



HAL
open science

La diachronie dans le discours politique. Le général de Gaulle

Cyril Labbé, Dominique Labbé

► **To cite this version:**

Cyril Labbé, Dominique Labbé. La diachronie dans le discours politique. Le général de Gaulle. David BANKS. Aspects diachroniques du texte de spécialité, L'Harmattan, pp.129-148, 2010. halshs-00117147

HAL Id: halshs-00117147

<https://shs.hal.science/halshs-00117147>

Submitted on 30 Nov 2006

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Université de Bretagne occidentale
Faculté des Lettres et Sciences sociales
Equipe de Recherche en Linguistique Appliquée

Aspects diachronique du texte de spécialité (Brest 17-18 novembre 2006)

La diachronie dans le discours politique **Le général de Gaulle**

Cyril LABBE
Université Grenoble I
cyril.labbe@imag.fr

Dominique LABBE
Institut d'Etudes Politiques de Grenoble
dominique.labbe@iep.upmf-grenoble.fr

Résumé :

Après avoir rappelé les règles de normalisation et de lemmatisation nécessaires pour le traitement informatique des textes en langue française, on présente une méthode capable de localiser avec précision les ruptures thématiques dans un texte ou un corpus. Une fois les sous-parties délimitées, on établit le vocabulaire et les thèmes caractéristiques de chacune d'elles. La méthode est d'abord appliquée à l'entretien accordé à M. Droit par le général de Gaulle entre les deux tours de l'élection présidentielle de 1965 puis à l'ensemble des interventions radiotélévisées du Général entre 1958 et 1969.

Mots clefs : France - Statistique lexicale - Discours politique - Vocabulaire politique - De Gaulle - Diachronie - Ruptures thématiques - vocabulaire caractéristique.

Trois couples de notions complémentaires - syntagme / paradigme ; synchronie / diachronie ; structure / histoire - renvoient aux trois niveaux d'appréhension possible de tout phénomène langagier. Le premier couple correspond à la micro-analyse qui ne va guère au-delà de la phrase ; le second est celui de la méso-analyse qui se situe au niveau d'un ou de quelques textes produits dans un court laps de temps ; le troisième celui de la macro-analyse qui envisage la langue ou les vocabulaires spécialisés et leurs évolutions à l'échelle historique.

En lexicologie, la notion de "diachronie" est devenue quasiment synonyme de "dimension temporelle" - et ceci quelle que soit l'échelle du temps en question - de telle sorte que cette notion s'applique à la fois aux niveaux méso et macro. Les communications acceptées par les organisateurs de ces journées montrent bien cette apparente ambiguïté de la "diachronie", même si la majorité d'entre elles se situent clairement au niveau "macro".

Au fond, tout phénomène langagier possède une dimension diachronique dès qu'il s'étend sur une certaine durée. Dès lors, pourquoi les mêmes outils ne pourraient-ils pas s'appliquer aux vastes corpus qui couvrent des décennies, voire des siècles, et à des textes ou des discours uniques quand leur étendue dépasse ce que l'analyse syntagmatique est capable d'embrasser ?

Dans les grands corpus, la dimension diachronique peut être étudiée de manière scientifique à l'aide d'outils performants. Notre communication présente l'un de ces outils : l'observation de l'accroissement du vocabulaire à l'aide de deux exemples.

D'une part, on utilise un texte singulier dans le discours politique français contemporain : l'entretien accordé par le général de Gaulle au journaliste Michel Droit en décembre 1965 (De Gaulle, 1970, p 412-440). Cet entretien a été réalisé en une seule prise d'une heure et demie et il a été diffusé en trois parties égales les 14 -15 et 16 décembre 1965, avec seulement deux brèves coupures (Lacouture, 1986, III, p 638-640). Ce texte comporte au total 9 390 mots. Il présente toutes les caractéristiques d'un oral improvisé et d'un "discours de combat" (Labbé, 2002).

D'autre part, l'analyse portera sur l'ensemble des 79 interventions radiotélévisées du général de Gaulle entre juin 1958 et avril 1969, soit 201 827 mots.

Nous allons tenter de répondre successivement à deux questions :

1. durant l'heure et demie de l'entretien de décembre 1965, est-il possible de repérer des continuités, des césures, des moments forts ou, au contraire des passages à vide ?

2. les outils statistiques qui permettent de répondre à cette première question peuvent-ils aussi être appliqués à de vastes corpus dans lesquels le temps peut se compter en années, voire en décennies ?

1. METHODES

Les règles de constitution des corpus seront rappelées succinctement car elles conditionnent toute opération statistique de qualité. Puis l'on présentera une approche possible de la diachronie : l'accroissement du vocabulaire.

Normalisation, lemmatisation et constitution des corpus

En premier lieu, il faut rappeler que les textes électroniques (ouvrages, revues, presse et autres documents souvent récupérés sur internet...) ne sont pas directement utilisables pour constituer des corpus et nécessitent un certain nombre de traitements préalables.

D'une part, ces textes comportent des indications marginales - dates et autres références, titres, intertitres, questions des journalistes pour les entretiens, interruptions et incidents de séance pour les discours à l'assemblée, etc. - indications qu'il ne faut pas confondre avec le texte proprement dit. Des "balises" serviront à isoler tout ce "para-texte" du texte (tableau I). On utilise pour cela les signes inférieur à (<) et supérieur à (>) qui se sont progressivement imposés, notamment dans les standards type XML. Ainsi peut-on conserver les informations intéressantes tout en évitant qu'elles viennent "polluer" le texte.

Tableau 1.

Le début de l'entretien du général de Gaulle avec M. Droit (décembre 1965)

<55. Général de Gaulle - Interview radiotélévisée avec Michel Droit - 14 décembre 1965>
<Texte électronique établi par Dominique Labbé (IEP Grenoble) 1990>

<Question : Michel Droit>

[Vos] adversaires de droite [vous] ont reproché de faire une politique de gauche.

<Réponse : Général de Gaulle>

C'est pas la gauche, la France ! C'est pas la droite la France ! Prétendre faire la France avec une fraction est une erreur grave, représenter la France avec une fraction est une erreur impardonnable.

Ce balisage est évidemment important : on s'intéresse seulement aux propos du général de Gaulle et non à ceux de son faire-valoir !

D'autre part, ce texte balisé (et soigneusement corrigé) est traité par un logiciel de normalisation des graphies et de lemmatisation. Le résultat est présenté dans le tableau 2.

Tableau 2 L'étiquetage des textes

ce, pronom	être, verbe	pas, adverbe	le, article	gauche, substantif
c'	est	pas	la	gauche

La procédure commence par normaliser les graphies multiples d'un même mot. Elle réduit les majuscules initiales de phrases (ou de vers) et celles qui sont données aux "faux noms propres". De même, le logiciel invite à déployer les abréviations à chaque fois que possible (M. : monsieur, Marcel, Maurice... mètre ?). Il contrôle l'orthographe des noms propres et de certains mots communs. Enfin, les chiffres sont convertis en lettres.

L'ensemble des règles de normalisation et de lemmatisation des textes ont été présentées dans Labbé (1990).

Dans les fichiers, en première position, on trouve la graphie normalisée : réduction des majuscules des mots communs, des graphies multiples, des abréviations ; en seconde position : l'entrée de dictionnaire et, en troisième, la catégorie grammaticale.

L'exemple ci-dessus comporte cinq mots et cinq problèmes :

- ce : pronom ou déterminant (adjectif démonstratif) ?
- est : verbe être ou substantif (le point cardinal) ?
- pas : adverbe ou substantif ?
- le : pronom ou déterminant (article) ?
- gauche : adjectif, substantif féminin ou substantif masculin ?

Cela n'a rien d'exceptionnel. Dans tout texte en langue française, plus du tiers des mots posent des problèmes de ce genre...

La lemmatisation respecte les conventions lexicographiques et les habitudes des usagers du français. Elle est **complète** et **univoque** (tout mot reçoit une étiquette et une seule). Elle est **stable** : les mêmes conventions sont strictement appliquées du début à la fin de l'opération.

Elle est aussi **automatique** que possible afin d'éviter un recours trop fréquent à un opérateur qui pourrait être source d'erreurs. A condition de s'en tenir à une nomenclature synthétique (type dictionnaire de langue), en moyenne 99% des mots peuvent être lemmatisés automatiquement. Le "résidu" n'est pas négligeable. Pour la lemmatisation du corpus de Gaulle, cela représente tout de même plus de 2 000 cas non résolus ! Les principales difficultés résident dans les mots les plus usuels ("suis", "tout", "même"...) et dans le caractère idiomatique de la langue : toute "règle" a des exceptions...

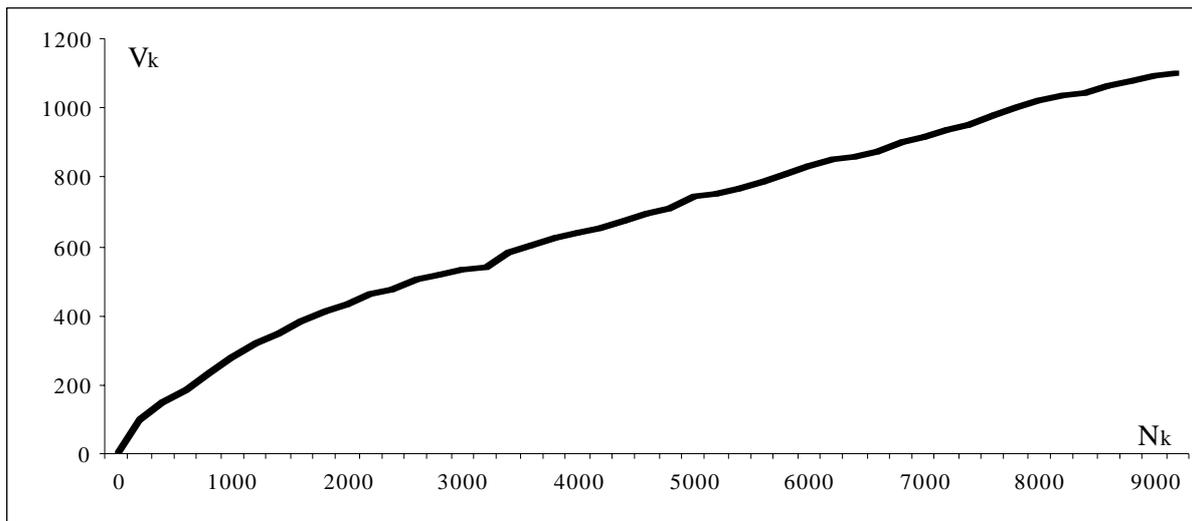
Enfin, la lemmatisation doit être **sans erreur** (du moins par rapport aux conventions retenues qui, elles-mêmes, devraient être entièrement explicites). C'est le principe le plus important. En effet, les étiquetages fantaisistes et lacunaires sont contre-productifs. Ils engendrent le scepticisme concernant l'ensemble des informations contenues dans la base (Hug, 2002).

Une fois le texte normalisé et lemmatisé, il est ajouté à la base existante. Ainsi peut-on, par étapes, constituer de grandes bases et capitaliser les résultats obtenus à chaque étape, contrairement à ce qui se passe dans les logiciels classiques d'analyse de "données textuelles" où chaque nouveau traitement exige que l'on recommence, à nouveaux frais, toutes les opérations...

L'accroissement du vocabulaire

La procédure consiste à ranger les textes par ordre chronologique pour former un corpus "ordonné" de taille N mots puis à découper ce corpus en K tranches - ici la taille des tranches est de 200 mots - et à observer, à la fin de chaque tranche, le nombre des V_k vocables différents apparus depuis le début du texte (tableau 3).

Tableau 3
Croissance du vocabulaire dans l'entretien de C. de Gaulle de décembre 1965
Nombre de vocables nouveaux par tranches de 200 mots.



Quelle que soit la longueur des textes ou des corpus, on obtient toujours ce profil assez particulier qui suggère la combinaison de trois phénomènes :

- le début de la courbe indique l'existence d'une tendance curviligne qui s'atténue progressivement, comme, par exemple, une exponentielle dont le multiplicateur serait inférieur à 1 (la pente d'une telle courbe diminue progressivement et tend vers l'horizontale),
- une tendance linéaire qui devient dominante vers 2 500 - 3 000 mots avec un coefficient lui aussi très inférieur à 1 (assez proche de l'horizontale),
- des petits accidents qui donnent au graphe un aspect "gondolé". Il existe donc une troisième dimension "accidentelle" ou "conjoncturelle".

Grâce aux deux premières dimensions, il est possible de calculer précisément le meilleur "ajustement" possible, c'est à dire la courbe "idéale" qui passe au plus près de tous les points. Cette possibilité avait été suggérée par Muller (1977).

Notre calcul combine donc :

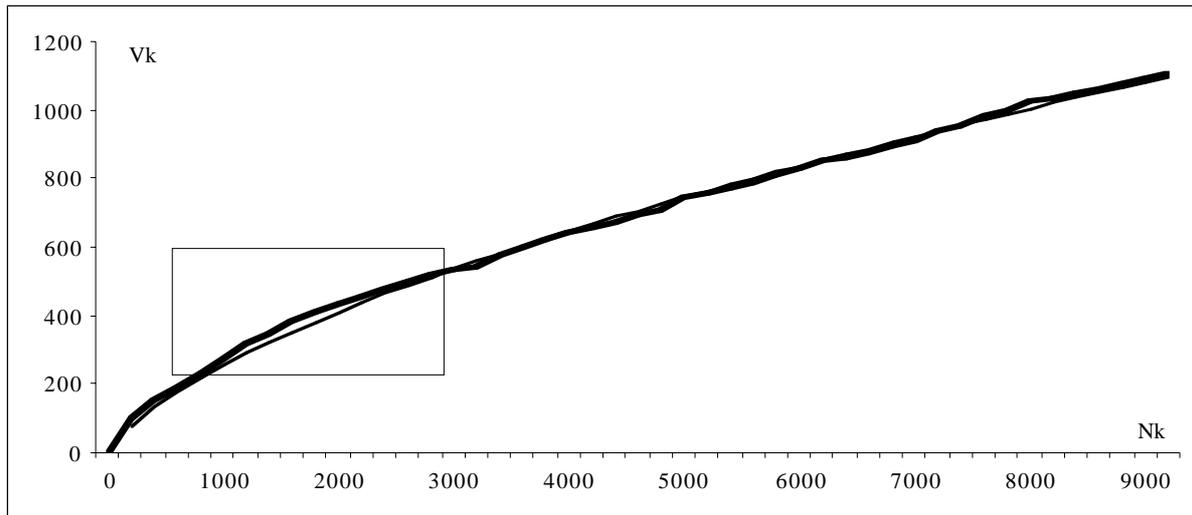
- un ajustement curviligne : une partie du texte est considéré comme le résultat d'un tirage exhaustif (sans remise) parmi les mots du texte issus du vocabulaire général de l'auteur (vocabulaire qu'il utilise quelles que soient les circonstances et les thèmes traités). Cette composante de l'accroissement du vocabulaire est décrite par la loi hypergéométrique ;

- une régression linéaire rend compte de l'apparition des mots du texte qui sont tirés d'un ou de plusieurs "vocabulaires spécialisés" (que l'auteur mobilise en fonction des circonstances et des thèmes qu'il aborde).

Les calculs et leurs propriétés sont présentés et discutés dans : Labbé et Al.(2004). Le modèle proprement dit a été présenté pour la première fois dans : Hubert et Labbé (1988).

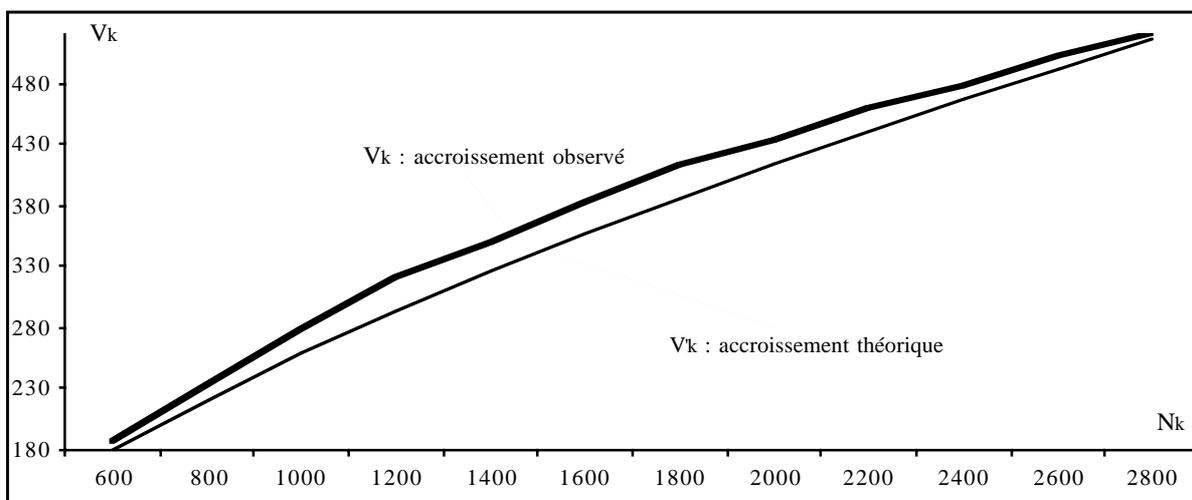
Appliqué à l'entretien du général de Gaulle avec M. Droit, ce calcul permet de tracer la courbe "ajustée" (trait maigre dans le tableau 4 ci-dessous). Cette courbe dessine un profil idéal qui passe au plus près de tous les points formant la courbe observée (trait gras).

Tableau 4 Croissance du vocabulaire dans l'entretien de C. de Gaulle de décembre 1965. Nombre de vocables nouveaux par tranches de 200 mots, valeurs observées (trait gras) et valeurs théoriques (traits maigres)



On pourrait presque dire que cet ajustement est "trop" parfait puisque les deux courbes se distinguent assez malaisément. Certes, il est possible d'utiliser une sorte de "zoom" sur les sections intéressantes. Par exemple, dans le graphique 5 ci-dessus, pour la section du texte comprise entre 600 mots et 3000 mots, les données observées (V_k) excèdent nettement ce que laisserait attendre une croissance régulière du vocabulaire (variable notée V'_k).

Tableau 5 Croissance du vocabulaire dans l'entretien de C. de Gaulle de décembre 1965. "Zoom" sur les tranches 600 à 3000 mots



Ici les deux traits se distinguent clairement mais une question essentielle demeure : quelle signification accorder à ces écarts entre la "ligne idéale" et les données observées ?

Pour répondre à cette question, on utilise un raisonnement courant pour analyser la conjoncture économique ou les phénomènes climatiques. Imaginons qu'il s'agit non pas de mots mais de températures. On veut savoir si un jour k , il fait un temps normal ou non pour la saison ? La température observée est comparée avec la température moyenne habituelle pour ce jour de l'année et l'on soustrait l'une de l'autre pour savoir si nous sommes au dessus, en dessous et de combien de degrés. **La variable est centrée sur la valeur théorique** ou "normale" ($V_k - V'_k$).

Autrement dit, la courbe "idéale" du graphique 4 passe à l'horizontale et les valeurs observées se répartissent autour d'elle. Si l'observation est supérieure à ce qui est attendu, c'est qu'en ce point, il est apparu un excédent de mots nouveaux par rapport au rythme moyen. L'orateur aborde probablement un sujet nouveau. A l'inverse, un point inférieur au rythme moyen signalerait un épuisement relatif du vocabulaire : l'orateur redit des choses qui ont déjà été énoncées auparavant.

Mais il surgit alors une seconde question : l'écart est-il exceptionnel ou non ?

Pour émettre ce deuxième jugement, il faut disposer d'une mesure de la variabilité "normale" des températures (ou, ici, du rythme d'apparition d'un mot nouveau au cours d'un texte). En effet, tout phénomène naturel (et le discours en est un) se trouve soumis à des fluctuations aléatoires qu'il faut distinguer des perturbations significatives.

Quand on dispose d'une longue série d'observations, on peut calculer empiriquement une plage de variation "normale". Sinon, la loi hypergéométrique permet de calculer cette plage de variation "normale". La mesure de ces variations s'appelle "écart type" ou "déviation standard" (noté par la lettre grecque σ). Les deux tiers des valeurs d'une variable gouvernée par le seul hasard (dite "variable aléatoire") sont situées à moins d'un écart type au-dessous ou au-dessus de la valeur moyenne ($\pm 1\sigma$) ; 95% de ces valeurs se situent à $\pm 2\sigma$; 99% à $\pm 3\sigma$.

Autrement dit, lorsqu'une observation s'écarte de la température attendue de $\pm 2\sigma$, le météorologiste peut conclure, avec moins de 5% de chances de se tromper, qu'il fait anormalement chaud ou froid pour la saison. De même, il aura moins de 1% de chances de se tromper si l'observation s'écarte de la valeur attendue d'au moins trois écarts types, etc.

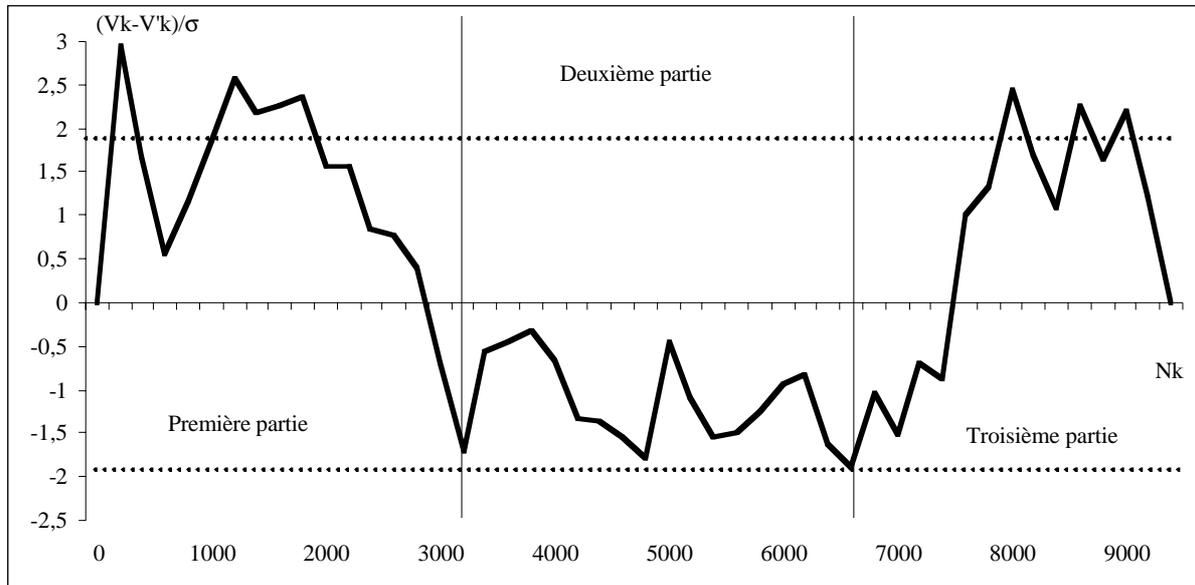
Pour faciliter la lecture des données, l'opérateur peut diviser la valeur centrée par l'écart type. Il obtient alors une **variable centrée et réduite** ou "normée".

Si le résultat est supérieur à $+ 2\sigma$ ou inférieur à $- 2\sigma$, l'opérateur sait immédiatement qu'il peut conclure avec très peu de chances d'erreur et que cette conclusion sera d'autant plus assurée que le résultat s'écartera de 2 en plus ou en moins.

Enfin si les observations s'étendent pendant plusieurs jours, l'opérateur pourra disposer les valeurs observées, centrées et réduites, sur un diagramme dont l'axe horizontal figurera le déroulement du temps.

L'accroissement du vocabulaire est observé de la même manière. À espace régulier, le logiciel calcule le nombre de mots nouveaux apparus puis il centre et réduit ces observations. Le graphique 6 donne le résultat de cette procédure appliquée à l'entretien du général de Gaulle de 1965.

Tableau 6
Croissance du vocabulaire dans l'entretien de C. de Gaulle de décembre 1965
(nombre de vocables nouveaux par tranches de 200 mots, variable centrée et réduite)



L'axe vertical est gradué en écarts types. La courbe théorique devient l'axe horizontal. L'écart entre les valeurs attendues et les valeurs observées ($V_k - V'_k$) est mesuré à la fin de chaque kième tranche de 200 mots et rapporté à l'écart type. Les seuils des deux écarts types sont figurés par des traits pointillés. La courbe étant contrainte aux deux extrémités - avant de commencer, la variable est nulle ; à la fin, elle est strictement égale au vocabulaire total - le premier et le dernier points de la courbe se trouvent donc sur l'axe.

Cette courbe peut être lue de deux manières complémentaires.

1. l'orientation des segments de courbe.

Une pente négative signifie qu'il apparaît "trop peu" de mots nouveaux : le thème a déjà été traité auparavant. À l'inverse, une pente positive signale un afflux de mots nouveaux supérieur à ce qui est attendu. Si le phénomène se prolonge sur plusieurs segments, on peut conclure à un renouvellement thématique, surtout si cette croissance succède à une portion de courbe négative. Dans le graphique ci-dessus, le phénomène est particulièrement visible au tout début de l'entretien (le général avait visiblement pensé son introduction) puis entre 600 et 1 400 mots et 7 400 et 8 000 (soit respectivement quatre et trois segments consécutifs dont la pente est fortement positive). A titre d'illustrations, ces deux passages sont reproduits en annexes 1 et 2. Ils sont effectivement très caractéristiques de cet entretien comme on va le voir.

Dans le graphique ci-dessus nous avons indiqué, par des barres horizontales les deux principaux points tournants tels qu'ils sont délimités à l'aide des ruptures de pente les plus importantes (3 200 mots et 6 600 mots), ce qui délimite trois parties dont la seconde apparaît comme singulièrement "pauvre" en apport de vocabulaire nouveau par rapport à la première et la dernière. On remarque immédiatement que ces coupures se situent exactement au tiers et au deux tiers de l'entretien. Or dès le départ, il avait été prévu que l'entretien serait diffusé en trois épisodes. Le "timing" a donc été scrupuleusement respecté. Dans ses fonctions

officielles, comme dans sa vie privée, le Général suivait toujours son emploi du temps quasiment à la minute près. Il n'empêche que l'entretien avait été quelque peu préparé, contrairement à ce que raconte la légende.

2. La position des points autour de l'axe

Les traits pointillés autour de l'axe horizontal, indiquent la plage de fluctuation entre $\pm 2\sigma$ que l'on peut considérer comme "normale". Les points situés en dehors de cette plage s'écartent significativement des valeurs attendues. En dessous, le discours politique se renouvelle moins que prévu. Au dessus de cette zone, on se trouve dans une période de renouvellement caractéristique.

De ce point de vue, le graphique 3 appelle une remarque évidente : il y a fort peu de points en dehors de la plage pointillée. C'est d'autant plus surprenant qu'il s'agit d'un entretien improvisé. On attendrait donc des hauts et des bas plus importants. En fait, durant les onze ans qu'il a été au pouvoir, les discours du Général ont tous été ainsi caractérisés par une régularité fondamentale.

II. DE LA DIACHRONIE A L'HISTOIRE

Naturellement, la méthode permet seulement de localiser les périodes de renouveau et les périodes de continuité. Elle n'indique pas quel est le contenu de ces changements ou de ces continuités. D'autres méthodes permettent de retrouver ce contenu. Nous allons les évoquer succinctement avant d'étendre la méthode à des intervalles de temps plus longs.

Le vocabulaire caractéristique des trois parties de l'entretien

Le raisonnement est le même que pour l'accroissement du vocabulaire : un vocable appartient au **vocabulaire caractéristique** d'un sous-corpus lorsque sa fréquence relative dans ce sous-ensemble s'écarte significativement, par le haut ou par le bas, de la fréquence observée dans l'ensemble du corpus. Dans la partie étudiée, si cette densité se situe au dessus de la plage de variation normale autour de la valeur attendue, on dit que le vocable est sur-employé (S+) ; dans le sens contraire, le mot est sous-employé (S-). En revanche, si la fréquence observée ne sort pas de cette plage de variation "normale", dans aucune des parties du corpus, on ne devrait rien conclure... mais l'on peut, par convention supposer que ce vocable est "non-caractéristique" (sur l'ensemble de la méthode : Labbé et Labbé 2005 ; pour le calcul proprement dit : Labbé et Labbé 1994 ; pour une application récente au discours politique : Monière et Al. 2005).

Le tableau 7 ci-dessous récapitule les mots significativement sur ou sous-employés dans la première partie par rapport aux deux autres.

Dans ce premier tiers de l'entretien, le destinataire est clairement identifié (les *Français*) et il s'agit de leur *niveau de vie*, des *problèmes sociaux* et *agricoles*, du *logement*, des *revenus* et de *l'enseignement*. Le discours privilégie la troisième personne impersonnelle ("*il faut*" : *augmenter, produire, faire...*). Sa tonalité dominante est nettement pédagogique (utilisation du démonstratif "ce... que"). Enfin, notons la forte fréquence de l'adverbe "plus" et du syntagme : *depuis 7 ans*. En effet, le Général est président depuis 7 ans. Tout son discours

consiste à dire : *Voilà ce qui a été fait depuis 7 ans dans le domaine économique et social et il faut faire plus...*

Tableau 7

Vocabulaire caractéristique de la première partie de l'entretien de décembre 1965

Vocables significativement suremployés au seuil de 5%
(Classement par catégories grammaticales et spécificité décroissante)

Noms propres : Français,
Verbes : augmenter, falloir, donner, produire, être, faire,
Substantifs : an, problème, vie, question, niveau, plan, logement, enseignement, revenu,
Adjectifs : social, vrai, agricole,
Pronoms : ce, il, que,
Adverbes : plus,
Déterminants : cinq, son, mon, notre, sept, cent,
Conjonctions et prépositions : pour, par, dans, depuis,

Vocables significativement sousemployés au seuil de 5%
(Classement par catégories grammaticales et spécificité décroissante)

Noms propres : Europe
Verbes : avoir,
Substantifs : pays, parti, république, régime, marché
Adjectifs : commun,
Pronoms : cela, nous,
Adverbes : là, alors, ne, pas,
Déterminants : même, le,
Conjonctions et prépositions : et, comme, à, entre,

Les vocables sous-employés dessinent "en creux" les thèmes peu ou pas traités : *l'Europe* (et le *marché commun*), la *République*, le *régime des partis*.

La seconde partie est dominée par *l'Europe* et *l'Allemagne*, le *marché commun*, la *coopération* et la *paix*. Il est donc question de politique internationale. Cette partie est dominée par le pronom "nous" (le vocable le plus significativement sur-employé dans cette seconde partie).

Enfin, la troisième partie est dominée par l'opposition entre le *régime des partis* et la *république* : ce sont les trois vocables les plus sur-employés, suivis du pronom "je", de *constitution* et de *peuple*. Il est donc question de politique intérieure...

Au passage, l'attitude du Général est bien résumée par le jeu des pronoms personnels : les problèmes économiques et sociaux sont pris en charge par l'impersonnel "il", les relations internationales par "nous" et la politique intérieure par "je". Tout au long des onze ans qu'il a passé à la tête de la France, le Général a utilisé le même système, notamment l'utilisation privilégié du "nous" quand il parlait de politique internationale et du "je" pour la politique intérieure (Labbé, 1997).

Naturellement, on peut juger que ces listes sont un peu abstraites et souhaiter les illustrer par des citations comme dans les dictionnaires afin d'illustrer le sens d'un mot. La procédure de sélection de ces phrases les plus caractéristiques est la suivante : après avoir

établi le vocabulaire caractéristique du texte ou de la partie de texte analysée, le logiciel relit ce texte ou cette partie, en recherchant les phrases contenant le plus grand nombre de mots sur-employés et le plus petit nombre de mots significativement sous-employés (le score est égal au nombre de vocables "S+" contenus dans la phrase diminué du nombre de vocables "S"). On peut lire ces phrases comme autant d'illustrations, voire de résumés des principaux thèmes privilégiés lors de l'échange considéré entre le général et le journaliste (Tableau 8).

Tableau 8
Phrases les plus caractéristiques de la première partie

"Je répète : depuis 7 ans le niveau de vie des Français a augmenté de 3.7% par an et d'après ce qui est décidé par notre cinquième plan et ce qui est devenu possible, grâce à nos progrès économiques, le niveau de vie des Français va monter à 4% par an" (score 17)

"C'est pourquoi je demande que l'on considère ce qu'il en était avant mon retour : les prix montaient, depuis 1946, jusqu'en 1958, automatiquement, en moyenne de 10% par an, et tout ce qu'on faisait pour les salaires, pour les traitements, et aussi pour les revenus paysans, était dévoré à mesure par cette augmentation" (score 15)

"Il faut beaucoup plus de logements, beaucoup plus d'autoroutes, beaucoup plus de téléphones, il faut un enseignement encore bien mieux organisé avec beaucoup plus d'écoles, d'universités, etc." (score 13)

Ces courtes listes fournissent trois résumés commodes pour cette heure et demie d'entretien et les phrases les plus caractéristiques éclairent chacun des thèmes traités.

Certes, on savait déjà que cet entretien avait été découpé en trois parties égales et à quel thème chacune était consacrée, mais le fait que l'algorithme ait permis de les localiser exactement "à l'aveugle" prouve son efficacité. Il fournit d'ailleurs des informations non-triviales. Par exemple, l'afflux de vocabulaire est nettement plus important quand le Général parle des questions économiques et sociales ou de politique intérieure que lorsqu'il se consacre à la politique internationale, ce qui est une surprise tant on a présenté de Gaulle comme surtout préoccupé de la place de la France dans le monde et fort peu soucieux d'"intendance" !

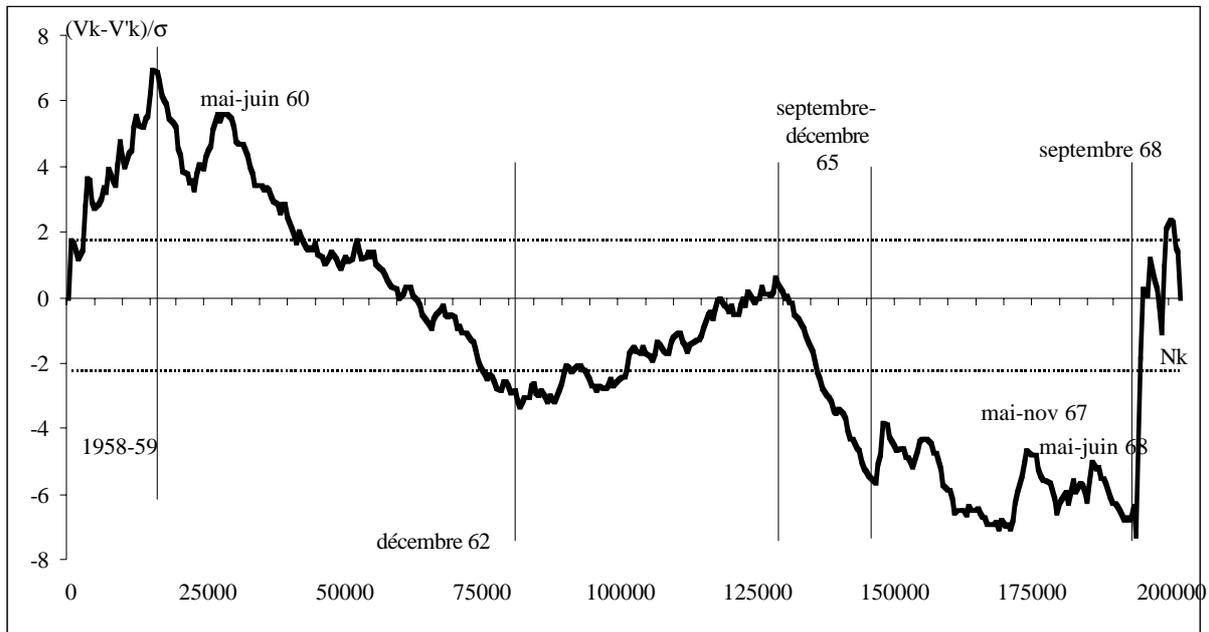
La méthode vient donc d'être appliquée avec succès à une sorte de "modèle réduit". Elle s'applique aussi fort bien à des corpus plus larges où la diachronie devient histoire.

Onze ans de pouvoir du Général : mai 1958 - avril 1969

Le général de Gaulle prononce son premier discours radio-télédiffusé le 13 juin 1958 et sa dernière apparition a lieu le 25 Avril 1969. Entre les deux dates, il a parlé 79 fois aux Français par l'intermédiaire de la radio et de la télévision (allocutions, conférences de presse, entretiens), soit 201 942 mots. Ces discours sont rangés par ordre chronologique, et la méthode présentée ci-dessus est appliquée à ce corpus. Elle permet d'isoler les principales ruptures (tableau 9).

Tableau 9

Croissance du vocabulaire dans le corpus total de C. de Gaulle 1958-69
(nombre de vocables nouveaux par tranches de 1000 mots, variable centrée et réduite)



Ce graphique appelle au moins deux observations :

- Les écarts sont beaucoup plus forts : la majorité des points se situent hors de la zone de variation normale de ± 2 écarts-types autour de la valeur attendue. Il en est toujours ainsi dans tout corpus comportant un nombre important de textes produits sur une longue période. Au contraire, par rapport aux autres hommes politiques français, les variations de l'accroissement du vocabulaire chez de Gaulle sont plutôt de relativement faible amplitude.

- il y a une tendance lourde à la décroissance de l'innovation entre le début de l'année 1960 et jusqu'après mai-juin 1968. Autrement dit, en dehors des premiers et des derniers mois de son mandat, le général de Gaulle a visiblement eu du mal à renouveler les thèmes de ses discours. Mais on peut aussi interpréter cela comme une fidélité à la mission qu'il s'était donnée une fois l'autodétermination de l'Algérie décidée : la stabilité des institutions, le développement économique et social de la France.

Toutefois, grâce à la méthode de découpage exposée ci-dessus, il est facile de distinguer trois grandes périodes, constituées chacune de deux phases successives :

1. 1958-1962 : 18 mois d'innovations et 3 ans de répétitions. La phase d'innovation initiale correspond à la mise en place des institutions de la Ve république puis au dévoilement progressif de l'autodétermination pour l'Algérie (dévoilement achevé dans l'allocution du 16 septembre 1959 et dans la conférence de presse du 10 novembre 1959). A partir de là, le Général semble se répéter. Cependant, deux allocutions apportent du neuf (mai et juin 1960) la première porte sur la croissance économique et le progrès social en France, la seconde sur la paix dans le monde et la première "conférence au sommet" des 4 superpuissances nucléaires à Paris en mai 1960. Les thèmes fondamentaux sont maintenant en place.

Contrairement à ce qu'on pourrait attendre, les événements d'Algérie (barricades d'Alger en janvier 1960, putsch des généraux d'avril 1961), ou la longue crise politique de l'automne

1962 n'apportent pas de renouvellement significatif. Mais, au fond, c'est assez logique puisque l'on sait déjà que le discours de combat utilise plutôt un vocabulaire simple et assez restreint.

2. *1963-1968* : trois ans d'innovations (à un niveau plus modeste que durant la première période) sont suivis d'un brusque effondrement lors de la campagne pour les élections présidentielles de 1965, spécialement lors de l'entretien qui a été analysé ci-dessus. En fait, le Général reprend des arguments déjà présentés, spécialement durant la première période. Là encore, il s'agit d'un discours de combat.

Après sa victoire de 1965, l'élan semble brisé. Aucun sursaut significatif se produit, même pas lors des législatives de mars 1967 (qui sont cependant suivies d'un très léger "pic" sans lendemain) ni lors des événements du printemps 1968 qui, "à chaud", ne génèrent pratiquement aucun renouvellement du vocabulaire.

3. *septembre 1968 - avril 1969*. Ces sept derniers mois sont marquées par sept interventions qui sont presque toutes le siège d'un renouvellement thématique. Ce renouvellement sera interrompu par la démission du Général après la victoire du "Non" lors du référendum d'avril 1969.

Pour chacune de ces périodes, le vocabulaire caractéristique sera isolé grâce à la méthode présentée ci-dessus qui permet également de retrouver les citations les plus illustratives et de caractériser ainsi les thèmes dominants de chacune de ces périodes.

Conclusions

L'intérêt essentiel de cette démarche réside dans le fait qu'elle ne fait intervenir aucune information extérieure au corpus, aucune hypothèse ad hoc, aucun découpage effectué arbitrairement en fonction d'évènements jugés a priori significatifs.

Au contraire, elle possède la première qualité que l'on peut attendre d'un outil scientifique, c'est-à-dire la **reproductibilité**. Toute personne appliquant la même procédure aux mêmes corpus aboutira à des résultats exactement semblables.

Cet outil a déjà été appliqué à un grand nombre de corpus, spécialement politiques. Par exemple, les discours des Premiers ministres français, canadiens et québécois depuis 1945 (Labbé et Monière, 2003) ou les discours du Premier ministre anglais Tony Blair (Arnold, 2005). Il pourra certainement s'appliquer à des corpus de plusieurs dizaines de millions de mots s'étendant sur plusieurs décennies voire plusieurs siècles. Il suffira pour cela que tous les textes aient été balisés, normalisés et lemmatisés en suivant rigoureusement les mêmes standards.

Naturellement, une analyse diachronique implique deux autres conditions :

- que le classement des textes se fasse dans l'ordre chronologique, ce qui peut poser quelques problèmes pour les textes les plus anciens dont la datation n'est pas toujours certaine,

- que le corpus soit relativement homogène du point de vue du genre, surtout s'il provient d'un grand nombre d'auteurs. Naturellement, les corpus présentés lors de ces journées présentent effectivement cette homogénéité. Mais il faut se souvenir que la plupart des grandes bases de "données textuelles" ne remplissent pas cette condition. Dans ce cas, d'autres outils, comme le calcul des distances intertextuelles combinées à la classification automatique aident à repérer des sous-ensembles homogènes.

Il existe aujourd'hui un grand nombre de collections de textes sous forme électronique : collections de journaux, de périodiques, de revues scientifiques, de textes juridiques, procès-verbaux des assemblées parlementaires, etc. Naturellement, il reste beaucoup de travail pour transformer ces collections en véritables corpus, c'est-à-dire en bases de données informatiques. Mais l'obstacle essentiel est ailleurs. Les lexicologues ignorent qu'il existe des outils informatiques et statistiques puissants, capables de traiter l'énorme masse d'informations contenues dans ces bases, pour les aider à retrouver les données pertinentes pour leurs propres recherches...

Références

- Arnold, Edward (2005) : "Le discours de Tony Blair", *Corpus*, 4, 2005, p 55-78..
- Gaulle, Charles de (1970) : *Discours et messages. Pour l'effort*, Paris, Plon, tome IV.
- Hubert, Pierre et Dominique Labbé (1988) : "Un modèle de partition du vocabulaire", in Dominique Labbé, D., Serant, D. et Thoiron, P. (eds.) : *Etudes sur la richesse et la structure lexicales*. Paris-Genève : Champion-Slatkine, p. 93-114.
- Hug, Marc (2002). "Désambiguïsation automatique d'homographes verbe/nom" in Morin, A. et Sébillot, P. (eds), *VIe Journées Internationales d'Analyse des Données Textuelles (Saint-Malo 13-15 mars 2002)*, Rennes : IRISA-INRIA, 2002, vol. 1, p 371-379.
- Labbé, Cyril et Dominique Labbé (1994) : *Que mesure la spécificité du vocabulaire ?*. Grenoble : CERAT. Reproduit dans *Lexicometrica*, 3, 2001.
- Labbé, Cyril, Dominique Labbé et Pierre Hubert (2004) : "Automatic Segmentation of Texts and Corpora". *Journal of Quantitative Linguistics*, December 2004, 11-3. p 193-213.
- Labbé, Cyril et Dominique Labbé (2005) : "How to measure the meanings of words ? Amour in Corneille's work", *Langage Resources Evaluation*, 2005, 39, p 335-351.
- Labbé, Dominique (1990) : *Normes de saisie et de dépouillement des textes politiques*. Grenoble: Cahier du CERAT.
- Labbé, Dominique (1997) : "Le «nous» du général de Gaulle", Communication au colloque *La comunicazione politica : aspetti socio-linguistici e pragmatici*, Rome, Université La Sapienza, 9-10 mai 1997, 16 p. Publié dans *Quaderni di studi linguistici*, 4/5, 1998, p 331-354.
- Labbé, Dominique et Denis Monière (2003) : *Le discours gouvernemental. Canada, Québec, France (1945-2000)*, Paris : Champion.
- Labbé, Dominique (2002) : "Le général de Gaulle en campagne". Communication aux IIIe Journées de l'ERLA, Brest, 15-16 novembre 2002, 16 p. Reproduit dans : Banks, D. (ed.) : *Aspects linguistiques du texte de propagande*. Paris: L'Harmattan, 2005, p 213-233.
- Lacouture, Jean (1986) : *De Gaulle*, Paris, Le Seuil.
- Monière, Denis, Cyril Labbé et Dominique Labbé (2005) : "Les particularités d'un discours politique : les gouvernements minoritaires de Pierre Trudeau et de Paul Martin au Canada". *Corpus*, 4, 2005, p 79-104.
- Muller, Charles (1977) : *Principes et méthodes de statistique lexicale*, Paris, Hachette.

Annexe 1

Le premier passage apportant un afflux exceptionnel de mots nouveaux (première partie de l'entretien de 600 à 1400 mots)

Il y en a, ce sont les démagogues, tranchons le mot, il faut bien le dire, ce sont les partis, c'est-à-dire ceux qui caressent les clientèles pour être élus, et souvent les journaux pour être lus, alors ceux-là, assez volontiers, et même très volontiers, disent : il en faut plus pour un tel, il faut faire davantage pour les ouvriers, il faut plus pour les paysans, il faut faire mieux pour les cadres, il faut arroser davantage les fonctionnaires, les agents des services publics, il faut même donner un petit peu plus aux chefs d'entreprise, avec un peu d'inflation. Il faut beaucoup plus de logements, beaucoup plus d'autoroutes, beaucoup plus de téléphones, il faut un enseignement encore bien mieux organisé avec beaucoup plus d'écoles, d'universités, etc. Ca, c'est commode.

Si vous voulez dire que je ne suis pas sur ce plan-là, vous avez raison. Moi, je suis sur le plan que voici : il faut que le peuple français soit prospère. Il le faut, parce que, s'il n'est plus prospère, la France dont nous parlions tout à l'heure ne pourrait pas jouer son rôle dans le monde d'aujourd'hui donc rien ne m'occupe davantage, depuis bien longtemps, que la prospérité nationale. Je dirai même, puisque vous avez parlé d'intendance, que ça me fait un peu sourire, car, si vous y faites attention, rien, rien, rien de ce qui a été fait d'important, au point de vue économique et social depuis la libération, n'a été fait excepté par mon gouvernement.

Après la libération, c'était le point de vue social immédiat qui l'emportait. Il fallait que nous retrouvions, après cette secousse épouvantable, notre équilibre. Alors mon gouvernement a fait les assurances sociales, les allocations familiales, les nationalisations, les comités d'entreprise.

Aujourd'hui ou plus exactement depuis que je suis revenu, depuis 7 ans, c'est l'économie qui me paraît l'emporter sur tout le reste, parce qu'elle est la condition de tout, et en particulier la condition du progrès social. C'est pas la peine de nous raconter des histoires, si nous sommes pauvres ou en désordre, économiquement parlant, nous ne progresserons pas socialement : au contraire, si nous sommes prospères, et à condition bien entendu que l'Etat fasse son devoir, tous les Français en profiteront et en profiteront.

Le niveau de vie des Français ? Pas plus tard que samedi j'ai donné des chiffres. à mon avis ils sont éclatants. Je répète : depuis 7 ans le niveau de vie des Français a augmenté de 3,7% par an et d'après ce qui est décidé par notre cinquième Plan et ce qui est devenu possible, grâce à nos progrès économiques, le niveau de vie des Français va monter à 4% par an. Cela veut dire que 18 ans après mon retour, le niveau de vie des Français a doublé. Ce n'est pas si mal. Alors c'est comme le bonheur, ça existe par comparaison. C'est pourquoi je demande que l'on considère ce qu'il en était avant mon retour : les prix montaient, depuis 1946, jusqu'en 1958, automatiquement, en moyenne de 10% par an, et tout ce qu'on faisait pour les salaires, pour les traitements, et aussi pour les revenus paysans, était dévoré à mesure par cette augmentation. Il est parfaitement exact que tout en faisant monter le niveau de vie, et je le dis bien, le niveau de vie réel, eh bien ! il est parfaitement exact que mon gouvernement a stabilisé les prix, dans toute la mesure où, à notre époque, c'est possible. Au lieu que les prix augmentent de 10% par an, comme c'était le cas avant 1958, ils augmentent maintenant de 2,5% par an.

Il n'y a pas un pays d'Europe occidentale où les prix sont aussi limités dans leur ascension que chez nous. Il n'y en a qu'un dans le monde : ce sont les Etats-Unis. Pourquoi ? parce que les Etats-Unis, grâce à ce qu'on appelle le gold exchange standard, ont la facilité d'exporter leur inflation. Mais nous n'avons pas cette facilité. Il a donc fallu faire le plan de stabilisation. Je reconnais que ce n'était pas facile. Je reconnais qu'il en résulte des difficultés, notamment au point de vue de l'expansion immédiate, mais maintenant c'est fait, et par conséquent c'est sur cette base solide, stable, avec des budgets en équilibre, une monnaie qui ne bouge pas, et des prix, je vous le répète, qui, pratiquement, n'augmentent pour ainsi dire plus et qui, comme vous le savez, ne doivent pas augmenter l'année prochaine, de plus de 1,5% par an, c'est sur cette base là que notre niveau de vie s'élève et que tout le monde y gagne.

Annexe 2

Le deuxième passage apportant un afflux exceptionnel de mots nouveaux (troisième partie de 7 400 à 8 000 mots)

Or, le régime des partis, c'est la pagaille. Evidemment, on l'a vécu avant la première guerre mondiale, pendant longtemps cela allait cahin-caha. A ce moment là on ne risquait pas grand-chose. à l'intérieur, on était très riche : je ne parle pas de tous les Français, bien entendu, il s'en faut, mais je vous parle de l'ensemble de ce que l'on appelait "la société", qui était riche. Il y en a qui disaient que c'était la "belle époque". Bien sûr ! on ne se transformait pas, on n'évoluait pas ! D'autres devenaient de grands pays industriels, comme l'Allemagne, l'Angleterre, qui avaient commencé avant tout le monde, les Etats-Unis qui avaient entrepris leur essor. Nous, nous restions, cahin-caha, comme nous étions. Et puis alors, au dehors, on ne risquait pas grand-chose non plus, bien sûr ! Il y avait la menace allemande à l'horizon pour le cas où... Mais il y avait l'alliance russe, et puis, après, il y a eu l'entente cordiale, enfin, ça allait comme cela.

Puis, il y a eu 1914 : désastre initial, auquel nous avons échappé par une chance inouïe, par un sursaut du tréfonds national qui nous a permis de nous en tirer, dieu sait d'ailleurs avec quelles pertes ! et encore de nous en tirer en 1914-1915 grâce au pouvoir personnel du père Joffre, et puis en 1917-1918, à la fin où ça devenait dramatique, et infiniment grave de nouveau, grâce au pouvoir personnel de Clemenceau ! Mais, dans l'intervalle, qu'est ce qu'ils avaient fait, les partis ? Ils n'avaient rien fait ! Ils renversaient les ministères comme à l'habitude, ils renversaient le ministère Viviani, puis le premier ministère Painlevé, et allez, donc ! Voilà ce qu'ils faisaient les partis !

Entre les 2 guerres, après qu'on eu liquidé Clemenceau - que les partis eurent liquidé Clemenceau - il y a eu alors ce qu'on sait, c'est-à-dire un régime de médiocrité, un régime d'impuissance où le désastre se dessinait à l'horizon sans qu'on fit en réalité rien pour l'empêcher.

Entre monsieur Clemenceau et monsieur Paul Reynaud, c'est-à-dire de 1920 à 1940, on a eu 47 ministères en 20 ans. Voilà le régime des partis ! Alors, naturellement on a été battus, écrasés en 1940. On n'avait rien préparé, on était divisés par les partis, on n'avait pas les armes nécessaires.

Vous me parlez actuellement d'un candidat du front populaire. Hélas, moi, j'ai connu le front populaire. Je vous dirai même qu'à cette époque-là, c'était en 1936, j'avais quelques idées sur la nécessité de rénover notre défense nationale, et je crois bien que l'expérience a prouvé que, si l'on m'avait écouté, on aurait évité le désastre et on aurait probablement tué l'entreprise d'Hitler dans l'oeuf avant même qu'elle ne pût s'étendre. Enfin, j'espérais un peu que ce mouvement du front populaire - qui m'apparaissait être la nouveauté, la réforme, etc. - allait effectivement s'emparer aussi de la défense. Eh bien ! il m'a fallu bien déchanter, et finalement le régime du front populaire, la majorité du front populaire, la chambre du front populaire, le gouvernement du front populaire, cela a fini par le désastre de Sedan, l'abdication de la république et la capitulation devant l'ennemi. Du reste, je me hâte de vous dire que, si au lieu d'être ce gouvernement du front populaire, c'est-à-dire des partis du front populaire, ç'avait été le gouvernement des autres partis, des partis conservateurs, ç'aurait été probablement la même chose. Les partis ne peuvent pas conduire la France, c'est trop dur, et c'est pourquoi d'ailleurs, après mon retour, en 1945, quand les partis ont reparu - tous contre moi, bien entendu - moi parti, ils n'ont plus rien fait du tout, excepté 23 crises ministérielles, dont j'ai parlé samedi dernier, et puis c'est tout.