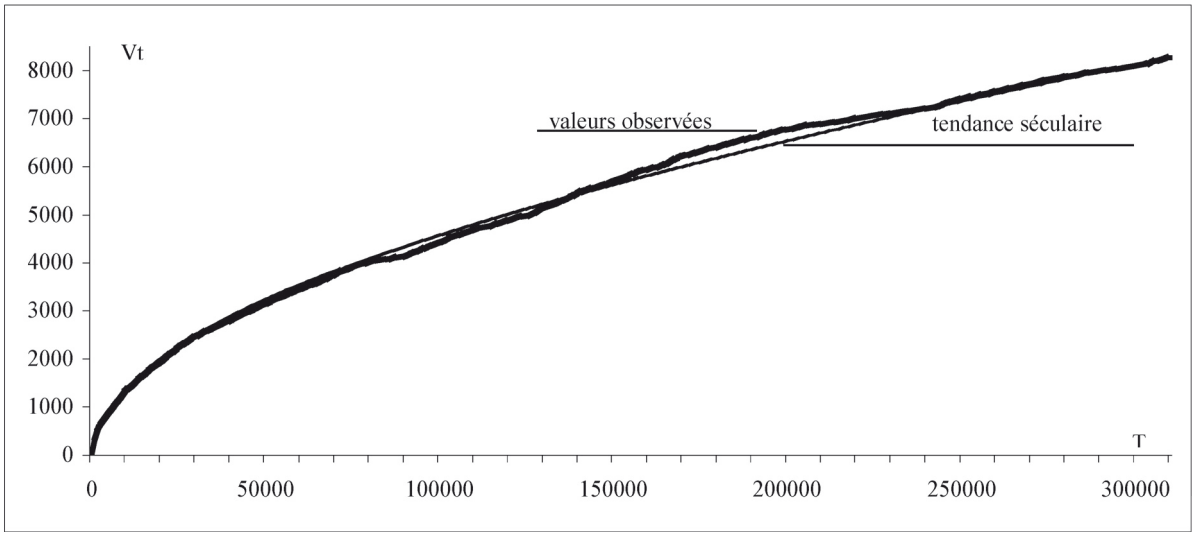


*Figure 1 : Accroissement du vocabulaire (ordre chronologique, intervalles de 1000 mots)*

$V'_t$ – nombre de vocables différents attendus dans un échantillon de $N'_t$ mots – est calculé grâce à la formule (1)

$$V'(u_t) = p \cdot u_t \cdot V + (1 - p)\left[V - \sum_{i=1}^{i=n} V_i Q_i(u_t)\right] \quad (1)$$

*Avec :* $u_t = \frac{N'_t}{N}$ $N'_t$ variant de 1 000 mots à $N$ (309.237), en augmentant selon un pas de 1000 mots,

$Q_i(u_t) = (1 - u_t)^i$

$p$ est le paramètre de partition qui estime, sur le corpus entier, la part respective du vocabulaire général et du vocabulaire spécialisé (pour le calcul de ce paramètre voir Hubert and Labbé, 1988b). Il varie entre 0 (pas de vocabulaire spécialisé) et 1 (pas de vocabulaire général). Le vocabulaire général est mobilisé de manière uniforme tout au long du corpus. Son accroissement peut être décrit grâce à la formule de Muller (1977) qui simule le résultat d'un tirage exhaustif – ie sans remise - de $N'_t$ mots dans une urne contenant $N$ mots (Hubert and Labbé, 1988a). C'est le seconde partie de la formule (1). Le vocabulaire spécialisé n'apparaît que pour traiter un thème spécifique selon une tendance linéaire sur l'ensemble du corpus – premier membre de la formule (1) – mais avec de brusques augmentations à chaque fois que l'on change de thème (Hubert and Labbé, 1988b). D'où la possibilité de découper le corpus en fonction des principales coupures thématiques en localisant ces brusques afflux.

Appliquée au corpus, la formule (1) fournit un excellent ajustement de l'accroissement effectif du vocabulaire durant la période sous revue (trait maigre sur Fig. 1). On constate que, par endroits, les deux courbes s'*écartent l'*une de l'autre. On peut mettre en valeur ces écarts et déterminer ceux qui ne peuvent être dus au hasard de la manière suivante.

### *2.3. Variations cycliques dans un corpus chronologique*

La tendance devient l'axe horizontal et les T valeurs observées sont centrées sur la tendance : $V'_*(u_t) - V'(u_t)$. L'*écart type de cette variable est calculée sur la partie générale du vocabulaire* $(1-p)V$, à l'aide de la formule 2.

$$\sigma(u_t) = \sqrt{Var[V'(u_t)]} = \sqrt{(1-p)\cdot\sum_{1}^{n} V_i Q_i(u_t)[1-Q_i(u_t)]} \quad (2)$$

On utilise cet écart type pour obtenir les *valeurs centrées et réduites* notées *V»* (formule 3)

$$V_*''(u_t) = \frac{V_*'(u_t) - V'(u_t)}{\sigma(u_t)} \quad (3)$$

La Figure 2 ci-dessous présente le résultat de ce calcul appliqué au corpus des déclarations gouvernementales québécoises. Un intervalle de fluctuation « normal » de ± 1,96σ est délimité par les traits pointillés. On peut considérer, avec moins de 5% de chance de se tromper, que tout point situé en dehors de cet intervalle s'écarte significativement de la tendance chronologique. Comme la grande majorité de la courbe se situe en dehors de cet intervalle, on en tire que l'accroissement n'a pas été régulier et qu'il est légitime de vouloir segmenter ce corpus en périodes homogènes.

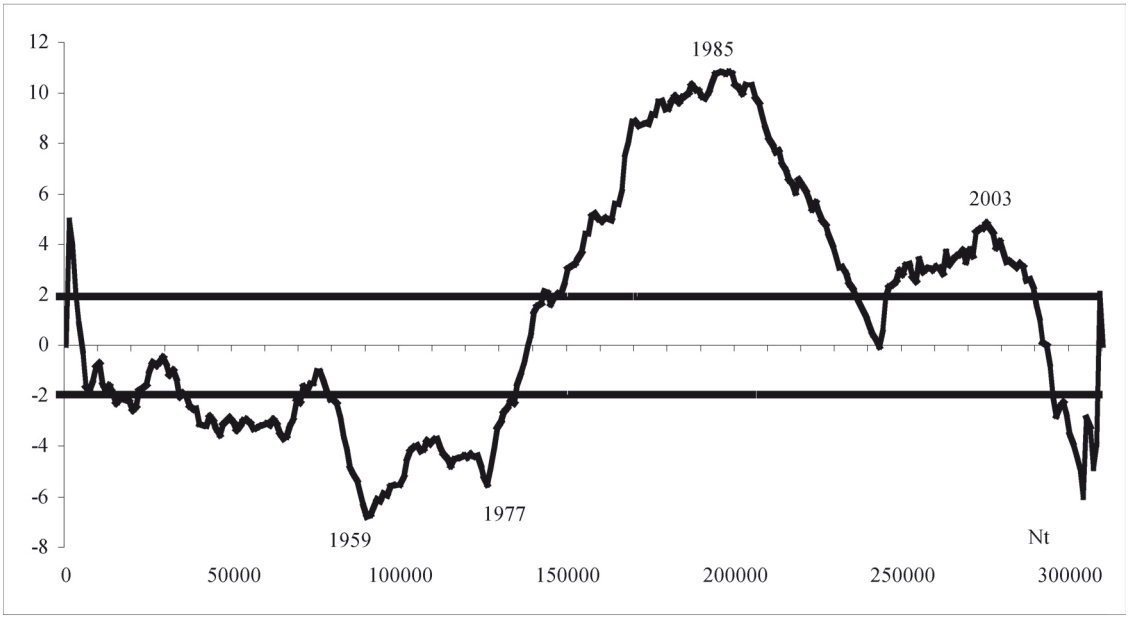


*Figure 2 : Accroissement du vocabulaire dans les déclarations gouvernementales (valeurs centrées et réduites)*

Nous avons daté les points les plus hauts et les plus bas (la fin de cette communication revient sur leur signification historique). Ils permettent d'identifier une difficulté : les 110 premières années (1867-1977) – la partie gauche de la figure – occupent une place moindre que les 32 dernières années (1978-2009). En effet, jusqu'en 1977 les premiers ministres suivaient l'usage anglais qui veut que le discours de la reine soit bref – de quelques centaines à 2.000 ou 3.000 mots au maximum –, annonçant les principaux projets inscrits à l'ordre du jour de la session sans les développer. Depuis 1977, les discours sont nettement plus longs (dépassant parfois les 10.000 mots) et présentent un véritable programme argumenté. En 2009, J. Charest est revenu à la tradition antérieure. Il faut donc découper le corpus en deux sous-parties. Dans la suite de cette communication nous allons concentrer l'analyse sur la première (1867-1977) en répondant à une seconde question : comment découper des périodes homogènes au sein de cette longue période ?

## 3. Mise en valeur des coupures historiques

### *3.1. Ajustement linéaire des écarts à la tendance*

On propose d'analyser la co-variation des valeurs de la variable *V''* et du temps. La succession des *t* observations réalisées à espaces réguliers fournit un indice temporel efficace étant donné que, durant la période 1867-1977, la longueur des discours varie assez peu.

On utilise la technique classique de l'ajustement linéaire d'une série chronologique. Pour un point d'abscisses *t*, l'ordonnée de la droite d'ajustement – notée $VL_t$ – est donnée par la formule :

$VL_t = a \,.\, V''_t + b$

Une telle droite passe par un point *M* dont les coordonnées sont les moyennes $\bar{t}$ et $\overline{v''}$, où $\bar{t}$ est le numéro d'ordre de l'observation médiane de la série, si *T* (nombre total d'observations) est impair, ou un point théorique situé à mi-chemin entre les deux observations médianes si *T* est pair. *a* est la « pente » de la droite d'ajustement (formule 4).

$$a = \frac{\sum_1^T (v''_t - \overline{v''})(t - \bar{t})}{\sum_1^K (t - \bar{t})^2} \text{ avec } \overline{v''} = \frac{\sum_1^T v''_t}{T} \tag{4}$$

Considérons les treize premières observations (tableau 1) qui correspondent à la période 1867-1878 et qui couvrent les 9 premiers discours (il n'y a pas eu de discours en 1868 et 1870). Remarques : nous expliquons plus bas la méthode de découpage ; le tableau 1 donne les deux premières décimales mais, comme dans tout calcul avec plusieurs itérations, on utilisera le maximum de précision.

| $t$ | $N'_t$ | $V''_t$ | $t - \bar{t}$ | $(v''_t - \overline{v''})(t - \bar{t})$ | $VL_t$ |
|---|---|---|---|---|---|
| 1 | 500 | 2,32 | -6 | -13,92 | 3,97 |
| 2 | 1000 | 3,96 | -5 | -19,80 | 3,39 |
| 3 | 1500 | 3,7 | -4 | -14,80 | 2,81 |
| 4 | 2000 | 2,93 | -3 | -8,79 | 2,22 |
| 5 | 2500 | 2,14 | -2 | -4,28 | 1,64 |
| 6 | 3000 | 1,01 | -1 | -1,01 | 1,06 |
| 7 | 3500 | -0,16 | 0 | 0,00 | 0,48 |
| 8 | 4000 | -0,22 | 1 | -0,22 | -0,10 |
| 9 | 4500 | -0,89 | 2 | -1,78 | -0,68 |
| 10 | 5000 | -1,22 | 3 | -3,66 | -1,26 |
| 11 | 5500 | -1,89 | 4 | -7,56 | -1,84 |
| 12 | 6000 | -2,61 | 5 | -13,05 | -2,42 |
| 13 | 6500 | -2,81 | 6 | -16,86 | -3,00 |
| *Somme* | | | | *-105,73* | |

*Tableau 1 : Tableau de calcul de l'ajustement linéaire des 13 premières valeurs de la série*

On a : $\overline{v''} = 0{,}4815$, $\bar{t} = 7$ et $\sum_1^T (t - \bar{t})^2 = 146$

La pente de la droite d'ajustement est égale à : $a = \frac{-105{,}73}{146} = -0{,}58$

Sur les 26 observations suivantes (1880-1893), les mêmes calculs donnent : $\overline{v''}$ = -1,99 et $a$ = 0,01. On reporte sur le graphique originel, les valeurs de $V»_t$ (écarts à la tendance centrés et réduits : trait maigre) et de $VL_t$ (leur ajustement linéaire : trait gras) (figure 3). La tendance séculaire est figurée par la ligne pointillée horizontale.

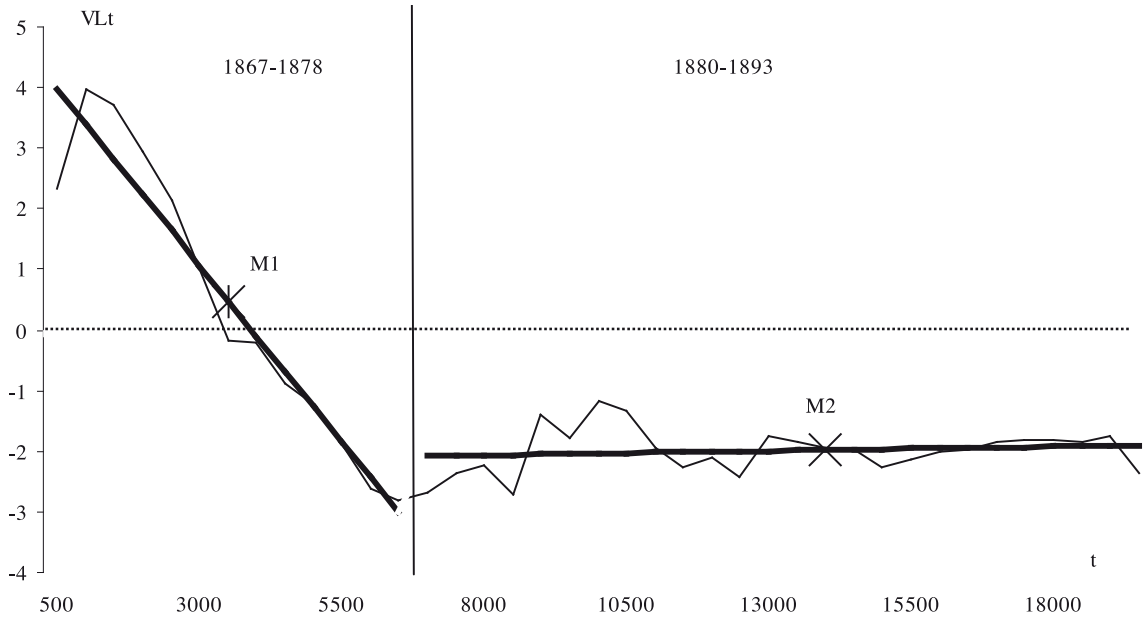


*Figure 3 : Ajustements des années 1867-1893 et détermination de la première coupure*

### *3.2. Interprétation des résultats*

Pour la première période, la valeur moyenne des rapports à la tendance (*M1* = 0,4815, soit environ un demi écart-type) signifie que, globalement, il est apparu un peu plus de mots nouveaux que ce que laisserait attendre la tendance séculaire, mais que ce léger excédent ne peut être considéré comme significatif. Cet apport s'est produit au tout début (plus précisément dans les 1000 premiers mots) ; la suite étant surtout de la répétition. En effet, la pente de la droite d'ajustement (-0,58) indique que, sur l'ensemble de la période, le rythme d'apparition moyen des mots nouveaux est de 58% inférieur à cette même tendance.

Le point moyen de la seconde période (*M2*) se situe à la limite des deux écarts types (-1,99), ce qui indique que l'apport en mots nouveaux depuis le début est significativement faible. Mais la pente de la droite quasiment horizontale (0,01) signale que, au cours de cette seconde période l'accroissement du vocabulaire s'est effectué au rythme de la tendance séculaire. Si l'on s'intéresse spécialement à cette seconde information, on peut négliger la hauteur relative de l'ajustement et ne retenir que la pente de la droite et sa longueur (la durée de la période).

Deux questions sont posées :

- comment juger de la qualité de l'ajustement ? Classiquement, pour un segment comprenant $T$ observations, on calcule l'écart $E$ entre toutes les valeurs « observées » ($V''_t$) et les valeurs ajustées ($VL_t$) correspondantes :

$$E = \sqrt{\frac{\sum_{1}^{T} (v''_t - VL_t)^2}{T}}$$

Les seuils de signification sont ceux de la loi normale réduite. Lorsque *E* est inférieur à 1,96, l'ajustement peut être considéré comme acceptable. Pour la première période, cet écart moyen est égal à 0,63 et, pour la seconde, à 0,36. Cela confirme le bon ajustement que l'on peut constater visuellement sur Fig. 3. Cependant, l'objectif est d'obtenir une valeur de *E* aussi faible que possible. Plus *E* sera élevé, plus on pourra mettre en doute l'homogénéité du segment ou encore la pertinence de sa non-segmentation. Ceci amène une question plus générale :

- comment estimer les risques d'erreur de première espèce – segmenter à tort une série homogène – et de seconde espèce – ne pas segmenter un corpus constitué de l'amalgame de deux ou plusieurs séries différentes ? En fait la réponse dépend d'abord d'une question apparemment simple : quel effectif minimal une série doit-elle comporter pour que l'on puisse valablement opérer un ajustement linéaire ? La question porte en fait sur le calcul de la moyenne, de l'écart-type, des coefficients de corrélation, etc. Des expériences simples apportent des réponses empiriques à ces 3 questions (effectif plancher, erreurs de première et de seconde espèce). A l'aide du générateur de nombres aléatoires de l'ordinateur, on constitue des séries dont on fait varier les effectifs et la dispersion et que l'on soumet à l'algorithme en comptant le nombre de fois qu'il les segmente à tort (erreur de première espèce). Puis l'on colle bout à bout plusieurs séries différentes – dont on fait également varier les effectifs et la dispersion - en comptant le nombre de fois que l'algorithme échoue à segmenter correctement ces séries hétérogènes (risque d'erreur de seconde espèce). De ces expériences actuellement en cours, il ressort qu'une dizaine d'observations est l'effectif plancher en dessous duquel on ne peut valablement calculer un ajustement linéaire comme ceux présentés ci-dessus. Les erreurs de première et deuxième espèce sont en cours d'évaluation.

### *3.3. Algorithme*

Après avoir calculé la tendance séculaire et avoir rapporté les valeurs observées à cette tendance, le programme parcourt toute la série en recherchant le « meilleur » découpage possible. Les contraintes imposées sont les suivantes : respecter l'ordre chronologique et ne pas constituer de segment comportant moins de 10 observations. Sous ces contraintes, le meilleur découpage est celui qui minimise les écarts entre les valeurs observées (V") et les valeurs ajustées (VL) pour chacun des segments découpés et qui maximise les contrastes entre les pentes (*a*) des droites d'ajustement pour chacune des paires de segments contigus.

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. Pour éviter l'explosion combinatoire, on utilise un algorithme d'optimisation de type « branch and bound » (Minoux, 1983).

## 4. Segmentation des discours du trône

### *4.1. Les périodes du discours gouvernemental québécois (1867-2009)*

Appliquée au corpus des discours du trône québécois sur la période 1867 à 1977, la méthode permet d'isoler 8 périodes dans le discours gouvernemental (Fig. 4).

Cette figure est une des présentations possibles des résultats. Les points moyens sont tous alignés au milieu du graphe qui ne figure plus que les pentes, la durée des périodes et l'ampleur des changements apportés. Ainsi, les mouvements les plus amples se constatent durant les deux

dernières périodes. Entre 1948 et 1958, M. Duplessis reprend quasiment les mêmes thèmes (Labbé and Monière, 2003 ; 2009) d'où la longueur de la flèche n° 7 et sa pente descendante très forte. Après sa mort, le discours se renouvelle profondément, ce qu'indique la flèche n° 8 de longueur quasiment équivalente et d'orientation inverse à la précédente.

Appliquée à la période 1977 à 2009, la même méthode délimite les 4 dernières périodes qui correspondent aux tournants de la vie politique québécoise après la coupure principale coïncidant avec l'arrivée au pouvoir du parti québécois en novembre 1976 (Tab. 2).

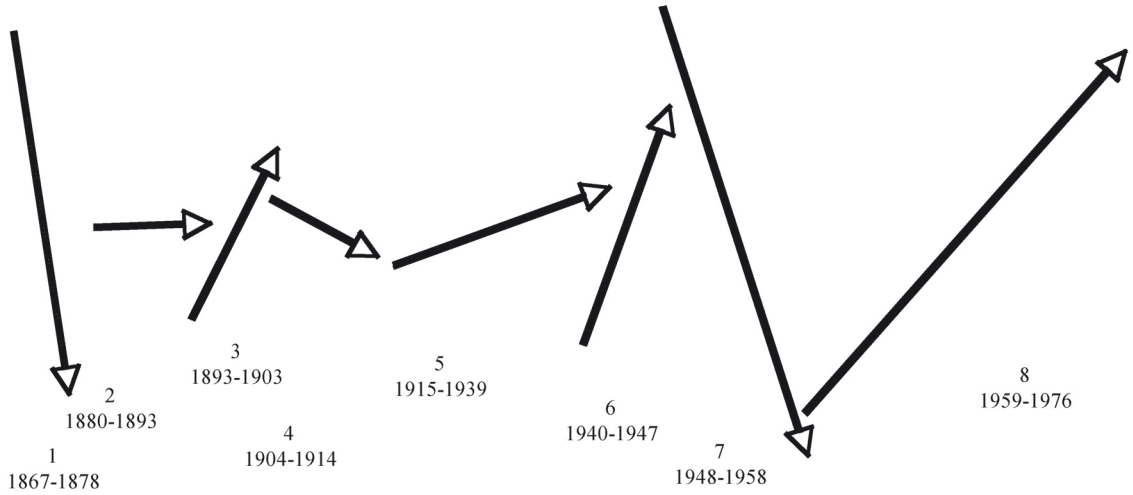


*Figure 4 : Les 8 périodes du corpus "Discours du trône québécois" entre 1867 et 1976*

| *N°* | *Début* | *Fin* | *Premiers ministres (ordre chronologique) et allégeance partisane* |
|---|---|---|---|
| 1 | 1867 | 1878 | Chauveau, Ouimet, Boucher de Boucherville (conservateurs), Joly de Lotbinière (libéral) |
| 2 | 1880 | 1893 | Chapleau, Mousseau, Ross (conservateurs), Mercier (libéral), Boucher de Boucherville (conservateurs) |
| 3 | 1893 | 1903 | Taillon, Flynn (conservateurs), Marchand, Parent (libéraux) |
| 4 | 1904 | 1913 | Gouin (libéral) |
| 5 | 1915 | 1939 | Gouin, Taschereau, Godbout (libéraux), Duplessis (union nationale) |
| 6 | 1940 | 1947 | Godbout (libéral), Duplessis (union nationale) |
| 7 | 1948 | 1958 | Duplessis (union nationale) |
| 8 | 1959 | 1976 | Sauvé (union nationale), Lesage (libéral), Johnson, Bertrand (union nationale), Bourassa (libéral). |
| 9 | 1977 | 1985 | Lévesque (parti québecois) |
| 10 | 1985 | 1994 | Bourassa (libéral) |
| 11 | 1994 | 2001 | Parizeau, Bouchard, Landry (parti québécois) |
| 12 | 2003 | 2009 | Charest (libéral) |

*Tableau 2 : Segmentation du corpus « discours du trône québécois (1867-2009) »*

### *4.2. Interprétation historique des résultats*

Le tableau appelle deux remarques. En premier lieu, il s'agit d'une étude à « gros grains ». Des découpages plus fins peuvent être effectués, selon la même méthode, à l'intérieur des périodes (Labbé and Monière, 2008). Deuxièmement, le tableau 2 montre que, avant 1948,

les leaders et les partis ne semblent pas être des variables très actives. Chaque période fait voisiner plusieurs premiers ministres de couleur politique différente : libéraux, conservateurs voire union nationale.

Le découpage des périodes recoupe certaines césures historiques comme la crise économique de 1878, le début des deux guerres ou des ruptures politiques comme celle de l'arrivée au pouvoir du Parti libéral en 1960 qui a inauguré ce qu'on a appelé la Révolution tranquille ou encore la prise du pouvoir du Parti québécois en 1977. La période 7 est plus surprenante car la coupure ne se produit pas à la fin de la guerre qui correspond au retour au pouvoir en 1944 de l'Union nationale qui succède alors au Parti libéral. Le changement lexical semble se produire avec un décalage de 4 ans et correspond à la réélection de l'Union nationale en 1948 qui obtient 90% des sièges de l'Assemblée législative et peut gouverner à sa guise sans être gênée par une opposition exsangue. Aucun des travaux portant sur la période duplessiste ne mentionne ce tournant de 1948 (voir, par exemple, Bourque and Duchastel, 1988 ; Bourque et al., 1994).

Quel a été le contenu lexical de chacune de ces périodes ?

### *4.3. Contenu lexical des périodes*

Pour déterminer le contenu lexical de chacune de ces périodes et sa singularité par rapport aux autres, on utilise la méthode du « vocabulaire caractéristique » (Monière et al., 2005) qui est inspirée des « spécificités du vocabulaire » (Lafon, 1984 : 64-66 ; Lebart and Salem, 1994 : 182).

Par exemple, pour la seconde période – 1880-1893 –, les substantifs les plus significativement suremployés indiquent que trois thèmes sont privilégiés : les *chemin de fer* et les *subsides* accordés aux *compagnies de*… deuxièmement, les autres secteurs de l'*économie* de la *province* : *phosphate*(*s*), *construction*, *mine*(*s*), *pont*(*s*), *exploitation*(*s agricoles et forestières*), *entreprise*(*s*) ; enfin l'état des *finances* de l'état provincial : *dépenses*, *emprunt*, *revenu*(*s*), *dette*, *recette*(*s*)… Certes la coupure de 1880 suit une bascule dans la chambre : J. Chapleau (conservateur) a succédé à un premier ministre libéral (de Lotbinière) qui n'est resté qu'un an au pouvoir. Le vocabulaire caractéristique montre que cette première césure provient surtout d'un interventionnisme accru de l'Etat provincial dans l'économie, interventionnisme provoqué par la crise économique qui a sévi à la fin des années 1870 et par l'état préoccupant des finances publiques qui en était résulté.

Les discours de la période 7 (1948-1958) – tous prononcés par le même homme : M. Duplessis – sont singuliers à beaucoup d'égards (pour une analyse des discours de M. Duplessis : Labbé and Monière (2009). Par rapport au reste du corpus, ces textes comportent très peu de verbes ; ils sont impersonnels – absence quasi-complète de pronoms personnels autres que la troisième personne et la première personne du pluriel (nous) – ils sont faiblement ancrés dans le temps et dans l'espace (très peu de dates, de noms propres, de chiffres). C'est un discours de principe plus que d'action. Les deux vocables les plus caractéristiques forment un syntagme répété de discours en discours : *droit*(*s*) *province*. La *défense des droits de la province* est en effet la principale thématique de cette période. L'*agriculture* (et l'adjectif *agricole*) forme le deuxième thème puis viennent : la *prospérité*, le *progrès*, la *santé*, le *bien-être* et l'*éducation*. Le logiciel a isolé deux phrases comme étant les plus caractéristiques des discours de cette période parce qu'elles contiennent la plus forte densité de vocables caractéristiques de cette période. Elles résument bien les deux thématiques principales du duplessisme :

Aussi le gouvernement continuera-t-il sa politique généreuse de prêt agricole qui répond aux besoins particuliers de notre agriculture et qui a produit, depuis une vingtaine d'années, de merveilleux résultats en affermissant la situation financière du cultivateur, en assurant le crédit de nos corporations municipales, scolaires et paroissiales, en

facilitant l'établissement de milliers de jeunes cultivateurs et en procurant à la classe agricole de notre province des conditions d'emprunt exceptionnellement avantageuses à notre époque de restriction du crédit et d'augmentation des taux d'intérêt (1957).

Nous sommes convaincus que le seul système gouvernemental approprié et juste est celui en vertu duquel l'état provincial et l'état fédéral, chacun dans sa sphère respective, possèdent les pouvoirs essentiels au gouvernement responsable et démocratique, et cela, tant au point de vue législatif et administratif qu'au point de vue financier (1955).

La même analyse a été conduite sur les autres périodes. Avant les années 1950, les crises économiques et les deux guerres mondiales sont les principales dates marquantes et le lexique de chaque période est marqué par l'empreinte de ces évènements majeurs. A partir de la fin des années 1940 – et au moins jusqu'en 2009 – la chronologie est politique et correspond à l'alternance des majorités parlementaires avec toutefois une confirmation : entre 1966 et 1970, l'Union nationale a poursuivi la politique de réformes de la « révolution tranquille » (Labbé and Monière, 2008).

## 5. Conclusions

Nous voudrions d'abord souligner que la méthode de segmentation présentée ci-dessus est aisée à mettre en œuvre, qu'elle se déroule de manière automatique, ce qui donne une grande solidité aux découpages obtenus.

Combiné avec la recherche des ruptures stylistiques dans le même corpus (Hubert et al., 2002), cet outil est bien adapté aux corpus de grandes dimensions, tels le discours gouvernemental québécois. Ces méthodes permettront de supprimer l'un des points « aveugles » de la statistique textuelle : le découpage des corpus en autant de sous-parties qu'ils en contiennent effectivement et non plus selon les *a priori* de l'observateur ou la doxa des historiens…

Naturellement, ces techniques sont « exploratoires ». Elles assistent le lecteur sans 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…

Dans le cadre restreint de cette communication, il n'était pas possible d'entrer dans le détail des discussions autour de l'analyse des séries chronologiques. Rappelons que ce type d'analyses rencontre un certain nombre de problèmes classiques :

- la robustesse des procédures – c'est-à-dire le degré de sensibilité à une modification marginale de certaines données ou à l'introduction d'une donnée aberrante – problème qu'il n'a pas été possible d'aborder ici ;
- la difficile appréciation des risques de première et de deuxième espèces : vaut-il mieux passer à côté d'une rupture réelle dans un corpus où découper celui-ci à tort ?
- par construction, le résultat dépend de l'ordre des données. En effet, l'algorithme ne s'applique qu'aux séries « ordonnées », comme les données chronologiques ou l'évolution du vocabulaire dans une œuvre.

Enfin, la constitution de vastes corpus – comme celui du discours politique québécois et canadien – est une nécessité pour le développement de l'analyse du discours. À ce propos, il faut souligner que ces corpus ne pourront être utilisables que s'ils répondent à certains critères formels. Il faut notamment que les graphies aient été soigneusement standardisées et, pour le français, que tous les mots aient été lemmatisés. A ces deux conditions, la statistique appliquée au langage pourra offrir des outils précieux non seulement pour les historiens et les politistes mais aussi pour les linguistes.

## Remerciements

Nous remercions Aline Korban qui a réalisé la numérisation des discours de 1867 à 1936, Cyril Labbé (Université de Grenoble I) et Jacques Savoy (Université de Neuchâtel) qui ont relu une première version de ce texte et dont les remarques nous ont été très utiles.

## Références

Bourque G. et Duchastel J. (1988). *Restons traditionnels et poigressifs.* Montréal : Boréal.

Bourque G., Duchastel J. and Beauchemin J. (1994). *La société libérale duplessiste, 1944-1960*. Montréal : Presses de l'Université de Montréal.

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

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

Hubert P., Labbé C. and Labbé D. (2002). Segmentation automatique des corpus. Voyages de l'autre côté de J.-M. Le Clézio. In JADT2002, Saint-Malo 13-15 mars, Rennes : IRISA-INRIA, Tome 1, pp. 359-369.

Labbé D. and Monière D. (2003). *Le vocabulaire gouvernemental. Canada, Québec, France (1945-2000).* Paris : Champion.

Labbé C., Labbé D. and Hubert P. (2004). Automatic Segmentation of Texts and Corpora. *Journal of Quantitative Linguistics*, 11-3 : 193-213.

Labbé D. and Monière D. (2008). *Les mots qui nous gouvernent*. Montréal : Monière-Wollank Editeurs.

Labbé D. and Monière D. (2009). Maurice Duplessis orateur : vocabulaire, style et axes de communication du chef de l'Union nationale. In Monière, D., editor, *Maurice Duplessis vous parle. Discours recueillis et présentés par Denis Monière*. Québec : Société du patrimoine politique du Québec, pp. 217-234.

Lafon P. (1984). *Dépouillements et statistiques en lexicométrie.* Genève-Paris : Slatkine-Champion.

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

Minoux M. (1983). *Programmation mathématique : Théorie et Algorithmes*. Tome 2. Paris : Dunod.

Monière D., Labbé C. and Labbé D. (2005). Les particularités d'un discours politique : les gouvernements minoritaires de Pierre Trudeau et de Paul Martin au Canada. *Corpus*, 4 : 79-104.

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

Müller D. (2002). Computing the Type Token Relation from the *A Priori* Distribution of Types. *Journal of Quantitative Linguistics*, 9-3 : 193-214.

Wimmer G. and Altmann G. (1999). Review Article: On Vocabulary Richness. *Journal of Quantitative Linguistics*, 6-1: 1-9.