LA REPARTITION DU VOCABULAIRE

Cyril Labbé, Dominique Labbé

Laboratoire d'Informatique de Grenoble

RAPPORT

LA REPARTITION DU VOCABULAIRE

septembre 2017

Cyril Labbé
Laboratoire d'Informatique de Grenoble (Université Grenoble-Alpes)
(cyril.labbe@imag.fr)

Dominique Labbé
PACTE (CNRS – Université Grenoble-Alpes)
dominique.labbe@umrpacte.fr
http://www.pacte-grenoble.fr/membres/labbe-dominique/

Résumé
La répartition d'un mot dans une collection de textes (corpus) est l'ensemble des emplacements où ce vocable apparaît. Cette dimension a été peu étudiée et uniquement pour des corpus constitués d'échantillons de longueurs égales. Cette note analyse le phénomène dans les corpus de textes entiers (longueurs inégales) et propose un indice dont les propriétés sont décrites à l'aide de plusieurs corpus de grandes dimensions. Une procédure simple permet d'isoler les vocables les plus régulièrement utilisés et ceux qui sont localisés en un point du corpus. Cette dimension complète la fréquence et apporte une information supplémentaire sur le vocabulaire du corpus.

Abstract
The distribution of a word in a collection of texts (corpus) is the set of locations where this term appears. This dimension has been little studied and only for corpora constituted of excerpts of equal lengths. This note analyzes the phenomenon in the corpora of entire texts (the lengths of which are unequal) and proposes an index whose properties are described using several corpora of large dimensions. A simple procedure makes it possible to isolate the words most regularly used and those which are located at a point of the corpus. In relation with the frequencies, repartition index provides some additional informations about the vocabulary of a corpus.

Cette étude a bénéficié de l'aide de Pierre Hubert (Université de Paris VI).
Edouard Arnold (Université de Dublin), Monique Bécue (Université de Barcelone) et Denis Monière (Université de Montréal) ont relu des versions antérieures de cette note et ont apporté de nombreuses remarques très utiles.

LA REPARTITION DU VOCABULAIRE
(septembre 2017)

Les **corpus** sont de vastes collections de textes réunis pour connaître de manière empirique leur vocabulaire et de leur syntaxe (Habert et al.1997).

Dans ces collections, chaque "mot" est caractérisé par deux dimensions. La première est le nombre de fois où il apparaît. On parle d'occurrences, d'**effectifs** ou de **fréquence** (absolue ou relative). La seconde dimension est la **répartition** du mot : apparaît-il de manière uniforme dans la totalité du corpus ou au contraire est-il plutôt localisé dans quelques documents ?

La première dimension est familière bien qu'assez peu étudiée empiriquement (les principaux travaux sont évoqués au début de cette note). En revanche, la seconde question a été moins explorée malgré son évidence. Cette note lui est consacrée. Après une présentation des principaux travaux et la position du problème (première partie), la **localisation** des vocables à la surface d'un corpus sera étudiée de manière empirique (deuxième partie) et un indice sera proposé afin de mesurer synthétiquement la répartition (troisième partie).

I. Fréquences et répartitions des mots

Malgré la riche tradition lexicographique française, il y a assez peu de travaux empiriques sur l'usage des mots en Français. A notre connaissance, deux seulement portent sur le français oral (Gougenheim et al. 1956 ; Beauchemin et al. 1992). La plupart ont étudié la littérature : un seul auteur (Muller 1967 ; Bernet 1983), une collection d'œuvres littéraires ("Trésor de la langue française", Imbs 1971) ou d'extraits d'œuvres (Juilland et Al. 1970, Engwall 1984). Plus récemment, de nouveaux types de corpus sont apparus.

Plusieurs types de corpus

Plusieurs distinctions sont nécessaires.

Premièrement, les corpus sont des collections. Un texte unique comme un roman, une pièce de théâtre, un discours, etc. n'est pas un corpus. Les éléments réunis dans un corpus ont souvent un lien entre eux : œuvres d'un même auteur, articles d'une même rubrique ou d'un même journal, entretiens sur un thème précis. La question de la répartition s'est principalement posée à propos de ces corpus, même si la plupart des méthodes présentées ci-dessous peuvent *a fortiori* s'appliquer à un texte unique, dès qu'il a une longueur minimale de quelques dizaines de milliers de mots.

Deuxièmement, les corpus ne sont pas (seulement) des listes de mots. Il existe plusieurs listes de mots pour le français. Par exemple, *Lexique 3* mise en ligne par l'Université de Savoie (New et Pallier. 2005) est une liste de mots élaborée à partir de la base Frantext (issue du "Trésor de la langue française") auquel ont été ajoutés des sous-titres de films (pour le français "oral). Les corpus présentent un

certain nombre de traits supplémentaires, notamment des outils permettant de retrouver le contexte d'un mot grâce aux "concordances" ou collocations" (Sinclair 1991).

Troisièmement, il faut distinguer les corpus contenant des textes entiers (que l'on peut nommer "exhaustifs") de ceux constitués d'extraits de textes (ou "corpus d'échantillons").

Les corpus exhaustifs sont constitués d'une ou plusieurs œuvres littéraires, d'articles de presse, des discours d'un homme politique, etc. Outre l'intégralité des textes (il ne s'agit pas d'extraits arbitraires), ces corpus ont deux caractéristiques notables. Premièrement, les textes qui les composent ont un ordre habituellement déterminé par leur date de composition ou de parution. On peut parler corpus "chronologiques" ou "corpus ordonnés". Seconde caractéristique : tous les textes ont des longueurs différentes.

Les corpus constitués d'extraits sélectionnés de manière plus ou moins arbitraires, ont habituellement pour but de constituer des échantillons de la langue générale ou de telle ou telle spécialité. Ici l'ordre de classement des textes est arbitraire.

Pour le français, le premier corpus de ce second type a été réalisé par G. Gougenheim et ses collaborateurs dans les années 1950, pour faciliter l'enseignement du français langue seconde. Il s'agissait de retranscriptions d'entretiens avec des locuteurs français natifs comptant au total 300 000 occurrences. Mais, du fait de l'absence d'outils informatiques à cette époque, l'opération a abouti à un dictionnaire et non à un corpus. L'autre échantillon du français oral est québécois. Réalisé par Beauchemin et Al (1992), il contient un million d'occurrences et 11 327 vocables différents. Mais là encore, seule la liste des vocables est consultable. Cette liste associe à chaque vocable un indice de dispersion et un indice d'usage sur lesquels nous revenons plus bas.

Avec les ordinateurs, les corpus d'échantillons ont pu atteindre des dimensions plus importantes. Le plus connu est le British National Corpus (Burnard 2007) créé en 1991 et contenant 100 millions de mots étiquetés. Les principales langues sont pourvues de corpus de ce type. Pour le français il existe plusieurs dictionnaires en ligne mais, à notre connaissance, un seul "corpus" (à l'Université de Liepzig : Eckart & Al. 2013). Cette base contient environ 1 500 millions d'occurrences du français collectées sur internet mais le dépouillement est fait sur les formes graphiques et n'aboutit qu'à des listes. Ces listes s'accompagnent de nombreuses informations statistiques mais aucune ne porte sur la répartition des "mots".

En effet, il est nécessaire de distinguer les collections de "formes graphiques" – comme celle de l'Université de Liepzig - des corpus, au sens propre du terme, dont le British National Corpus est le modèle. Dans les dépouillements sur "formes graphiques", *L', Le, La, Les, l', la, le, les* sont des unités différentes. De plus, *la* (article) n'est pas distingué de *la* (pronom) ou du *la* (nom masculin note de musique). En revanche, dans un corpus, chaque occurrence est dotée d'une étiquette indiquant son entrée de dictionnaire – par exemple, l'infinitif auquel sont

rattachées les différentes flexions d'un verbe – et sa catégorie grammaticale (Habert et Al. 1997).

Les corpus étiquetés sont homogènes et comparables entre eux, ce qui n'est pas le cas des collections de formes graphiques. Par exemple, dans l'œuvre de Marcel Proust (1871-1922), soit 1,386 millions de mots, les deux substantifs les plus fréquents sont "madame" et "monsieur", la plupart du temps écrits sous forme abrégée (Mme, M.) mais pas toujours. Cela fait donc quatre formes graphiques différentes – et même dix puisqu'on rencontre aussi : "mesdames", "messieurs", "MM.", "Mmes", "Madame" et "Monsieur", notamment en début de phrase. Cette instabilité graphique complique les statistiques et la comparaison entre textes et entre auteurs. Mais surtout, les dépouillements en formes graphiques ne permettent pas de réaliser des concordances et des collocations exhaustives qui sont les outils de base de la lexicologie (Sinclair 1991).

Pour atteindre cette exhaustivité, on attache à chaque mot du corpus une étiquette comportant sa graphie normalisée (ici *madame* ou *monsieur*), son "lemme" ou entrée de dictionnaire – par exemple, l'infinitif du verbe pour toutes ses flexions - et sa catégorie grammaticale (pour les conventions suivies : Labbé 1990).

Toutes les expériences présentées dans ce rapport sont réalisées sur des corpus exhaustifs étiquetés – comptant au total 52 millions de mots à la date de rédaction de cette note - et les statistiques portent sur les vocables et non les "formes graphiques".

Jusqu'à maintenant, l'étude du français a peu fait appel à ces corpus (Cobb et Al. 2004) et la question de la répartition reste largement inexplorée, sauf dans les corpus d'échantillons.

Les indices de répartition dans les corpus d'échantillons

Pour les corpus d'échantillons, un consensus s'est établi pour utiliser comme mesure de la répartition, "le nombre de textes (ou de tranches) de ce corpus où ce vocable est attesté" (Engwall 1984, p. XXXVIII). Ainsi P. Lafon propose que l'"indice de répartition" "décompte les parties dans lesquelles la forme est présente" (Lafon 1983, p. 51. Voir également Lafon 1980).

Naturellement, cette convention n'a de sens que si toutes les parties du corpus ont la même longueur.

Par exemple, le corpus du français québécois se divise en dix tranches égales pour les différentes régions du pays. La répartition d'un vocable est le nombre de tranches dans lesquelles il apparaît. Afin de pouvoir comparer divers corpus, l'indice est exprimé en valeur relative.

Soit T le nombre total de tranches ou extraits et t_i le nombre de tranches dans lequel apparaît le *ième* vocable. L'indice de répartition de ce *ième* vocable sera :

$$(1) \quad R_i = \frac{t_i}{T}$$

R varie uniformément entre 1 (vocable présent dans toutes les tranches) et 1/T (toutes les occurrences sont localisées dans la même tranche), à condition toutefois que ce nombre d'occurrences soit au moins égal à T (problème examiné plus bas, à propos des limites de l'indice).

En pondérant les effectifs des vocables par leur répartition, on obtient un indice d'usage (Juilland et Al. 1970).

Soit :

- N le nombre total de mots (en anglais : tokens) composant le corpus (ou sa longueur) ;
- n_i, le nombre d'occurrences du *ième* vocable dans l'ensemble du corpus,
- f_i, la fréquence relative de ce vocable.

Dans la suite de cette note, le terme "fréquence" désigne toujours la fréquence relative et "effectif" le nombre d'occurrences. Conventionnellement, les effectifs seront notés n et les fréquences f :

$$(2) \quad f_i = \frac{n_i}{N}$$

- U_i, son indice d'usage sera :

$$(3) \quad U_i = f_i * R_i$$

L'hypothèse implicite est la suivante : un vocable serait d'autant plus important qu'il aurait à la fois une fréquence élevée et une répartition régulière sur l'ensemble du corpus. A l'inverse, un vocable fréquent mais localisé en un point du corpus, et absent du reste, serait caractéristique à un milieu, une période, un thème et n'aurait pas le même poids dans la langue.

Limites des indices traditionnels de la répartition

Ces indices comportent des limitations dont certaines semblent avoir échappé aux auteurs qui les ont élaborés. Les deux principales concernent la liaison entre répartition et fréquence :

- un vocable dont le nombre d'occurrences est inférieur au nombre de tranches ($n_i < T$), aura nécessairement un indice inférieur à l'unité. Par exemple, le corpus du français québécois est découpé en dix tranches. Seuls les vocables apparaissant au moins dix fois dans ce corpus peuvent donc faire l'objet du calcul. De ce fait, une proportion importante du vocabulaire ne peut faire l'objet du calcul. Par exemple sur les 20 761 vocables différents de *A la recherche du temps perdu* (M. Proust) – utilisé dans la suite de cette note – seulement 5 144 apparaissent dix fois ou plus. Les trois quarts du vocabulaire sont donc exclus.

- plus le vocable a des effectifs élevés, plus il a de chance de figurer dans un grand nombre de tranches et donc d'avoir une répartition régulière. Par exemple, dans le corpus québécois, un vocable d'effectif 100 a théoriquement dix fois plus de chances d'atteindre la répartition maximale qu'un vocable utilisé 10 fois.

Loin d'apporter une information indépendante de la fréquence, ces indices sont donc largement redondants.

De plus, le décompte en tranches égales s'applique difficilement aux corpus exhaustifs dans lesquels tous les textes ont des longueurs différentes. Par exemple, dans le corpus des interventions radio-télévisées du président Mitterrand (1981-1988) qui sera présenté plus loin, le texte le plus long contient presque 40 fois plus de mots que le plus petit. Dans ce cas, il faut découper le corpus en tranches égales, mais certaines tranches comporteront la fin d'un texte et le début du suivant ("enjambement"). Ce découpage est acceptable lorsque le corpus est chronologique et que les textes concernés par l'enjambement ne sont pas séparés par des durées trop grandes ou trop inégales. Le corpus Mitterrand remplit ces conditions et pourra donc être découpé en tranches égales (résultats dans la section suivante).

Enfin, la formule 3 comporte un jugement de valeur : à fréquence égale, le vocable employé dans un grand nombre de tranches est plus important que celui qui n'apparaît que dans un petit nombre. Cette hiérarchie semble discutable. La répartition indique simplement que le premier vocable appartient au vocabulaire commun à l'ensemble du corpus et le second à un vocabulaire propre à l'une ou l'autre des parties de ce corpus.

De cet examen critique, on tire que *la mesure de la répartition d'un vocable dans un corpus doit être indépendante du nombre d'occurrences de ce vocable* afin de donner, à côté de la fréquence, une information sur la distribution de ses apparitions.

Dans cette perspective, C. Muller proposait de définir la *répartition* comme "la façon dont les occurrences d'un vocable sont réparties dans l'étendue du texte" ou du corpus (Muller 1977, p 55. Voir également Muller 1985a et 1985b).

A sa suite, nous définirons *la répartition d'un vocable dans un texte comme l'ensemble des emplacements où ce vocable apparaît.*

Ces emplacements sont identifiés en numérotant les mots du texte au fur à mesure de leur apparition (pour un exemple, voir ci-dessous le tableau 2 qui donne la répartition de trois mots dans l'œuvre de M. Proust). Quand l'apparition est unique, cette localisation est en elle-même significative et il n'est pas besoin d'information supplémentaire. En revanche, dès qu'un vocable revient en plusieurs endroits d'un texte, plusieurs questions sont soulevées. Ce retour est-il régulier et comment mesurer cette régularité ?

On remarquera que la répartition d'un vocable ne doit pas être confondue avec son éventuelle "spécificité". Cette seconde notion répond à la question : peut-on considérer que le vocable est caractéristique de tel ou tel passage ou tel ou tel locuteur ? (Lafon 1980 ; Labbé & Labbé 1994).

La deuxième partie de cette note examine la répartition à travers quelques exemples. La troisième partie présente un indice de répartition qui quantifie le degré de régularité d'un vocable à l'intérieur d'un corpus.

II. Répartition des vocables à la surface d'un corpus ordonné

La répartition d'un vocable est définie comme sa **localisation** en certains points de la surface du texte. Comme tout phénomène spatial, celui-ci est susceptible d'une représentation graphique et de diverses mesures.

Représentation graphique de la localisation d'un vocable

Soit un vocable *j* apparaissant n_j fois dans le corpus comptant au total *N* mots. On découpe le corpus en T tranches de longueur t_j, entier supérieur à N/n_j afin que, en cas de répartition homogène, le vocable soit présent au moins une fois dans chacune des tranches.

Par exemple, la répartition des pronoms *je* et *nous* dans les interventions radiodiffusées de François Mitterrand (président de la république française) entre mai 1981 et mai 1988[1], soit :

n = 101 textes,
N = 405 242 mots (longueur du corpus)
V = 8 583 : nombre de vocables différents (en anglais "types") ou vocabulaire du corpus.

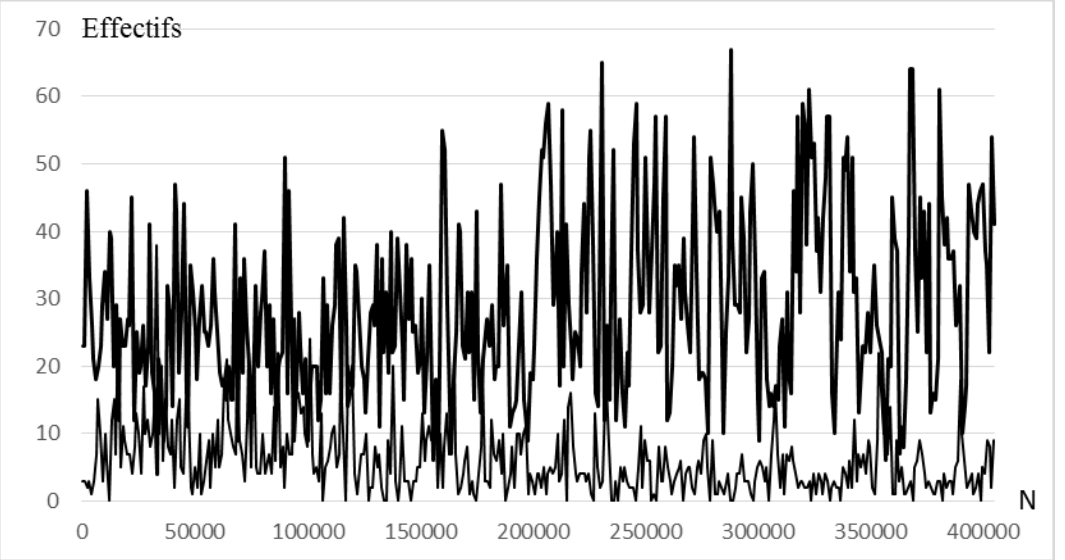
Le pronom personnel le plus employé est "je" : 11 592 occurrences, soit avec la formule (2), une fréquence de 28.6‰. Ce ratio signifie que, dans n'importe quelle tranche de mille mots extraits dans ce corpus, ce mot a une probabilité d'occurrence de 28,6. Le pronom personnel "nous" est utilisé 2 524 fois soit une fréquence de 6.2‰.

Pour visualiser la localisation de ces deux vocables à la surface du corpus, celui-ci est découpé en 405 tranches de 1 000 mots. Pour chaque tranche, on relève le nombre d'apparitions de "je" et de "nous" (figure 1). En ordonnées du graphique : les 405 observations (nombre d'occurrences dans chaque tranche de mille mots) et en abscisses le nombre de mots lus depuis le début du corpus au moment de l'observation.

Remarque : le président ayant observé une certaine régularité dans ses interventions radio-télévisées, l'axe horizontal est aussi celui du temps.

[1] Nos deux études de 1990 (Hubert & Labbé ; Labbé) portaient sur ce corpus sans la campagne pour l'élection présidentielle, soit 35 textes qui ont été ajoutés à la présente étude.

Figure 1. Localisation des occurrences des pronoms "je" (trait gras) et "nous" (trait maigre) dans les interventions radio-télévisées du président Mitterrand entre 1981 et 1988 (Effectifs absolus, corpus découpé en tranches de 1000 mots)



La densité d'utilisation du "je" est systématiquement supérieure à celle du "nous" (sauf à une dizaine de reprises, soit 2,5% des observations). De plus, la densité du "je" semble plutôt ascendante et celle du "nous", descendante A part cela, les deux séries paraissent si chaotiques qu'il semble difficile d'en tirer quelque chose, du moins par un simple examen visuel. Trois questions principales sont posées.

- Peut-on quantifier l'ampleur des variations ?
- Au-delà de ces variations, peut-on considérer que F. Mitterrand fait le même usage de ces deux pronoms tout au long de ces 7 ans ou bien le temps exerce-t-il une influence, comme le suggère le graphique ?
- S'il y a une dimension temporelle, est-il possible de localiser des ruptures délimitant des périodes ?

La densité d'emploi du "je" chez F. Mitterrand sera utilisée pour illustrer les réponses statistiques à ces trois questions,

Mesure de l'hétérogénéité de la répartition d'un vocable

Pour répondre à la première question, la voie classique consiste à calculer le nombre moyen d'apparition du vocable dans les tranches ($\bar{n}$) puis la dispersion standard des observations autour de cette moyenne.

Soit n_t nombre de fois que le vocable apparaît dans la tranche *t*, avec *t* variant de 1 à *T* (nombre de tranches).

$$(4)\quad \bar{n} = \frac{\sum_{1}^{T} n_t}{T} = \frac{11592}{405} = 28{,}6$$

Comme les tranches comptent ici mille mots, on retrouve la fréquence du "je" dans l'ensemble du corpus.

La dispersion des observations autour de cette moyenne est mesurée grâce à la "déviation standard" ou écart-type (σ), c'est-à-dire la moyenne quadratique des écarts des *T* observations (n_t) à la moyenne arithmétique $\bar{n}$ (formule 5). Cette valeur est rapportée à la moyenne pour obtenir le coefficient de variation relative : v% (formule 6).

$$(5)\ \sigma = \sqrt{\frac{\sum_{1}^{T} (n_t - \bar{n})^2}{T}} = 13{,}1$$

$$(6)\ v\% = \frac{\sigma}{\bar{n}} = 45{,}8\%$$

Le coefficient de variation signifie qu'environ les deux tiers des observations sont comprises dans un intervalle de ± 45,8 % autour de la moyenne, ce qui est important. Cela signifie que :

- la répartition du vocable à la surface du corpus est hétérogène,
- en conséquence, la fréquence moyenne représente mal le phénomène. Elle est assez peu prédictive du nombre de vocables que l'on rencontrera dans un texte quelconque appartenant à ce corpus (*a fortiori* dans la langue générale, ou un langage de spécialité, dans le cas d'un corpus construit).

Ce qui amène la question de la stabilité de l'usage au cours du temps de l'emploi de la première personne par le président. Ce corpus s'étend sur sept années. Dès lors, le temps est-il l'un des facteurs qui "perturbe" la densité d'emploi des pronoms de la première personne par le président ?

Mesure de l'influence du temps sur la localisation d'un vocable dans un corpus

Pour traiter cette question, la technique de **l'ajustement** linéaire d'une série chronologique sera utilisée. Les observations sont rangées par ordre chronologique (*t* variant de 1 à *T*), et représentées sur un graphique comme celui de la figure 2. Sur ce graphique, chaque observation est représentée par un point ayant comme abscisse *t* (numéro d'ordre de la tranche) et ordonnée n_t (nombre d'occurrences du vocable dans cette tranche).

Il existe une droite, dite "droite d'ajustement", passant au plus près de tous les points, qui rend compte de l'évolution chronologique de la variable. Cette droite passe par un point "moyen" dont les coordonnées sont $\bar{t}$ et $\bar{n}$. Pour une tranche *t*, l'ordonnée de la droite d'ajustement – notée n'_t - est donnée par la formule 7.

$$(7)\ n'_t = a\,.t + b$$

- $\bar{t}$ est le "milieu" de la période, c'est-à-dire le numéro d'ordre de la tranche médiane, si *T* (nombre total de tranches) est impair, ou un point théorique situé à mi-chemin entre les deux tranches médianes si *T* est pair ;

- $\bar{n}$ est le nombre moyen d'apparition dans chaque tranche du vocable sous revue (formule 4) ;

- le coefficient *a* est le coefficient directeur, ou "pente" de la droite d'ajustement (formule 8) ;

- *b* est l'origine de cette droite, c'est-à-dire le point où elle coupe l'axe des ordonnées ou encore la valeur théorique de la variable pour l'année zéro (formule 9).

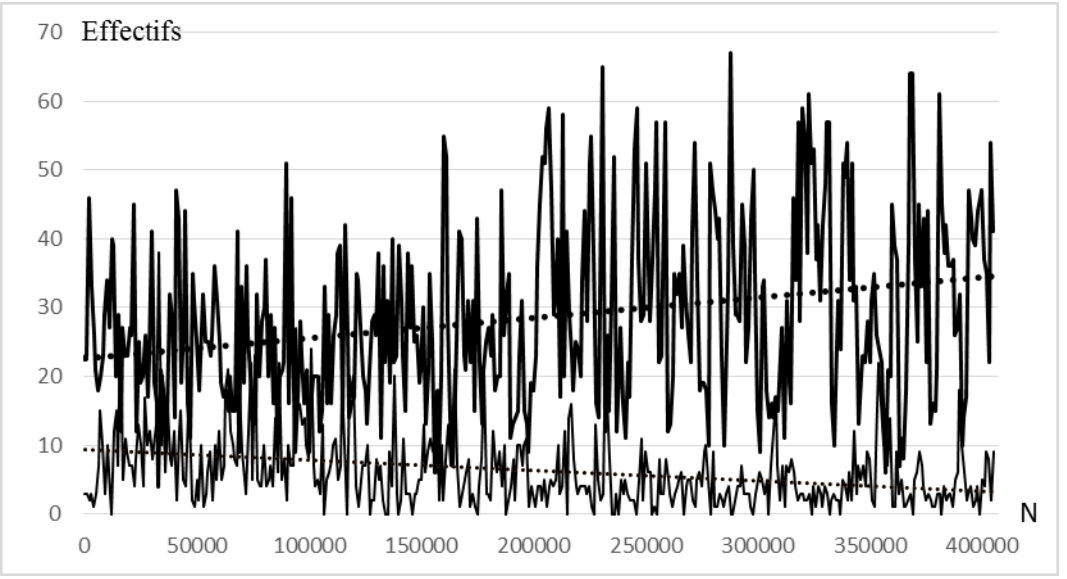
$$(8)\quad a = \frac{\sum_{1}^{T}(n_t - \bar{n})(t - \bar{t})}{\sum_{1}^{T}(t - \bar{t})^2} = 0{,}0294$$

$$(9)\quad b = \bar{n} - a.\,\bar{t} = 22{,}60$$

Une valeur positive du coefficient *a* indique que la variable s'accroît avec le temps. A l'inverse, une valeur négative révèle une tendance au recul. C'est le cas du pronom "nous" (*a* = - 0,0153).

Ces valeurs permettent de tracer les deux droites d'ajustement (Figure 2)

Figure 2. Droites d'ajustement des occurrences du "je" (trait gras) et du "nous" (trait maigre) dans le corpus des interventions radio-télévisées de F. Mitterrand (1981-1988)



On peut estimer cette tendance (δ) sur l'ensemble de la période en rapportant *b* - origine de la droite d'ajustement, ou valeur théorique pour le début du corpus – à *f'*$_{405}$ ou valeur théorique de la variable dans la dernière tranche. Voici le calcul pour le pronom de la première personne :

$$\delta = \frac{n'_{405} - n'_0}{n'_0} = +0{,}527$$

Le résultat signifie que, entre mai 1981 et mai 1988, la propension moyenne de F. Mitterrand à dire "je" a augmenté de 52,7%. Le même calcul appliqué à "nous" donne -66,4% : en moyenne, la densité de la première personne du pluriel a diminué des deux tiers. Il s'est donc produit un changement considérable dans le discours du président entre le début et la fin de son mandat. Celui-ci s'est centré sur sa propre personne et le collectif s'est effacé (nous dirons plus bas le contenu de ce collectif).

Cependant, cette conclusion ne peut être acceptée sans avoir auparavant examiné la **qualité** de l'ajustement de *n* par *t*. Cette qualité est donnée par le coefficient *r*, dit coefficient de détermination de la variable par le temps (formule 10) qui est une application particulière du coefficient de corrélation de Bravais-Pearson. Il consiste à rapporter la covariation de la variable et du temps (numérateur) à la variation totale (produit de leurs écart-types respectifs, au dénominateur). Les résultats varient entre ±1 (liaison linéaire positive ou négative) et 0 (absence de liaison).

$$(10)\quad r = \frac{\dfrac{\sum_{1}^{T}(n_t - \bar{n})(t - \bar{t})}{T}}{\sqrt{\dfrac{\sum_{1}^{T}(n_t - n)^2}{T}} * \sqrt{\dfrac{\sum_{1}^{T}(t - \bar{t})^2}{T}}}$$

Pour les pronoms 'je' et 'nous', les coefficients de détermination sont respectivement égaux à +0,263 et -0,331.

Avec 403 degrés de liberté (*T*-2), ces deux coefficients sont élevés. La table du coefficient de corrélation[1] indique que l'on peut accepter l'existence de la liaison linéaire avec moins de 1% de chances d'erreur. Enfin, le calcul n'est pas fait sur un échantillon mais sur la population entière, ce qui élimine cette source possible d'erreur et renforce la conclusion.

On en conclut donc que, au fur et à mesure de son premier mandat (1981-1988), F. Mitterrand a fait preuve d'une propension croissante à utiliser la première personne du singulier et d'une tendance inverse pour le *nous*.

Cependant, les coefficients indiquent l'existence d'un "résidu" (1 - r) que le temps n'explique pas. Dans le cas précis, ce résidu est important : 74% des fluctuations pour le *je* et 67% pour le *nous*. Autrement dit, l'action du temps est indéniable mais d'autres facteurs sont à l'œuvre (ceci est marqué par le profil chaotique des courbes dans les figures ci-dessus).

Ce constat conduit logiquement à la troisième question posée au début de ce paragraphe : peut-on isoler des "périodes", c'est-à-dire des sous-groupes plus homogènes au sein de cette collection de textes ?

---

1 Fisher Ronald A. & Yates Frank. *Statistical Tables for Biological, Agricultural and Medical Research*.1949. Ces tables figurant en annexe de la plupart les manuels de statistiques.

Localisation des ruptures et des périodes

Les méthodes et calculs sont présentés dans Hubert, Labbé & Labbé 2004.

Quatre remarques préalables :

- le calcul porte sur des tranches de longueurs égales mais la méthode peut être appliquée à des textes de longueurs inégales ;

- il s'agit d'un corpus exhaustif dont l'ordre ne peut être bouleversé. L'algorithme ne peut grouper que des tranches voisines.

- l'opérateur doit choisir une longueur minimale pour un segment donné. Dans le cas présent, un segment doit comporter au minimum 10 tranches voisines.

- il doit également choisir un seuil de significativité des résultats (souvent appelé "risque d'erreur"), ici $\alpha = 5\%$ et $\alpha = 1\%$. La répétition des opérations avec deux seuils différents permet de juger de la solidité de la partition obtenue.

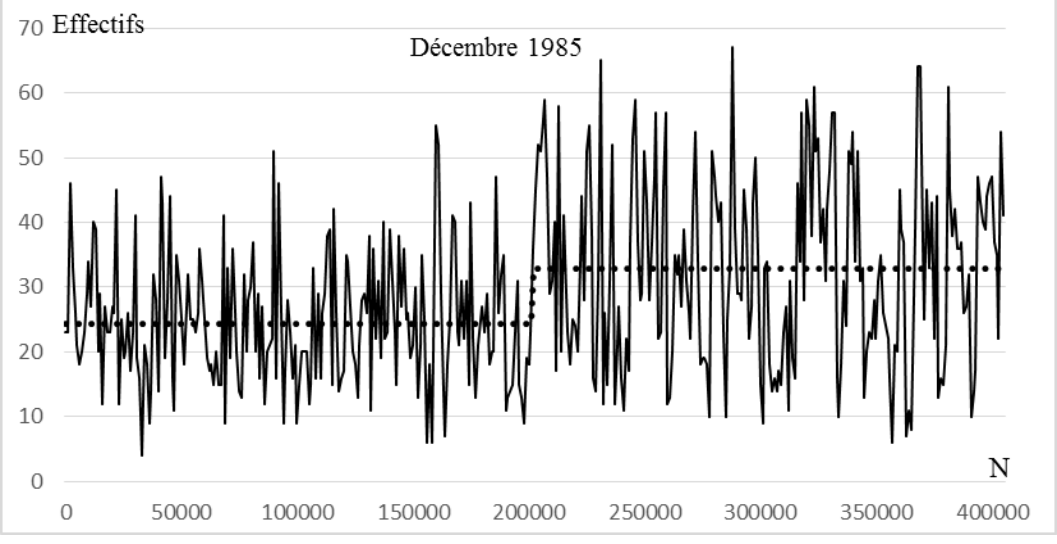
L'algorithme teste les principaux regroupements possibles en cherchant à constituer des segments aussi homogènes et contrastés que possible. La première condition revient à rechercher le groupement qui minimise l'écart-type entre les tranches regroupées au sein d'un même segment. La seconde signifie que les moyennes de tous les segments voisins, considérés deux à deux, doivent être significativement différentes au seuil choisi.

L'étude s'est déroulée en deux étapes.

Dans un premier temps, les conditions les plus fortes sont imposées à l'algorithme : le seuil le plus élevé pour la comparaison entre les moyennes de deux segments possibles ($\alpha = 1\%$) et au moins 30 tranches pour former un segment (soit un bloc de 30 000 mots).

Pour la première personne du singulier, cette première procédure aboutit à une segmentation du corpus en deux parties égales, la coupure se situant le 15 décembre 1985 (figure 3).

Figure 3. Segmentation des interventions radiotélévisées du président F. Mitterrand (1981-1988) en fonction de la densité des pronoms de la première personne ($\alpha = 1\%$; longueur minimale des segments : 30 000 mots).



Les graphiques et le calcul présenté dans le paragraphe précédent pouvait laisser penser à une croissance du "je" plus ou moins étalée sur l'ensemble des sept ans. En fait, le 15 décembre 1985, la propension du président à dire "je" a augmenté d'un coup de 32,8% - par rapport à la moyenne du début du septennat -, à l'occasion de l'émission de TF1 "Ca nous intéresse, Monsieur le président", animée par Y. Mourousi. C'est le tournant essentiel du premier septennat de F. Mitterrand. Il correspond au lancement de la campagne pour les élections législatives de mars 1986 (ce tournant avait déjà été signalé dans Labbé 1990a).
Alors que le calcul indique un phénomène stationnaire sur les années 1981-1985, il laisse planer un doute sur la période postérieure à décembre 1985.

Dans un deuxième temps, le seuil α est fixé à 5% et la longueur minimum des segments est réduite à 10. Neuf périodes apparaissent (figure 4 et tableau 1).

Figure 4. Segmentation des interventions radiotélévisées du président F. Mitterrand (1981-1988) en fonction de la densité des pronoms de la première personne (α = 5% ; longueur minimale des segments : 10 000 mots).

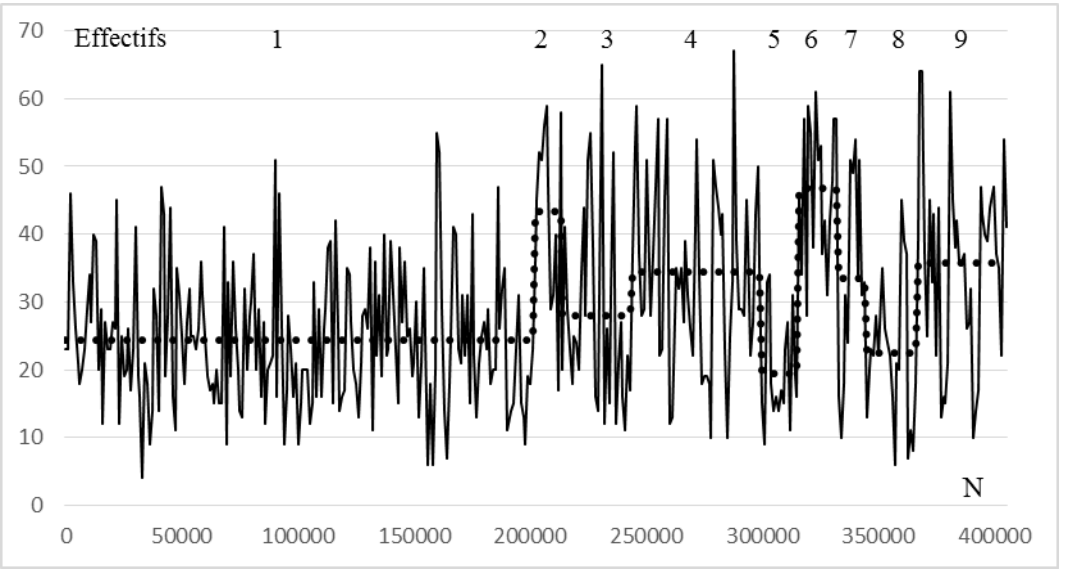


Tableau 1. Les neufs périodes des interventions radio-télévisées de F. Mitterrand (1981-1988). Découpage en fonction de la densité en pronoms de la première personne

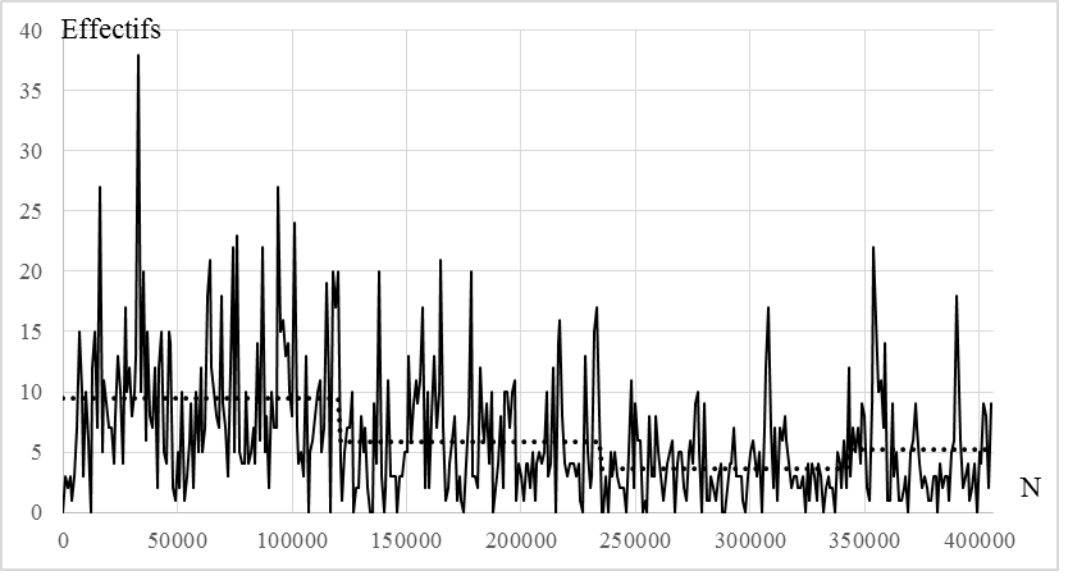
Segment	Dates	Longueur (mots)	densité moyenne du je (‰)
1	Mai 1981-décembre 1985	201 000	24,4
2	Décembre 1985 – février 1986	12 000	43,4
3	Mars-mai 1986	30 000	27,9
4	Mai 1986 – décembre 1987	56 000	34,4
5	Décembre1987 - janvier 1988	16 000	19,5
6	Février 1988	17 000	46,8
7	Mars 1988	12 000	33,5
8	Mars-avril 1988	22 000	22,5
9	Avril-mai 1988	33 000	35,8

La coupure principale ne change pas. La première période est donc homogène, c'est-à-dire que la propension moyenne à dire "je" du président Mitterrand peut être considérée comme stable entre mai 1981 et décembre 1985, soit la moitié du corpus.

La deuxième période est découpée en huit sous-périodes. En premier lieu, la campagne pour les élections législatives de mars 1986 qui voient la victoire de la droite. Pendant les deux années suivantes, le président devra "coexister" avec un gouvernement dirigé par J. Chirac (la troisième sous-période est celle de l'installation de ce gouvernement, la quatrième, celle de la "coexistence" proprement dite). Les quatre dernières sous-périodes correspondent à la campagne pour la présidence qui oppose J. Chirac à F. Mitterrand, campagne d'abord officieuse (sous-période 6) puis déclarée (sous-périodes 7 - 9).

Le pronom "nous" présente un profil un peu plus régulier (figure 5) avec une descente en trois paliers jusqu'en février 1988 et connaît alors un modeste ressaut durant la campagne pour l'élection présidentielle du printemps 1988.

Figure 5. Segmentation des interventions radiotélévisées du président F. Mitterrand (1981-1988) en fonction de la densité des pronoms de la première personne du pluriel (α = 1% ; longueur minimale des segments : 30 000 mots).



L'intérêt de cette procédure est évident. Contrairement aux découpages traditionnels (en années civiles ou à l'aide de dates jugées *a priori* décisives), cette segmentation échappe à l'arbitraire de l'opérateur et aux hypothèses réductrices.

Une fois ce découpage acquis, l'étude du vocabulaire caractéristique de chaque période et sous-périodes (Labbé & Monière 2012) et celle des "univers lexicaux" du "je" et du "nous" (Labbé & Labbé 2006) donnent la clef du phénomène. En résumé, F. Mitterrand sur-emploie la première personne du singulier lorsqu'il parle des institutions - spécialement des pouvoirs du président ou de ses relations avec le gouvernement – et de la politique étrangère, en particulier de l'armée et de

la force de frappe. Naturellement, cette caractéristique peut s'expliquer par les institutions de la Ve république qui font du président la clef de voûte des institutions et le chef des armées, unique responsable de la dissuasion nucléaire. Mais, comparé aux autres présidents, notamment le général de Gaulle, F. Mitterrand apparaît bien comme plus attaché aux pouvoirs du président, aux choses militaires et à une vision très XIXe siècle de la politique internationale (Labbé 1998). A l'inverse, lorsque les questions économiques et sociales sont abordées, le "je" semble s'absenter, ces questions étant prise en charge par le gouvernement, les ministres, par le pronom "nous", voire par un "il" impersonnel.

On en conclut que, entre 1981 et 1985, la densité de la première personne est stationnaire car, dans les interventions radio-télévisées, le poids des différents sujets est à peu près le même. A partir de décembre 1985, la politique l'emporte puis, à partir d'avril 1986, face à un gouvernement hostile, le président se replie sur ses pouvoirs constitutionnels et son rôle coutumier en politique étrangère où sa propension à dire "je" se manifeste fortement.

Ce premier examen suggère donc qu'il existe des liens entre les répartitions de plusieurs vocables.

Liaison entre les principaux vocables d'un corpus

L'hypothèse est la suivante : s'il existe des rapports d'exclusion et d'association entre certains vocables, ceux-ci devraient se traduire dans leur répartition à la surface du corpus. Lorsqu'ils sont associés dans l'esprit du locuteur, deux (ou plusieurs) vocables devraient être localisés dans les mêmes passages. A l'inverse, en cas d'opposition, la présence de l'un exclurait plus ou moins la présence de l'autre à proximité.

Le corpus utilisé est *La recherche du temps perdu* de M. Proust. Cette oeuvre se compose de 6 romans parus sur une durée de 15 ans (1913-1927) : *Combray*, *Du côté de chez Swann*, *A l'ombre des jeunes filles en fleur*, *Le côté de Guermantes*, *Sodome et Gomorrhe*, *La prisonnière*, *La fugitive*. Outre le narrateur et sa mère, les quatre personnages principaux sont Albertine Simonet (2 280 occurrences, soit une fréquence de 1,94 ‰), Charles Swann (1 554, 1,32 ‰), les Guermantes (1 449, 1,23 ‰ ; 780 occurrences désignent explicitement la duchesse Oriane) et le baron de Charlus (1 190, 0,94‰).

L'œuvre est découpée en 117 tranches égales de 10 000 mots (la dernière est négligée). La densité d'emploi des quatre principaux noms dans les tranches est décrite dans les figures 6 et 7).

La lecture des graphiques est la suivante. Albertine apparaît à la 25^{e} tranche (*A l'ombre des jeunes filles en fleur*) et sa densité maximale (133 pour 10 000 mots) est atteint dans la 112^{e} tranche qui correspond au passage de *La Fugitive* où le narrateur apprend la mort de la jeune fille et repense à leurs relations, etc.

Figure 6. Localisation de "Albertine" et de "Guermantes" dans la *Recherche du temps perdu* (Effectifs = nombre d'occurrences par tranches de 10 000 mots)

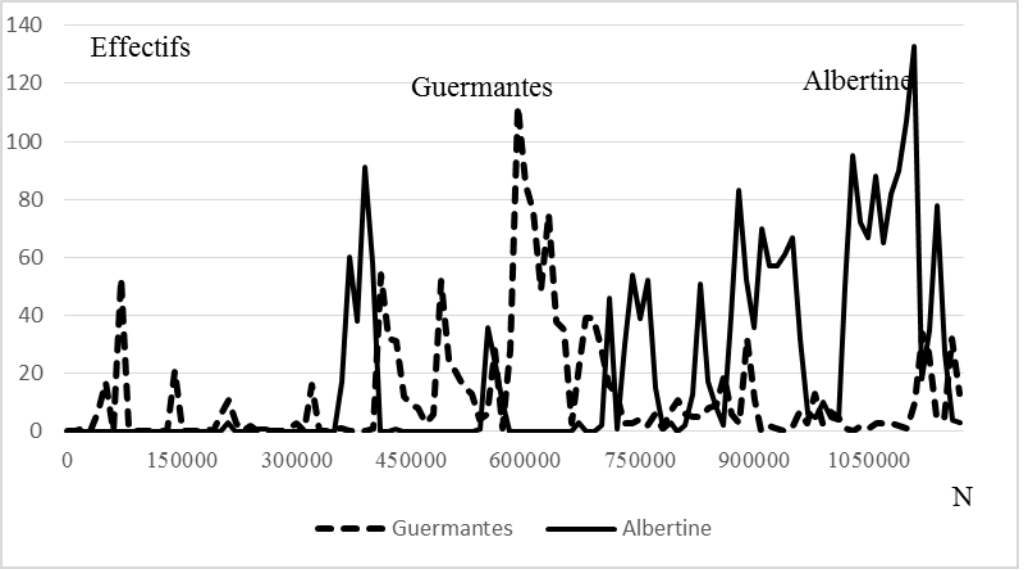
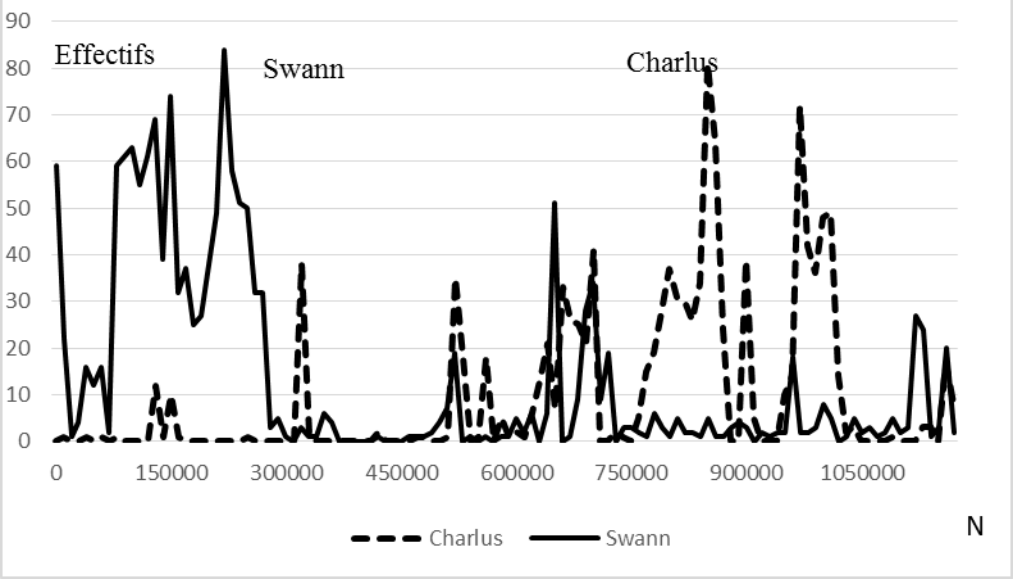


Figure 7. Localisation de "Swann" et de "Charlus" dans la *Recherche du temps perdu*. (Effectifs = nombre d'occurrences par tranches de 10 000 mots)



Ces deux graphiques suggèrent une sorte d'exclusion mutuelle entre ces quatre personnages pris deux à deux : les passages consacrés aux Guermantes, spécialement à la duchesse, ne parleraient pas d'Albertine (d'origine bourgeoise et maîtresse de l'auteur) et cette exclusion n'est pas chronologique. Certes, Albertine n'apparaît pas dans le premier quart de l'œuvre, mais à partir de ce moment, les deux femmes figurent dans des tableaux distincts et assez alternatifs. Pour Swann, il est surtout présent au début et réapparaît ensuite assez rarement mais là encore, il y a une antinomie apparente avec la présence de Charlus.

Le coefficient de corrélation permet de tester cette hypothèse (le principe du calcul a déjà été présenté à propos de la qualité de l'ajustement chronologique d'une variable). Soit $\overline{n_A}$ et $\overline{n_B}$, respectivement les effectifs moyens des vocables

Albertine et Guermantes dans l'ensemble des tranches ; n_{Ai} et n_{Gt} le nombre d'occurrences respectivement d'Albertine et de Guermantes dans la *tième* tranche. La formule 10 devient :

$$r = \frac{\frac{\sum_{1}^{T}(n_{At} - \overline{n_A}) * (n_{Gt} - \overline{n_G})}{T}}{\sqrt{\frac{\sum_{1}^{T}(n_{At} - \overline{n_A})^2}{T}} * \sqrt{\frac{\sum_{1}^{T}(n_{Gt} - \overline{n_G})^2}{T}}} = -0{,}246$$

Avec 115 degrés de liberté, la table du coefficient de corrélation indique que l'on a moins d'une chance sur 100 de se tromper en affirmant que les Guermantes – spécialement la duchesse Oriane - n'apparaissent pas dans les passages concernant Albertine et vice-versa ou encore que ces personnages appartiennent à des univers – la bourgeoisie et l'aristocratie - qui sont assez étrangers dans l'œuvre de Proust mais aussi à des périodes différentes du récit.

Avec Charles Swann et le baron de Charlus, le coefficient est égal à -0.198 soit une significativité au seuil de 5%.

A ce propos, l'empan étant assez large (10 000 mots), certaines tranches peuvent chevaucher plusieurs passages, ce qui peut expliquer pourquoi les coefficients ne sont pas plus élevés.

Certes, cette double opposition n'est pas une découverte. Tout lecteur attentif aura noté qu'il y a dans le monde social de Proust deux "côtés" qui n'ont guère de communication entre eux. Il y a aussi chez M. Proust, deux types principaux de femmes, d'hommes et de sentiments amoureux. Chacun de ces "côtés" s'incarnent dans quelques personnages principaux qui n'ont pas vocation à se rencontrer, du moins avant l'affaire Dreyfus puis le *Temps retrouvé*. Il s'agissait de montrer comment il est possible, à l'aide d'instruments statistiques simples, d'accéder rapidement à la structure essentielle d'une œuvre d'une certaine ampleur.

Peut-on résumer ce phénomène à l'aide d'une dimension unique ?

III. L'indice de répartition

Cette dimension unique doit rendre compte le mieux possible du phénomène, être indépendante de la fréquence et présenter les caractéristiques attendues pour un indice statistique. En particulier, les valeurs doivent évoluer entre un minimum et un maximum connus - ici 0 et 1 - de manière uniforme sans saut ni seuil en couvrant tous les cas possibles. Il est également souhaitable que cet indice soit simple à mettre en œuvre et à interpréter.

Après avoir exposé le calcul de l'indice, sa portée et ses principales propriétés, plusieurs applications seront présentées, ce qui conduira à définir une procédure précise pour exploiter cette dimension dans les grands corpus étiquetés.

Calcul

Une première présentation de cet indice a été faite dans Hubert & Labbé 1990a ; Hubert & Labbé 1990b.

Dans un corpus de N mots, considérons un vocable j ayant n_j occurrences et associons à ce vocable une dimension t_j égale à l'inverse de la fréquence relative ($1/f_j$), soit N / n_j. La répartition de j évoluera entre deux situations extrêmes :

- l'uniformité : j apparaît à espace régulier. En cas de régularité parfaite, l'espace séparant deux occurrences du mot est t_j. Dans ce cas, l'indice doit atteindre sa valeur maximum, c'est-à-dire 1 ;

- la localisation : toutes les occurrences de j surviennent en un seul point du texte. En cas de localisation parfaite, toutes les occurrences sont contigües et l'espace séparant la dernière occurrence de la première est égal à $N - n_j$. Dans ce cas, l'indice doit atteindre son minimum, soit 0. On remarque que pour atteindre ce résultat, il faut soustraire l'effectif de j à N et "boucler" le texte entre la dernière occurrence et la première.

Entre ces deux situations extrêmes, l'indice doit varier uniformément en rendant compte exactement de la répartition du mot. Par exemple, à mi-chemin entre les deux pôles décrits ci-dessus, un fragment de longueur t_j sur deux contient deux occurrence de j. Dans ce cas, l'indice doit être égal à 0,5 (du moins si ces occurrences "contigües" sont exactement séparées par un intervalle égal à $t_j/2$. Etc.

La procédure de calcul de l'indice découle de ces caractéristiques (le tableau 2 présente trois exemples de répartitions et la procédure de calcul).

Premièrement, calcul des n_j intervalles d séparant chaque occurrence du vocable. Soit d_i le nombre de mots séparant les *ième* et *(i+1)ème* occurrences de ce vocable j. Les bornes des intervalles étant comprises dans le calcul, la somme de ceux-ci sera égale à N.

$$\sum_{1}^{nj} (d_i) = N$$

Deuxièmement, les intervalles d_i sont classés par longueurs croissantes (i variant de 1 à n_i).

Troisièmement, de combien la répartition observée s'écarte de la répartition uniforme ? L'indice i est incrémenté de 1 à (k-1) tant que d_i est inférieur ou égal à t_j. Soit k la valeur de i lorsque d_i devient supérieur à tj [1]. L'intervalle d_k contient un certain nombre de fragments dans lequel le vocable considéré ne figure pas, contrairement à une répartition uniforme. Leur nombre est égal à (d_k - t_j). Il y a (n_j - k) intervalles où une telle situation est possible.

Appelons M_j le nombre de mots contenus dans ces fragments d'où le vocable j est absent, alors qu'il devrait être présent en cas de répartition uniforme. M_j est égal à :

$$M_j = \sum_{k}^{nj} (d_i - t_j) = \sum_{k}^{k-1} d_i - \left[(n_j - k)* t_j\right]$$

Quatrièmement, calcul de N'_j : nombre de mots contenus dans des intervalles de longueur inférieure à t_j contenant le vocable j :

$$N'_j = N - M_j$$

L'indice de répartition, pour le vocable j, repose sur la comparaison de N et de N'_j (formule 11)[2].

$$(11)\quad R_j = \frac{N' - n_j}{N - n_j}$$

R_j variera entre :

- 1 lorsque $k = nj$ alors $N'_j = N$ (rappel : les bornes des intervalles sont comprises dans le calcul de d). Le numérateur de la formule 11 est strictement égal à son dénominateur. La répartition est *uniforme*.

- 0 si $N' = n_j$: toutes les occurrences sont contiguës et sont contenues dans un intervalle de nj mots. M est égal au dernier intervalle (N - n_j). Le numérateur de la formule 11 est nul. *La répartition est strictement localisée*.

Entre ces deux bornes, R varie uniformément sans saut ni seuil. Il atteint 0,5 lorsque la répartition est exactement à mi-chemin entre la localisation et l'uniformité. Ses résultats correspondent sans biais au phénomène à décrire et sont faciles à interpréter.

Lorsque le corpus a une dimension suffisamment grande, on peut calculer la répartition moyenne des vocables et associer à cette moyenne, un écart type empirique (formule 5 ci-dessus). Ces valeurs servent à juger de l'usage d'un vocable singulier.

---

[1] Généralement t_j n'est pas entier. Il convient donc de prendre comme longueur de d_k le premier intervalle au moins égal à l'entier immédiatement supérieur à t_j.

[2] La notation retenue (R pour répartition) est en majuscule pour éviter une confusion avec le coefficient de corrélation de Bravais-Pearson traditionnellement noté r.

Par exemple, dans l'ensemble de la *Recherche du temps perdu*, l'indice de répartition moyen est de 0,563 avec un écart-type de 0,106. Pour les quatre principaux personnages, ces indices sont : Albertine : 0,261 ; Charlus : 0,227 ; Guermantes : 0,306 ; Swann : 0,275. Sous réserve de ce qui est dit plus bas à propos de la répartition des mots à majuscule, ces quatre personnages sont significativement plus localisés que la moyenne : ils n'apparaissent qu'en certains passages du livre, comme l'avait montré leurs localisations.

Trois exemples, également tirés de Proust vont illustrer la procédure de calcul avant une application à un nouveau corpus.

Procédure de calcul

Ci-dessous, le tableau de calcul utilisé par l'ordinateur. Il s'agit de trois vocables extraits du vocabulaire de *La recherche du temps perdu* : le vocable le plus localisé (le nom féminin *pâtissière*), le plus régulier (l'adverbe *mentalement*) et un des vocables qui se situent approximativement à mi-chemin entre la régularité et la localisation parfaites (le nom féminin *claustration*). Ces trois exemples ont également été choisis car leurs effectifs sont faibles, ce qui permet de donner chacun de leurs emplacements dans la *Recherche*, contrairement aux quatre principaux personnages analysés précédemment.

Tableau 2. Calcul de l'indice de répartition (*R*) pour trois vocables caractéristiques de *A la recherche du temps perdu* (M. Proust, 1 175 840 mots).

Vocables	*mentalement* (adv.)		*claustration* (n.f.)		*pâtissière* (n.f.)	
n_j	11		12		13	
Rang	Occurrences	Intervalles	Occurrences	Intervalles	Occurrences	Intervalles
1	61 185		338408		1 056 561	
2	177 186	116 001	344710	6 302	1 056 576	15
3	287 346	110 160	467735	123 025	1 056 621	45
4	382 468	95 122	573147	105 412	1 056 667	46
5	586 904	204 436	962096	388 949	1 056 712	45
6	675 806	88 902	965663	3 567	1 056 724	12
7	686 300	10 494	1038348	72 685	1 056 747	23
8	797 888	111 588	1039480	1 132	1 056 777	30
9	919 421	121 533	1041242	1 762	1 056 950	173
10	996 979	77 558	1050947	9 705	1 057 042	92
11	1 158 612	161 633	1080874	29 927	1 057 134	92
12			1157221	76 347	1 057 163	29
13					1 057 227	64
bouclage		78 413		35 7027		1 175 174
t_j		106 896		97 987		90 449
n_j-k		6		4		0
M		183 983		582 466		1 175 174
N'		991 856		593 374		666
R_j		0,8435		0,5046		0,0006

Pour chaque vocable, en première colonne, l'emplacement des occurrences : par exemple, la première occurrence de "mentalement" survient au 61 185e mot du texte et la seconde au 177 186e, soit un intervalle de 116 001 mots supérieur à t_j (106 896). Il y a ainsi six intervalles supérieurs à cette valeur repère. Le total de ces six intervalles est égal à 641 367 mots auxquels il faut déduire six fois t_j pour obtenir *M* (183 983) puis *N'* (991 856) et, grâce à la formule 11 : *R* (0,8435). C'est l'indice le plus élevé observé dans *La recherche du temps perdu*. Etc.

On peut vérifier que "claustration" apparaît pratiquement dans un intervalle sur deux de dimension t_j.

Il faut ensuite éditer les concordances de ces mots pour comprendre leur singularité (leurs effectifs sont trop faibles pour calculer leurs univers lexicaux). Par exemple, "pâtissière" apparaît seulement dans une anecdote de *La fugitive* où l'auteur est soupçonné de pédophilie par une commerçante (la patissière) qui prévient la police. Quant à "mentalement", sa régularité ne surprend pas : les mécanismes de la mémoire et des sentiments sont le principal thème de la *Recherche du temps perdu*.

Signification de l'indice de répartition

Statistiquement, l'indice de répartition est une bonne approximation de la *probabilité pour qu'un segment, de t_j mots contigus, prélevé aléatoirement dans le corpus, contienne le vocable j*. Pour le vérifier, en 1990, nous avons effectué un test sur les propos de F. Mitterrand lors de son face-à-face d'avril 1988 avec J. Chirac. Ce test a porté sur les vocables d'effectifs supérieurs à 9. Pour chaque vocable *j*, il a été prélevé aléatoirement 10 000 tranches de longueur t_j parmi lesquels ont été décomptés le nombre de tranches où apparaissait le vocable considéré. Puis ces résultats ont été rapportés à l'indice R_j. Les résultats sont donnés dans le tableau en annexe, pour les vocables les plus fréquents. Le nombre relatif de tranches de taille t_j prélevées aléatoirement et où la recherche du vocable a été positive (R_j') se voit associer un intervalle de confiance de deux écarts types.

Les valeurs de l'indice R_j tombent pratiquement toutes dans cet intervalle de confiance. Il est possible d'en conclure que le modèle donne la description assez exacte du phénomène estimé grâce aux tirages aléatoires, c'est-à-dire la probabilité de rencontre du vocable *j* dans une tranche quelconque de t_j mots prélevée dans le corpus.

L'indice s'approche d'autant plus de cette probabilité que n_j est très petit par rapport à *N*. Dans ce cas, on peut écrire :

$$R = \frac{N'_j - n_j}{N - n_j} \approx \frac{N'_j}{N} \text{ avec } N >> n_j$$

En effet, le rapport N'_j/N est le nombre de cas favorables (rencontre du vocable *j* dans un segment quelconque de longueur t_j) sur le nombre total de cas possibles

(les *N* combinaisons possibles, soit autant que de mots contenus dans le texte). Ce rapport tend vers une limite inférieure égale à n_j/N (dans le cas d'occurrences contiguës du vocable).

Théoriquement, cela conduit *R* à être indépendant de la fréquence des vocables. Cette propriété va être discutée à l'aide d'une question décisive pour l'étude d'un corpus :

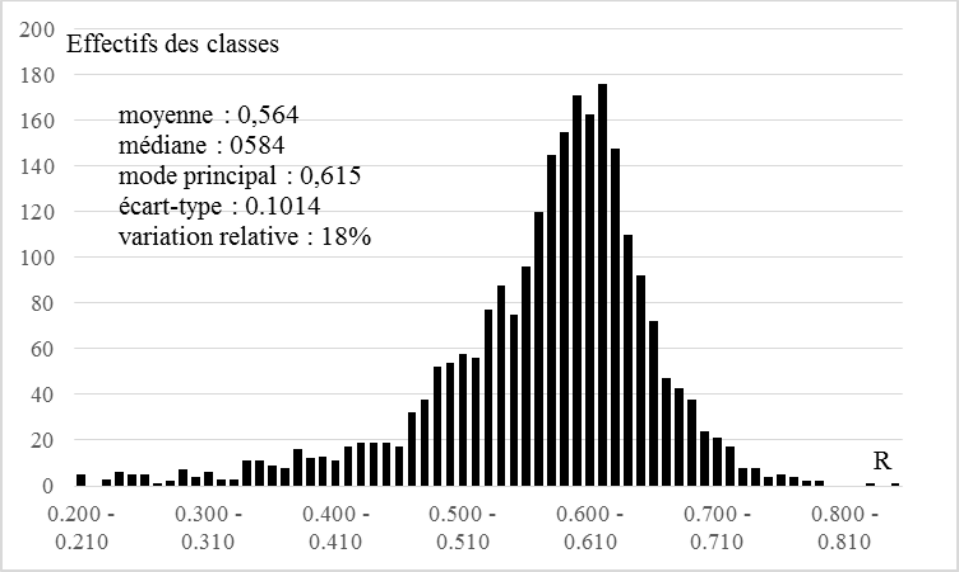
Comment déterminer les vocables singuliers (les plus régulièrement utilisés et les plus localisés) ?

Pour répondre à cette question, on utilise le corpus des discours et messages du général de Gaulle entre son retour au pouvoir (juin 1958) et sa démission (mars 1969), soit 451 textes comptant au total 407 332 occurrences et un vocabulaire de 9 094 vocables dont 2 466 ayant plus de 9 occurrences.

Les vocables anormalement répartis

Les 2 466 indices *R* sont rangés par ordre croissant dans des classes d'intervalles égaux (0,01) dont les effectifs sont représentés dans le graphique de la figure 8.

Figure 8. Histogramme des indices de répartition des vocables ayant au moins dix occurrences dans les discours et messages du général de Gaulle (1958-1969) classement par ordre croissant (intervalle des classes 0,01)



La figure présente grossièrement un profil "en cloche" – dit "gaussien" – mais avec plusieurs caractéristiques notables. D'une part, il n'y a pas un mode unique (valeur la plus fréquente, notée *Mo*). Ici le mode principal (milieu de la classe dont les effectifs sont les plus importants : 0 ,61-0,62) est accompagné de plusieurs modes secondaires (notamment 0,595 et 0,535). D'autre part, dans une série "gaussienne", on s'attend à ce que les trois valeurs centrales : la médiane (valeur de l'indice pour l'individu du milieu, notée *Me*), la moyenne ($\overline{R}$) et le mode

soient confondues. Ici elles sont assez décalés avec une relation ($\bar{R} < Me < Mo$) qui indique une asymétrie à gauche, bien visible sur la figure 8 avec une importante "queue de distribution" dans les plus basses valeurs.

Dans une population, où un caractère est distribué de manière gaussienne, la procédure classique de recherche des individus "singuliers", quant à ce caractère, est la suivante : choix d'un seuil ou "risque d'erreur" (α = 5% ou 1 %) et repérage des vocables significativement réguliers ou irréguliers par rapport à ces valeurs limites.

Par exemple, quand un caractère est distribué de manière gaussienne dans une population donnée, il y a moins de 1% des individus qui se trouvent en-deçà ou au-delà des valeurs limites ($\bar{R} \pm 2{,}56\ \sigma$ pour α = 1 % et $\bar{R} \pm 1{,}96\ \sigma$ pour α = 5%). L'indice de répartition ne semble pas répondre à ces critères (tableau 3)

Tableau 3. Effectifs théoriques et constatés pour le nombre de vocables dont la répartition serait "anormale" dans le corpus des discours et messages du général de Gaulle (1958-1969).

	Valeurs seuils de R	Effectifs attendus	Effectifs constatés
Bornes inférieures (1%)	0,304	13	73
(5%)	0,365	62	118
Bornes supérieures (5%)	0,762	62	11
(1%)	0,823	13	2

Comme il y a 2 466 vocables dotés d'un indice *R*, une distribution gaussienne laisse attendre 13 vocables au-delà de chacune des deux bornes de l'intervalle à 1% et 62 en dehors de l'intervalle à 5%. La dernière colonne du tableau donne les effectifs constatés et montre que le raisonnement ne correspond pas à la réalité. Il y a trop de vocables en dessous des seuils inférieurs et pas assez au-dessus des seuils supérieurs. C'est une conséquence logique du profil asymétrique de la figure 8. De plus, l'examen de ces vocables singuliers (tableaux 4 et 5) suggère quelques constats.

Les vocables "anormalement" répartis présentent trois caractéristiques principales.

- A part un adjectif (*canadien*), tous les vocables les plus irrégulièrement répartis sont des noms propres alors qu'il n'y en a aucun dans les plus réguliers. La présence des noms de pays ne surprend pas. Par exemple, sauf lors des visites de leurs chefs d'Etat à Paris, le Général ne parle des pays d'Amérique Latine que lors de son célèbre voyage du 21 septembre au 16 octobre 1964. De même, *Bizerte* n'est évoquée que lors de la crise de juillet 1961. Cette localisation souligne en même temps la place des mots à majuscule initiale, à l'interface entre la langue et les lieux, les personnes, les institutions sur lesquels porte le discours (on parle parfois d'"embrayeurs").

Tableau 4. Les vocables les plus irrégulièrement répartis avec leur indice de répartition et leurs effectifs dans les discours et messages du général de Gaulle (1958-1969).

Vocable	Répartition	Occurrences
Paraguay	0.002	12
Finlande	0.002	10
Danemark	0.002	13
Bizerte	0.004	17
Venezuela	0.004	12
Colombie	0.005	11
Polynésie	0.007	20
Equateur	0.008	14
Bolivie	0.014	10
canadien (adj)	0.043	16
Brésil	0.054	19
Israël	0.065	11
Chili	0.080	35
Argentine	0.107	20
Pérou	0.113	20
Québec	0.114	49
Turquie	0.120	46
Mali	0.130	16
Iran	0.136	23
Mexique	0.136	36

Tableau 5. Les vocables les plus régulièrement répartis dans le corpus de Gaulle avec leurs indices de répartition et leurs effectifs dans les discours et messages du général de Gaulle (1958-1969).

Vocables	Répartition	Occurrences
hostile	0.761	13
encouragement (nom)	0.761	15
inébranlable (adj)	0.761	10
divisé (adj)	0.766	11
inconsistance (nom)	0.767	12
arrivée (nom)	0.778	11
aventure (nom)	0.779	15
ébranler (verbe)	0.782	10
respecté (adjectif)	0.783	10
asiatique (adjectif)	0.823	10
rattacher (verbe)	0.848	10

- Aucun "mot outil" (adverbe, pronom, déterminant, conjonction) ne figurent dans ces deux listes.

- Aucun vocable fréquent n’est présent dans ces deux listes. Le plus employé est *Québec* : 49 occurrences, quasiment toutes lors de son voyage de juillet 1967 et de son célèbre "vive le Québec libre" (hôtel de ville de Montréal, le 24 juillet 1967).

Au passage, ceci souligne la nécessité du retour au texte pour interpréter ces listes. Par exemple il est frappant de trouver dans la liste des vocables les plus

régulièrement employés, l'opposition entre les qualités de la Ve république et l'état de la France à son "*arrivée*" (*divisé, inconsistance, aventure, ébranler*)…

Ajoutons que ces écarts entre le comportement de *R* et le modèle gaussien se retrouvent dans tous les corpus avec les trois mêmes caractéristiques : prédominance des noms propres dans les vocables les plus irréguliers, absence des mots outils et des vocables les plus fréquents dans les deux populations "anormales". Ces constats soulèvent au moins deux questions. D'où vient cette dissonance entre le modèle de la distribution aléatoire et la distribution effective de R ? L'indépendance de la répartition et de la fréquence est-elle remise en cause ?

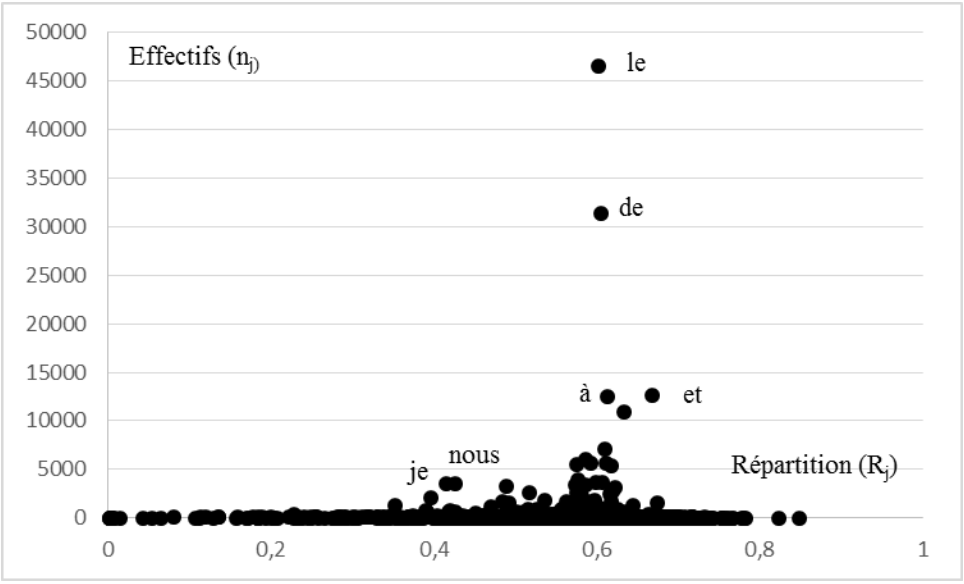
La réponse à cette seconde question permettra de découvrir une autre propriété de *R*.

Relation entre la répartition des vocables et leurs effectifs

L'indépendance de *R* et de n_j a été vérifié systématiquement sur tous les corpus, en calculant les coefficients de corrélation entre ces couples de valeurs pour l'ensemble des vocables appartenant aux corpus. Tous ces coefficients indiquent l'indépendance des deux variables.

Cependant, il existe bien une relation entre la répartition et les effectifs comme les figures ci-dessous permettent de le comprendre. Sur ces deux graphiques, chaque vocable *j* (avec $n_j > 9$) est représenté par un point d'abscisses R_j et d'ordonnée n_j. Par exemple, pour le déterminant "le", les coordonnées sont $n = 46\,600$ et $R = 0.603$.

Figure 9. Relation entre les effectifs et la répartition des vocables (discours et messages du Général de Gaulle, vocables de 10 occurrences et plus).

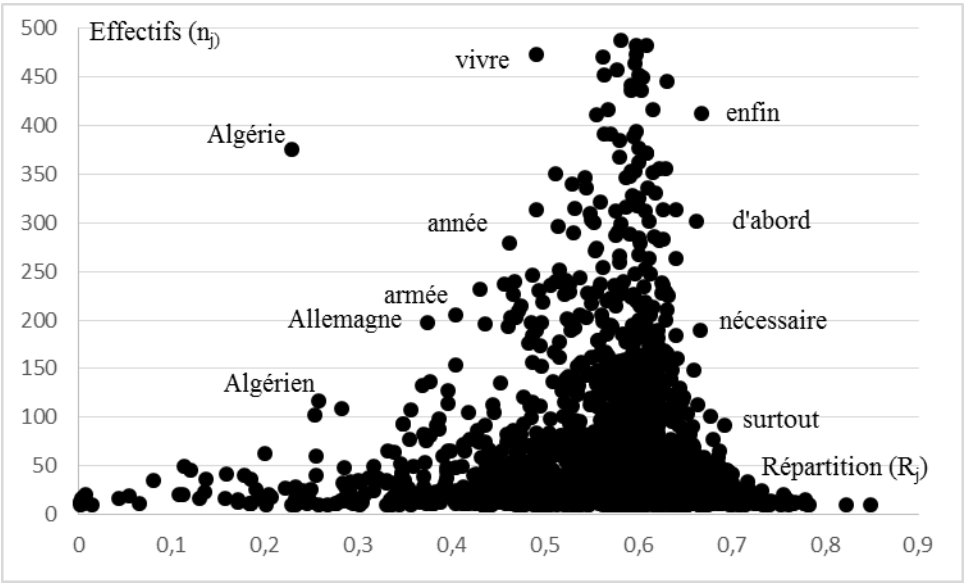


Du fait de la présence de quelques vocables très utilisés, le phénomène n'est pas lisible pour les effectifs moyens et faibles. On remarque cependant que, pour

les trois vocables les plus utilisés (le, de, à) l'indice de répartition est sensiblement égal à la valeur modale (comprise dans la classe 0,61-0,62).

Pour rendre visibles les basses occurrences, la figure suivante effectue un "zoom" sur celles-ci (10-500).

Figure 10. Relation entre le nombre d'occurrences et la répartition des vocables (discours du Général de Gaulle, vocables de 10 à 500 occurrences).



Les caractéristiques de la distribution des indices R en fonction du nombre d'occurrences des vocables sont les suivantes : un étalement considérable de *R* dans le bas du tableau, une forte asymétrie faisant apparaître une "queue" importante à gauche et un resserrement progressif vers le mode au fur et à mesure que le nombre d'occurrences augmente.

La relation entre répartition et fréquence se formule ainsi : *plus le nombre d'occurrences d'un vocable est élevé dans un corpus, plus sa répartition tendrait vers la répartition modale du vocabulaire dans ce corpus.* Ou encore, la dispersion des indices de répartition est d'autant plus faible que le nombre d'utilisation des vocables est élevé.

Les points remarquables dans les deux figures permettent également de supposer que le phénomène dépend des catégories grammaticales.

Certains noms propres auraient un usage "localisé" comme nous l'avons vu chez Proust à propos des quatre principaux personnages de *la Recherche*. L'*Algérie* et les *Algériens* en donnent une bonne illustration : très présents entre 1958 et 1962, ils disparaissent totalement du discours du Général après l'été 1962 (indépendance de l'Algérie). De même pour l'*Allemagne*, dont la majorité des occurrences surviennent entre fin 1958 - première rencontre entre de Gaulle et le chancelier Adenauer - et la démission de ce dernier (octobre 1963).

Tout ceci suggère une liaison entre les catégories grammaticales (ou "parties du discours") et la répartition des vocables.

Relation entre la répartition des vocables et leurs catégories grammaticales

Le tableau 6 ci-dessous récapitule les principales caractéristiques de la répartition des vocables en fonction de leur catégorie grammaticale. A l'indice de répartition moyen de chaque catégorie est associé un écart type empirique et un coefficient de variation relative (formules 5 et 6 ci-dessus) qui indique l'importance de la dispersion des indices autour de la moyenne propre à chaque catégorie grammaticale.

Tableau 6. Principales caractéristiques de la répartition des vocables dans les discours et message du général de Gaulle en fonction de leurs catégories grammaticales.

Catégorie	Effectifs ($n_j > 9$)	Nombre moyen d'occurrences	Répartition moyenne	Coefficient de variation de R (V%)
Mots à majuscule initiale	130	59,8	0,338	49
Verbes	536	99,0	0,603	9
Substantifs	1026	65,2	0,558	16
Adjectifs	437	50,7	0,586	14
Pronoms	48	983,9	0,559	14
Pronoms personnels	*15*	*469,8*	*0,538*	*15*
Adverbes	181	139,5	0,594	9
Déterminants	49	1530,6	0,530	15
Numéraux	*29*	*201,2*	*0,517*	*16*
Indéfinis	*11*	*416,8*	*0,576*	*7*
Prépositions	35	1874,3	0,592	6
Conjonctions	18	1473,5	0,571	13
Coordination	*9*	*1814,9*	*0,571*	*17*
Subordinations	*9*	*1132,1*	*0,572*	*6*

Il y a 130 mots à majuscule utilisés dix fois ou plus (deuxième colonne). Ils sont en moyenne utilisés 59,8 fois (deuxième colonne). Leur répartition moyenne est de 0,338, bien loin de la moyenne générale de 0,564. De plus, cette répartition est très étalée (coefficient de variation relative : 49%). Les noms propres tranchent nettement par rapport à toutes les autres catégories. Comme indiqué ci-dessous, ces noms assurent une sorte d'interface entre le discours et la réalité extérieure où il est censé s'appliquer… Exemple : l'*Algérie* et les *Algériens* ou l'*Allemagne* et Adenauer…

Après les noms propres, les nombres présentent la répartition moyenne la plus basse et un coefficient de variation important car eux aussi rattachent le discours avec le monde extérieur grâce aux dates et à la quantification.

A l'opposé, les verbes, les adverbes et les prépositions apparaissent comme nettement plus réguliers car ils sont au cœur de la langue.

Au sein même de chaque catégorie, la variation autour de la moyenne peut être significativement différente. Par exemple, les conjonctions de subordination ont la même répartition moyenne que les coordinations, mais la dispersion des premières est bien moindre que celle des secondes (6% de variation standard autour de la moyenne contre 17%). Ces écarts sont considérables : si l'on prend comme base

les prépositions et les coordinations de subordination, l'étendue de la distribution des indices des substantifs est 2,5 fois plus grande, comme celle des nombres ; celle des conjonctions de coordination est 2,7 fois plus grande et celle des noms propres :7,6.

Enfin, il est intéressant de signaler que tous ces ordres de grandeur se retrouvent dans les autres corpus de la grande bibliothèque électronique. Il s'agit donc de caractéristiques du français et non pas de singularités du général de Gaulle.

Le problème statistique posé est le même que celui rencontré pour le calcul du vocabulaire caractéristique (Labbé & Monière 2012 ; Labbé & Labbé 1994). Plus un vocable est utilisé, dans un corpus donné, plus les indices (de spécificité, comme de répartition) convergent vers la valeur modale pour leur catégorie grammaticale. Mais cette convergence est moins rapide que ce que laisse attendre la distribution au hasard.

La solution proposée par Labbé & Monière en 2012, pour le calcul du vocabulaire caractéristique, s'applique également à la répartition. Cette solution tient compte à la fois de la catégorie grammaticale et de la fréquence.

Comportement des répartitions en fonction de la catégorie grammaticale

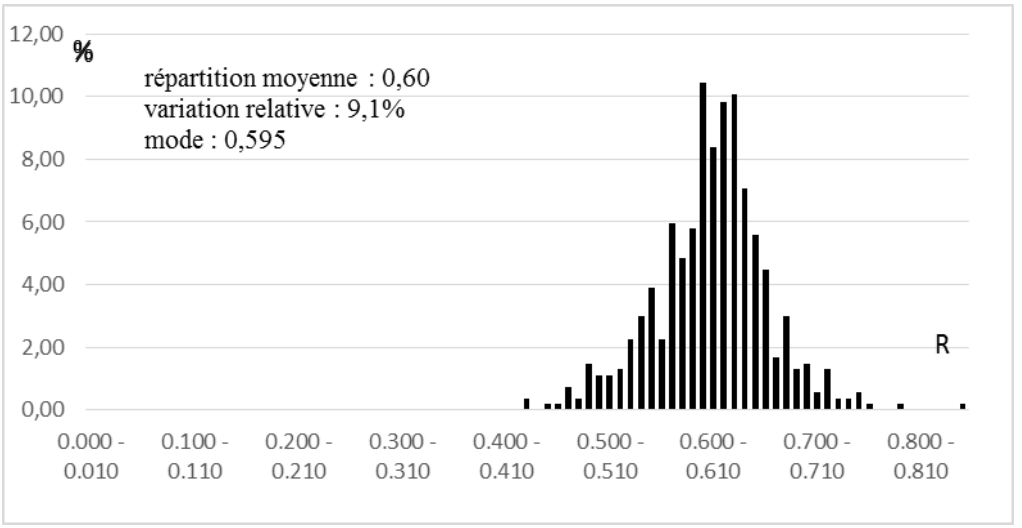
Le corpus est découpé en autant de sous-corpus que de catégories grammaticales significatives. Dans la présente expérience, le recensement se limite aux mots à majuscule initiale, verbes, substantifs, adjectifs et mots-outils. Cette dernière catégorie groupe les pronoms, les adverbes, déterminants, prépositions et conjonctions.

Les figures 11 présentent la distribution des indices R, par catégories grammaticales, rangés par valeurs croissantes dans des classes d'intervalles égaux. Par exemple dans la figure 11 - 3, 6,2% des vocables à majuscules initiales ont des indices de répartition compris entre 0,00 et 0,01, c'est-à-dire qu'on peut les considérer comme extrêmement localisés. En revanche, aucun verbe n'a d'indice R inférieur à 0,420.

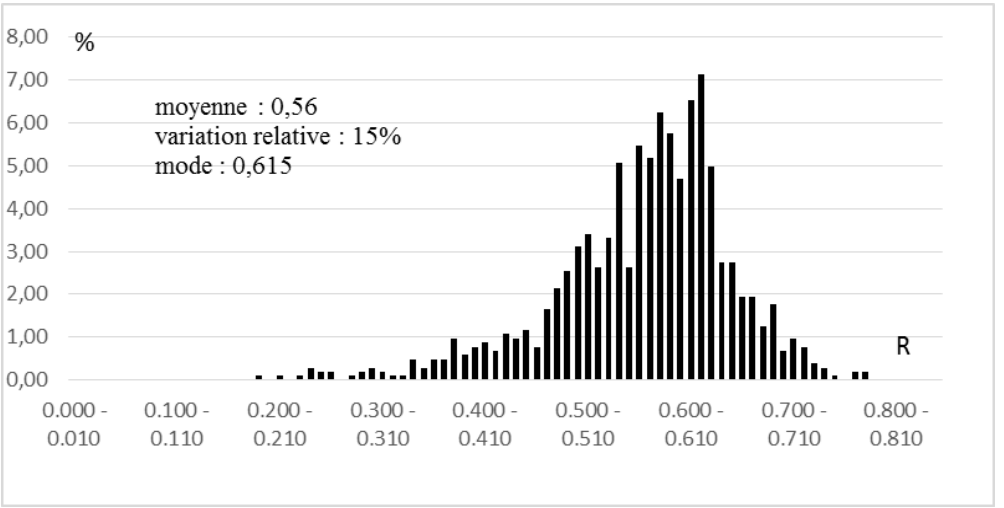
L'échelle horizontale identique, ainsi que les effectifs relatifs de chaque classe en ordonnées, permettent de visualiser les différences dans les répartitions en superposant les trois graphiques.

Figures 11. Répartition des vocables les plus fréquents du général de Gaulle classés par catégories grammaticales, par ordre croissant et par classes d'intervalles égaux.

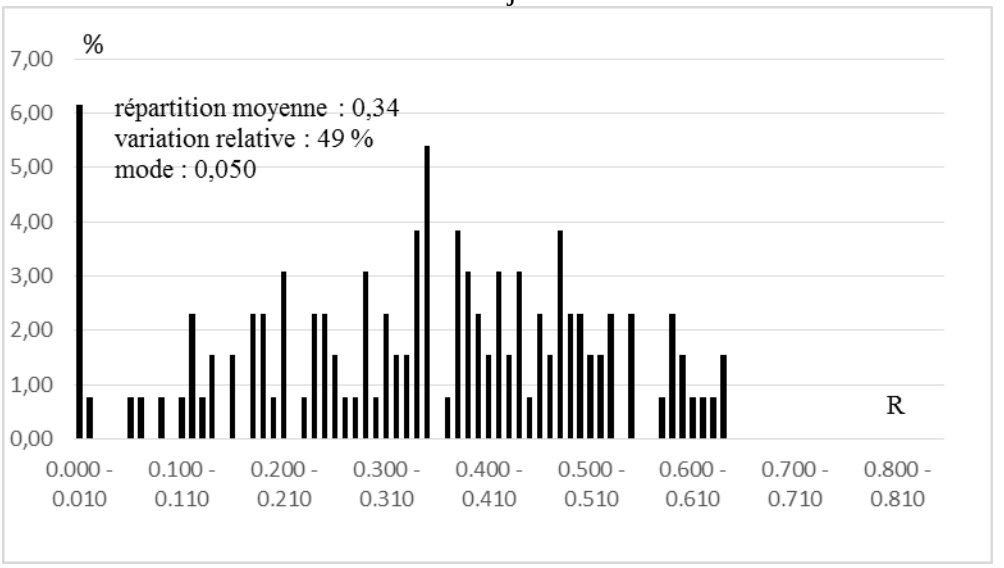
1. Les verbes



2. les substantifs



3. les mots à majuscule initiale



La répartition des verbes (Figure 11- 1) est en moyenne plus régulière que celle des substantifs (Figure 11 - 2) et elle est nettement plus resserrée autour de cette moyenne.

Le profil des adjectifs – non reproduit - est assez proche de celui des substantifs ; celui des mots-outils – également non reproduit - assez semblable à celui des verbes. Enfin, les "mots à majuscule" sont complètement à part (Figure 11 - 3), vérifiant ainsi que leur nature n'est pas comparable au reste du lexique.

Sur ces graphiques, chacun des modes secondaires correspond à une sous-catégorie, essentiellement en fonction des effectifs des vocables qui composent cette sous-catégorie. Par exemple, sur le graphique 11 - 1, le mode central contient la plupart des verbes les plus usuels (*être, avoir, faire, pouvoir, falloir, dire, devoir…*).

En résumé : les indices de répartition sont corrélés aux catégories grammaticales et, pour chacune de ces catégories, ils sont d'autant moins dispersés que les effectifs du vocable sont élevés.

On tire de ces deux constats, une procédure de repérage des vocables anormalement répartis.

Procédure de repérage des vocables anormalement répartis

Au sein de chacune des catégories grammaticales, les vocables sont rangés par effectifs croissants dans des classes dont l'intervalle varie à peu près de manière logarithmique. Ceci correspond à une particularité dite "loi de Zipf" - ou "Zipf-Mandelbrot" - selon laquelle la fréquence d'un mot dans un texte est lié à son rang (Zipf 1935 ; Mandelbrot 1957). Cette division assure à chacune des classes un effectif suffisamment important. Au sein de chacune de ces classes, une répartition moyenne et un écart-type sont calculées, ce qui permet de repérer les vocables anormalement répartis (aux deux extrémités de la distribution). Naturellement, le calcul de l'écart-type ne peut être fait que lorsque les classes de fréquence sont suffisamment peuplées (nous avons retenu ici un effectif minimal de 20 en dessous duquel la recherche est interrompue). Cela revient à dire que cette démarche ne peut s'appliquer qu'à des corpus de plusieurs centaines de milliers de mots.

Le tableau 7 ci-dessous présente le résultat de ces calculs sur les substantifs du général de Gaulle

Tableau 7. Les substantifs anormalement répartis dans les discours et messages du général de Gaulle (vocables classés par effectifs croissants)

Classe	Effectif de la classe	Répartition moyenne	Ecart type
10-50	359	0.564	0.102
- Vocables les plus localisés : patriarche (0.185), délégué (0.237), insurgé (0.238), continental (0.246), cardinal (0.249), livre (0.258)			
- Vocables les plus uniformes : déclin (0.739), trésor (0.740), encouragement (0.760), inconsistance (0.767), arrivée (0.778), aventure (0.779).			

Classe	Effectif de la classe	Répartition moyenne	Ecart type
50-100	501	0.552	0.083
- Vocables les plus localisés : sénat (0.200), front (0.222), rébellion (0.233), sire (0.254), outremer (0.284), monseigneur (0.291), musulman (0.292), dollar (0.307), altesse (0.311), autodétermination (0.320), bloc (0.330), désarmement (0.339)…			
- Vocables les plus uniformes : intention (0.688), rendement (0.692), dieu (0.694), instinct (0.696), habitude (0.701), fermeté (0.703), champion (0.707), risque (0.717), relief (0.717), discipline (0.717), essentiel (0.718), signification (0.732).			

Classe	Effectif de la classe	Répartition moyenne	Ecart type
100-200	100	0.561	0.063
- Vocables les plus localisés : majesté (0.282), parti (0.356)			
- Vocables les plus uniformes : particulier (0.630), contraire (0.635)			

Classe	Effectif de la classe	Répartition moyenne	Ecart type
200-500	56	0.570	0.048
- Vocables les plus localisés : armée (0.404), communauté (0.463)			
- Vocables les plus uniformes : condition (0.641), égard (0.645)			

Par exemple le dernier cadre indique qu'il y a 56 substantifs dont le nombre d'occurrences est compris entre 200 et 500, leur répartition moyenne est 0,570 avec un écart type de 0,048. Dans cette classe, le substantif le plus localisé est *armée* avec un indice de 0,404. L'essentiel de ses occurrences surviennent entre 1958 et 1962 (guerre d'Algérie). Pour les mêmes raisons, *communauté* est surtout employée pour désigner la Communauté française rapidement dissoute après les indépendances africaines. Pour désigner la communauté européenne, le général préfère dire "marché commun" ou Europe.

On remarque que les indices moyens – compris entre 0.55 et 0.57 - ne varient pas en fonction des effectifs. Il en est ainsi quelle que soit la catégorie grammaticale. Cela vérifie l'indépendance de la répartition par rapport à la fréquence. En revanche, plus les effectifs sont élevés, plus l'écart-type se resserre.

Il en est également ainsi pour toutes les catégories grammaticales (autres que les noms propres).

Ensuite, il faut se reporter aux contextes d'utilisation de chacun de ces vocables anormalement répartis, pour comprendre la raison de leur position singulière. On voit ainsi apparaître des familles de mots. Les plus localisés sont liés à des réceptions de dignitaires à l'Elysée (*patriarche, cardinal, sire, monseigneur, altesse, majesté*) ; ou encore à la période de la guerre d'Algérie (1958-1962) et à la décolonisation : *insurgé, front, rébellion, musulman, autodétermination, armée, communauté*…).

De même, on trouvera dans les noms propres des pays visités une seule fois ou l'*Algérie* et les *Algériens* (dont on ne parle plus après 1962). Curieusement, on trouve aussi *Allemagne* dans les mots à majuscules significativement localisés, alors que *Adenauer* est le nom propre le plus également réparti, car le chancelier allemand a été reçu régulièrement jusqu'à sa mort (avril 1967). A chaque rencontre avec Adenauer, le Général faisait un discours avec toast. C'est un cas unique qui souligne l'estime particulière que de Gaulle portait à Adenauer. En effet, les substantifs les plus régulièrement répartis révèlent certaines préoccupations constantes du Général comme le *déclin*, l'*inconsistance*, l'*aventure*, le *risque* opposés à la *fermeté* et à la *discipline*. Il y a aussi certaines expressions qui sont presque des "tics" de langage comme "au contraire", "en particulier", "à la condition", "à cet égard", etc.

Une remarque : nous avons choisi des seuils sévères afin de réduire les listes. En abaissant ces seuils, on obtient des listes plus longues sans pouvoir exclure certaines présences moins significatives.

Conclusions

La répartition d'un vocable dans un corpus est la dimension oubliée de la plupart des études statistiques "textuelles" au profit de la seule fréquence. Cela ne saurait surprendre puisque la plupart de ces études portent sur des corpus trop restreints et sur les seuls mots "bruts" sans standardisation des graphies ni étiquetage. Or, dans les vastes collections de textes, les fluctuations de graphie des mots sont rarement aléatoires. De plus, l'étiquetage des mots ("lemmatisation") permet d'accéder aux catégories grammaticales qui sont un facteur explicatif essentiel de la répartition. Le calcul n'a pas beaucoup de sens si on ne les prend pas en compte (il en va de même avec le calcul dit des "spécificité du vocabulaire").

Nous espérons avoir montré que la répartition est une dimension intéressante, du moins quand il s'agit de véritables corpus exhaustifs.

A titre exploratoire, l'étude de la localisation des vocables les plus fréquents permet de visualiser les répartitions pour les caractériser, formuler des hypothèses et choisir les calculs les plus aptes à vérifier ces hypothèses (ou à les invalider).

Dans un second temps, l'indice de répartition attache à chaque vocable une dimension complémentaire de la fréquence. Les indices moyens, et la dispersion des observations autour de celles-ci, permettent de caractériser le style d'un auteur ou d'un groupe et d'isoler les vocables significativement localisés en un point du corpus ou, à l'inverse, les plus uniformément répartis sur l'ensemble.

Enfin, l'indice sert à resserrer la présentation de l'index du vocabulaire (une seule ligne par vocable) sans perdre totalement l'information concernant la localisation des occurrences (voir l'index du vocabulaire de F. Mitterrand dans Labbé 1990b).

Il reste maintenant à adapter la méthode au cas des corpus d'échantillons dans lesquels il n'y a pas d'ordre naturel organisant les textes. Rappelons que l'indice de répartition donne la probabilité pour qu'un segment de t_j mots contigus, prélevé aléatoirement dans le corpus, contienne le vocable j. Dès lors, l'une des pistes les plus prometteuses consiste à extraire aléatoirement, dans ce type de corpus non-ordonnés des échantillons de longueur égale à la dimension caractéristique (t_j) et à compter le nombre de ces échantillons qui contiennent le vocable considéré, en évaluant la stabilité de la mesure grâce aux ré-échantillonnages (*bootstraps*). Cependant, la difficulté croît rapidement avec la dimension des corpus et la méthode semble difficilement applicable à plusieurs millions de mots et plusieurs dizaines de milliers de vocables.

Bibliographie

Beauchemin Normand, Martel Pierre & Théoret Michel (1992). *Dictionnaire de fréquence du français parlé au Québec : fréquence, dispersion, usage, écart réduit.* New York : Peter Lang.

Bernet Charles (1983). *Le vocabulaire des tragédies de Racine (Analyse statistique)*. Genève-Paris : Slatkine-Champion.

Burnard Lou (1995 réédition : 2007). *Reference Guide for the British National Corpus*. Oxford University Computing Services.

Eckart Thomas, Elmiger Daniel, Kamber Alain & Quasthoff Uwe (2013). *Frequency Dictionary French / Dictionnaire de fréquence du français*. Leipzig Leipziger Univeristätsverlag.

Engwall Gunnel (1984). *Vocabulaire du roman français*. Stockholm : Sture Allen.

Gougenheim Georges, (1958). *Dictionnaire fondamental de la langue française*. Paris : Didier.

Gougenheim Georges, Michéa René, Rivenc Paul, & Sauvageot Aurélien (1956). *L'élaboration du français élémentaire*. Paris : Didier. Rééditions augmentées en 1959 et 1967, sous le titre *L'élaboration du français fondamental*. Paris ; Didier.

Habert Benoît, Nazarenko Adeline & Salem André (1997). *Les linguistiques de corpus*. Paris : A. Colin.

Hubert Pierre & Labbé Dominique (1990a). Note sur l'indice de répartition utilisé dans l'index du vocabulaire de F. Mitterrand. Annexe à Labbé Dominique (1990), p.

Hubert Pierre & Labbé Dominique (1990b). La répartition des mots dans le vocabulaire présidentiel. *Mots*, 22, mars 1990, p. 80-88.

Hubert Pierre, Labbé Cyril & Labbé Dominique (2004). Automatic Segmentation of Texts and Corpora. *Journal of Quantitative Linguistics*, december 2004, 11-3, p. 193-213.

Imbs Paul (dir.) (1971). *Dictionnaire des fréquences. Vocabulaire littéraire des XIXe et XXe siècles*. Nancy : Centre de recherche pour un trésor de la langue française.

Juilland Alphonse, Brodin Dorothy, Davidovitch Catherine (1970). *Frequency Dictionnary of French Words*. La Haye : Mouton.

Labbé Cyril & Labbé Dominique (1994). *Que mesure la spécificité du vocabulaire ?* Grenoble: CERAT, décembre 1994 & juin 1997. Disponible en ligne sur le site de la revue *Lexicometrica*. 3-2001.

Labbé Dominique (1990a). *Normes de saisie et de dépouillement des textes politiques*. *Cahier du CERAT n° 7*. Grenoble : CERAT-IEP, avril 1990.

Labbé Dominique (1990b). *Le vocabulaire de François Mitterrand*. Paris : Presses de la Fondation Nationale des Sciences Politiques.

Labbé Dominique (1998). La France chez de Gaulle et Mitterrand. In Fiala Pierre et Lafon Pierre (dir). *Des mots en liberté. Mélanges Maurice Tournier*. Fontenay-aux-Roses : ENS Editions, p. 183-193.

Lafon Pierre (1980). Sur la variabilité de la fréquence des formes dans un corpus. *Mots*, 1, octobre 1980, p. 127-165.

Lafon Pierre (1983). *Dépouillements et statistiques en lexicométrie*. Paris-Genève : Slatkine-Champion.

Mandelbrot Benoît (1957). Étude de la loi d'Estoup et de Zipf : fréquences des mots dans le discours. In Apostel L. Mandelbrot B. & Morf. *Logique, langage et théorie de l'information*. Paris, PUF, p. 22-53.

Monière Denis & Labbé Dominique (2012). Le vocabulaire caractéristique du Premier ministre du Québec J. Charest comparé à ses prédécesseurs. In

Dister Anne, Longrée Dominique, Purnelle Gérald (éds). *Proceedings of the 11th International Conference on Textual Data Statistical Analysis*. Liège : LASLA - SESLA, 2012, p.737-751.

Muller Charles (1967). *Etude de statistique lexicale. Le vocabulaire du théâtre de Pierre Corneille*. Paris : Larousse. (réédition : Genève-Paris, Slatkine-Champion, 1979).

Muller Charles (1977). *Principes et méthodes de statistique lexicale*. Paris : Hachette, 1977, p. 55.

Muller Charles (1985a). Sur les répartitions lexicales. *Langue française, linguistique quantitative, informatique*. Genève Paris : Slatkine-Champion, p. 87-101

Muller Charles (1985b). La répartition lexicale : problèmes et solutions". Genève Paris : Slatkine-Champion, p. 103-113.

New Boris, Pallier Christophe (2005*). La documentation officielle de Lexique 3*. Chambery : Université de Savoie.

Sinclair John (1991). *Corpus, Concordance, Collocations*. Oxford : Oxford University Press.

Zipf George K. (1935). *La psychobiologie du langage*. Paris : CEPL, 1974.

Annexe - Répartition des vocables les plus fréquents dans les interventions de F. Mitterrand face à J. Chirac (avril 1988). Calcul sur des tranches de texte prélevées aléatoirement (R') puis simulation à l'aide du modèle de partition (R)

	Effectifs	Echantillons aléatoires		Indice de répartition
		R'	2 écarts types	R
le (article)	852	0,673	0,016	0,670
de	561	0,680	0,018	0,684
être (verbe)	349	0,654	0,016	0,650
avoir (verbe)	338	0,653	0,014	0,643
je	299	0,553	0,018	0,556
à	210	0,684	0,017	0,682
et	180	0,665	0,016	0,661
il	173	0,595	0,013	0,593
que (conj.)	159	0,583	0,012	0,595
vous	157	0,498	0,019	0,497
ne	144	0,607	0,012	0,601
un (article)	144	0,612	0,022	0,607
ce (pronom)	143	0,609	0,017	0,611
pas (adverbe)	126	0,595	0,013	0,607
qui	123	0,615	0,017	0,609
ce (article)	89	0,659	0,016	0,657
en (préposition)	86	0,618	0,014	0,616
dans	78	0,615	0,010	0,618
le (pronom)	76	0,662	0,017	0,655
dire	73	0,671	0,017	0,687
cent	72	0,427	0,015	0,429
pour	70	0,579	0,015	0,584
que (pronom)	66	0,644	0,016	0,647
se	62	0,600	0,017	0,602
nous	57	0,590	0,015	0,576
faire	56	0,576	0,015	0,574
y	55	0,568	0,013	0,569
mille	53	0,411	0,014	0,413
falloir	52	0,476	0,015	0,469
quatre	52	0,503	0,009	0,513
on	51	0,541	0,021	0,538
vouloir	50	0,651	0,013	0,650
vingt	49	0,524	0,015	0,524
neuf (numéral)	46	0,495	0,017	0,497
plus	46	0,629	0,016	0,628
mais	45	0,615	0,016	0,613
par	45	0,569	0,018	0,573
bien (adverbe)	44	0,624	0,019	0,620
monsieur	43	0,573	0,016	0,561
premier	41	0,525	0,013	0,526
avec	39	0,611	0,020	0,607
sur	38	0,537	0,018	0,542
ministre	36	0,495	0,015	0,497
pouvoir (verbe)	36	0,641	0,016	0,638
tout (déterminant)	35	0,647	0,015	0,639
là	34	0,673	0,013	0,678
ils	32	0,444	0,016	0,450
très	32	0,626	0,013	0,634
cela	31	0,597	0,011	0,590
si (conj.)	31	0,548	0,010	0,549
en (pronom)	30	0,615	0,014	0,614