La structure du vocabulaire du général de Gaulle

Pierre Hubert, Dominique Labbé

LA STRUCTURE DU VOCABULAIRE DU GENERAL DE GAULLE

Dominique Labbé
CERAT-Institut d'études politiques de Grenoble
Grenoble
Dominique.labbe@iep-grenoble.fr

Pierre Hubert
Ecole des Mines de Paris 35 rue Saint Honoré
77.305 FONTAINEBLEAU
hubert@cig.ensmp.fr

Summary

Description of the vocabulary structure of a corpus. The links between words are calculated using the hypergeometric law to compare the number of their co-occurences with their whole frequencies in the corpus. This calculus has been applied to the televised speeches and press conferences by General de Gaulle between 1958 and 1969.

KEY WORDS : Lexical Statistics — Vocabulary Structure — Co-occurences — de Gaulle.

Description de la structure du vocabulaire d'un corpus. Les liens entre les mots (force et direction) sont calculés grâce à la loi hypergéométrique qui permet de comparer le nombre de leurs co-occurrences avec leur fréquence totale dans le corpus. Ce calcul a été appliqué au discours radiotélévisés et aux conférences de presse du général de Gaulle entre 1958 et 1969.

Mots clefs : statistique lexicale – structure du vocabulaire – cooccurence – discours politique - Général de Gaulle

Publié dans BOLASCO Sergio, LEBART Ludovic et SALEM André. *III Giornate internazionali di Analisi Statistica dei Dati Testuali*. Rome : Centro d'Informazione e stampa Universitaria, 1995, tome II, p. 165-176.

La notion de "champ lexical" est classique en lexicologie : elle décrit la manière dont est organisé le lexique de la langue. Elle est plus rarement appliquée à l'analyse du vocabulaire d'un auteur ou alors il s'agit de reconstitutions plus ou moins "artisanales" à l'aide des tables de concordances.

Que peut tirer la statistique lexicale de l'hypothèse selon laquelle le vocabulaire d'un auteur est organisé à la manière d'un champ lexical ? En premier lieu cela suggère que les mots entretiennent entre eux des relations (d'association, de substitution, d'opposition voire d'exclusion) plus ou moins fortes et que l'ensemble des places et des relations constitue la *structure du vocabulaire* de l'auteur. Il faut donc recenser les relations que chaque mot entretient avec tous les autres. Mais cela risque d'aboutir à des tableaux démesurés. Pour éviter cet inconvénient, on suppose que le vocabulaire, comme le lexique de la langue, est organisé en quelques sous-ensembles groupés autour d'un ou plusieurs noyaux formés par les mots — ou paradigmes — qui suscitent le plus grand nombre de liaisons. Il suffit d'étudier ces mots pour repérer l'essentiel de la structure du vocabulaire.

Il s'agit de transposer, dans la statistique textuelle, des notions classiques en lexicologie (champ, synonymie, antonymie, paradigme, etc). Toutefois, afin de ne pas masquer cette transposition, il est sans doute préférable de ne pas utiliser la terminologie lexicologique et, notamment le mot "champ" qui s'applique à la langue, alors que nous travaillons sur des corpus, des discours, de la parole. C'est pourquoi nous avons proposé d'y substituer le terme d'«univers» (Labbé, 1990a).

En définitive, il s'agit de répondre à des questions simples : dans un corpus donné, quels sont les mots qui sont employés ensemble — on supposera qu'ils appartiennent au même univers — et quels sont ceux qui s'excluent mutuellement ? Par exemple, quand l'auteur dit "je", quels verbes préfère-t-il et lesquels évite-t-il ? On se posera la même question pour les adjectifs, les substantifs, etc. Cette opération apportera un avantage subsidiaire en offrant une vision synthétique du "contexte" de certains mots. En effet, pour connaître le sens qu'un auteur donne à certains mots polysémiques, il faut relever les emplois qu'il en fait et les comparer au reste de son vocabulaire. Pour les "grands" mots de la politique — démocratie, peuple, liberté, nation… — cette possibilité de clarification n'est pas négligeable !

Ces idées ne sont pas neuves mais les tentatives passées d'application statistique ont mis en lumière plusieurs difficultés. Citons les "cooccurrences", les "rafales" et les "segments répétés" développés au laboratoire de Saint-Cloud (Lafon 1981 et 1984 ; Tournier, 1985 ; Salem, 1987). Dès que sont abordés de vastes corpus, les cooccurences débouchent sur des tableaux pléthoriques et des "lexicogrammes" fort complexes. La notion de "segments répétés" représente un progrès mais se heurte aussi à quelques problèmes :

— Beaucoup de formes sont ambiguës (Salem, 1987, p 158-170). C'est pourquoi nous proposons de travailler sur des *corpus lemmatisés* (Muller, 1977 ; Labbé, 1990a). Dans la suite de cette communication, nous utilisons donc la terminologie et la notation proposée par Muller.

— Beaucoup de segments répétés proviennent simplement des contraintes "formelles" ou "syntaxiques" du français (par exemple, le substantif est

généralement précédé d'un déterminant ou d'une préposition). Autrement dit, le calcul doit neutraliser, autant que possible, la *composante syntaxique* — puisqu'elle pèse également sur tous les usagers de la langue —, pour mesurer la seule *composante stylistique* ou *lexicale* propre à l'auteur étudié

— Enfin, dès que l'on dépasse l'étude du "slogan", la construction des syntagmes est rarement figée : des locutions toutes faites, des inversions, des incidentes peuvent rompre la contiguïté entre les éléments sans pour autant détruire l'association syntagmatique (ainsi l'auxiliaire et le verbe sont rarement contigus dans les textes un peu élaborés). En définitive, dans les subtiles variations du maître de rhétorique de Monsieur Jourdain, on ne trouvera que deux segments répétés — "belle marquise" et "beaux yeux" — alors que, naturellement, le traitement statistique adéquat sera celui qui reconnaîtra, dans toutes les phrases, une même association entre "beau", "marquise", "oeil", "mourir" et "amour"… Par conséquent, *l'environnement pertinent d'une occurrence d'un mot, c'est la phrase* dans laquelle il se trouve sans que soit prise en compte la construction de cette phrase.

Le procédé de calcul et l'application que nous présentons ci-dessous sont directement déduits de ces quelques postulats.

I. CALCUL DES LIAISONS ENTRE LES MOTS

Considérons, dans un corpus composé de N mots, l'ensemble P formé des phrases contenant le vocable dont on recherche les associations. Cet ensemble comporte N_p mots. Le vocable a une fréquence absolue F dans le corpus et F_p dans P. Si la loi qui règle l'apparition des mots dans le corpus est la même que celle en oeuvre dans P, la fréquence des mots dans le sous-corpus sera proportionnelle à leur fréquence totale pondérée par la taille de P. L'espérance mathématique d'apparition du vocable considéré dans le sous-corpus P est de :

$$(1) \quad E_p = F * \frac{N_p}{N}$$

E_p devant nécessairement être un entier, la formule (1) devient :

$$(2) \quad \frac{(F+1)(N_p+1)}{N+2} - 1 \leq E_p \leq \frac{(F+1)(N_p+1)}{N+2}$$

A quel moment peut-on dire que F_p s'écarte significativement, en plus ou en moins de E_p ? Le raisonnement probabiliste est maintenant familier en statistique lexicale. La discussion a longtemps porté sur la loi à utiliser pour le calcul (normale, binomiale, hypergéométrique, Poisson…) Un accord presqu'unanime se dégage en faveur de la loi hypergéométrique sous certaines réserves techniques en grande partie levées grâce à la puissance de calcul des machines actuelles (Brunet, 1980). Nous utilisons ci-dessous la formulation donnée par P. Lafon (pour mesurer la répartition d'une forme dans les différentes parties d'un corpus : Lafon, 1984 et, pour une discussion : Labbé, 1994).

Soit X la variable aléatoire mesurant le nombre d'apparitions du vocable considéré dans P. On calcule la probabilité pour que ce mot apparaisse K fois :

$$(3)\quad P(X=K) = \frac{\binom{F}{K}\binom{N-F}{N_p-K}}{\binom{N}{N_p}}$$

Les factorielles de la formule (3) ne peuvent être utilisées directement à cause des grands effectifs des corpus linguistiques. Pierre Lafon a proposé une formule aisément programmable :

$$(4)\ \log P(X=F_p) = \log F! + \log(N-F)! + \log N_p! + \log(N-N_p)! - \log N! - \log F_p! - \log(F-F_p)! - \log(N_p-F_p)! - \log(N-F-N_p+F_p)!$$

Si la fréquence observée (F_p) n'est pas égale à la fréquence attendue (E_p), la probabilité pour qu'on rencontre une fréquence telle que celle observée sera (selon la procédure de calcul proposée par P. Lafon) :

$$\text{avec } F_p > E_p : P(X \geq F_p) = \sum_{k=F_p}^{Min(F,N_p)} P(X=k)$$

$$\text{avec } F_p < E_p : P(X \geq F_p) = \sum_{k=0}^{F_p} P(X=k)$$

Puisqu'on ne calcule pas la probabilité de l'ensemble des possibles mais, au plus, la moitié de ceux-ci (quand la valeur de F_p est très proche de celle de E_p), la valeur maximum de la probabilité est de 0.5.

Si cette probabilité est inférieure à un certain seuil choisi par l'analyste — généralement 0.01 ou 0.05 —, on dira que les deux mots analysés sont reliés positivement ou négativement suivant les cas. Le calcul nous permet donc de "prédire" que, dans le corpus, lorsque l'auteur emploie tel mot, il l'associera avec tel ou tel autre ou, au contraire, il évitera d'autres mots qui font pourtant partie de son vocabulaire mais qui appartiennent à d'autres univers…

Dans la formulation donnée par P. Lafon, la présentation des résultats pose deux problèmes. Premièrement, la liaison entre les mots est d'autant plus forte que la probabilité est faible, ce qui peut surprendre le lecteur. Deuxièmement, la probabilité nulle est, en gros, égale à 0.5 et non à 1 comme on pourrait s'y attendre en fonction de la proposition précédente. Pour faciliter la lecture et l'interprétation des résultats, nous proposons de les formuler sous la forme d'un indice de liaison (L) qui fluctue entre 0 — pour l'absence de liaison — et ± 1 pour exprimer une liaison quasi-stochastique de même sens ou de sens inverse (Labbé, 1994). Ceci peut être obtenu grâce à une formulation simple :

$$L = \frac{0{,}5 - P}{0{,}5}$$

Si l'on choisit un seuil de 1%, la liaison positive sera attestée avec $L \geq 0{,}99$ et la liaison négative avec $L \leq -0.99$. Avec un seuil de 5%, ces valeurs seront de ±0.95,

II. APPLICATION

Nous avons appliqué ce calcul au vocabulaire du général de Gaulle entre 1958 et 1969. Le corpus se compose des allocutions et des conférences de presse, soit 79 textes, 201.907 mots dont 12.640 différents ("formes") et 6480 vocables.

On trouvera en annexe les résultats pour deux des trois pronoms personnels les plus significatifs (les premières personnes du singulier et du pluriel) ainsi que pour "France" qui est le substantif le plus employé par le général de Gaulle. Ces mots forment les noyaux de trois des principaux univers de son vocabulaire. Les tableaux sont limités aux liaisons significatives au seuil de 1%.

En tête des trois tableaux se trouvent logiquement un certain nombre de vocables. Par exemple l'article défini devant *France* ou encore avec "je" : l'article possessif *mon,* les pronoms *mien* et *moi-même*. Avec "nous" : *nôtre, nous-même* mais aussi : *chez nous, avec nous* et *notre siècle* qui sont de véritables stéréotypes dans la bouche du Général. Leur présence en tête des tableaux prouve que le procédé restitue bien les segments répétés et qu'il le fait de manière synthétique.

Les interlocuteurs privilégiés du "je" sont *Madame* et *Monsieur* : la quasi-totalité des *Madame* et 60% des *Monsieur* sont utilisés dans des phrases construites avec "je" alors qu'on en attendrait seulement 17% si la répartition obéissait au hasard. C'est ainsi que le Général s'adresse aux journalistes qui le questionnent. D'où également la présence, dans le premier tableau, de : *question, réponse, sujet, poser* et *répondre*.

Dans ses allocutions, de Gaulle s'adresse aux *Françaises* et aux *Français* — qui figurent également en associations positives —, mais surtout à *vous* et il s'agit d'abord de : *je vous demande votre confiance*. Les phrases construites avec la première personne servent donc d'abord à interpeller les auditeurs. Cette «tension» du "je" et du "vous" avait déjà été notée par Cotteret et Moreau (1967) bien qu'ils aient travaillé sur les allocutions entières — et non sur chaque phrase — et sur un corpus de seulement 70.000 mots. C'est pour cette raison qu'ils avaient baptisé les allocutions où la première personne dominait : les «discours appel».

En laissant de côté, les voeux de nouvel *an,* les substantifs qui gravitent autour du "je" dessinent un univers essentiellement politique : *gouvernement, président, élection, constitution, ministre, fonction, mandat, vote*… ou fortement valorisé : *foi, honneur, esprit, mission,* etc. Une date est reliée à la personne du Général : *mille neuf cent quarante* (il estime incarner la "légitimité nationale" depuis son appel du 18 juin).

Les verbes liés à la première personne se répartissent en trois ensembles. Le premier, par l'importance, est constitué de verbes d'«énonciation». Certains de ces verbes s'expliquent par l'interraction des questions et des réponses lors des conférences de presse (*répondre, répéter, poser, remercier*…). Il faut y ajouter des associations avec *dire, parler, (s')adresser, déclarer,* etc. Incontestablement, dans

l'esprit du fondateur de la Ve République, la *fonction* présidentielle consiste d'abord à *parler*. On retrouve d'ailleurs ce trait dans le vocabulaire de F. Mitterrand (Labbé, 1990a). A cela s'ajoute la dimension de la connaissance ou de la pensée : *croire, savoir, penser, douter, expliquer, connaître,* etc… La troisième dimension est celle de la volonté : *vouloir, espérer, souhaiter*… Cependant, la fonction présidentielle se réduit-elle à parler et à commander ? La présence du verbe *faire* au bas du premier tableau peut en faire douter (d'ailleurs, quand on est investi de l'autorité légitime, *dire* n'est pas déjà *faire* ?) A contrario, on relèvera au moins la présence de *action* dans les vocables significativement sous-employés avec la première personne (tableau 4). Enfin, les verbes à la première personne présentent deux caractéristiques intéressantes. Premièrement, la forte présence de l'auxiliaire *avoir* indique que de Gaulle privilégie le passé composé quand il parle à la première personne. Deuxièmement, avec "je", le Général utilise beaucoup la construction négative (*ne… pas*). Dans la théorie transformationnelle, la négation signale un exposé construit en opposition à un énoncé prononcé antérieurement par d'autres. Cette forte utilisation de la négation peut être la marque d'un discours pédagogique (on argumente contre les idées fausses) ou d'une sensibilité aux critiques contre ce qu'il a *dit, déclaré, voulu* ou *fait* auparavant (d'où le passé).

Il faudrait encore signaler tous les vocables qui ne sont ni attirés ni repoussés par la première personne et que l'on pourrait baptiser "neutres". Entre autres, chez de Gaulle : *chef de l'Etat* et *président de la République* qui seront les deux syntagmes préférés de F. Mitterrand quand il dira "je" (ce qu'il fait d'ailleurs beaucoup plus que de Gaulle)…

Les associations négatives sont aussi importantes. Comme l'indique le tableau 5, beaucoup de mots que le général évite, quand il parle à la première personne, apparaissent dans les liaisons positives du *nous.* Les univers des premières personnes du singulier et du pluriel sont presque parfaitement symétriques, opposés, exclusifs l'un de l'autre.

L'univers du *nous* est, en gros, constitué de deux domaines distincts.

Le premier domaine du *nous* comporte les principales dimensions du vocabulaire *économique* et *social* : le *progrès (technique),* les *moyens,* la *production (industrielle),* l'*économie,* les *mesures,* l'*inflation,* la *recherche,* l'*industrie,* la *finance,* les *échanges,* la *concurrence*… Dans ce cas, la première personne du pluriel signifie : *nous les Français*. On a prêté au Général la formule "l'intendance suivra" qu'il a récusée en disant qu'il s'agissait de "blagues pour journalistes". Il n'en reste pas moins que c'est *nous* qui avons en charge les questions économiques et sociales et que le "je" est significativement peu présent lorqu'il est question d'intendance !

Le second domaine du *nous* est celui de la politique étrangère, de l'*international* : la *défense* du *territoire,* la *coopération,* les *grandes puissances,* le *militaire,* l'*Allemagne,* l'*Asie,* l'*Afrique,* l'*Amérique,* l'*univers*… La première personne du pluriel a ici une nuance différente : il s'agit de "nous, la France". Au passage, cet exemple montre que l'on ne peut s'en tenir au constat statistique : suivant les contextes, une liaison négative peut signifier exclusion mutuelle — dans

ce cas, il faut que les univers soient effectivement antinomiques — ou substitution possible et dans ce cas, une intersection plus ou moins importante doit exister entre les deux univers. Tel est le cas entre *France, nous* et *Français* comme l'indiquent les tableaux 6 à 8 en annexe.

Cinq formules clefs illustrent la conception gaullienne de la France : *la France vit avec son temps, le destin de la France, l'indépendance de la France, au nom de la France* et, enfin, *une politique digne de la France.* Pour les verbes : la France est *concernée* et elle *proclame.* Enfin, le président évoque souvent les *rapports entre la France et… les Algériens, l'Italie,* et en prolongeant le tableau : les *Etats-Unis* (0.98), l'*Afrique* (0.97) et l'*Allemagne* (0.97).

Enfin, *nation* n'est pas associée à *France* dans le vocabulaire de de Gaulle et *national* figure dans les associations négatives. La présence, dans le dernier tableau en annexe, d'une partie de l'univers du *nous,* ainsi que celle du vocabulaire politique propre à au *je* (les élections, le gouvernement, le pouvoir, la constitution), souligne les intersections existant entre les trois univers. Ces intersections sont logiques étant donné la philosophie structuraliste qui soutend la notion de champ et inspire les calculs. Ces complémentarités rappellent aussi la fonction des pronoms dans la langue et permettent de comprendre le sens que de Gaulle donne à ces "symboles vides" dont parlait Benveniste.

Nous espérons que ces quelques résultats auront convaincu de l'intérêt de la méthode proposée. Quant au fond, ils ne surprendront pas trop puisqu'ils recoupent les conclusions des recherches sur le vocabulaire ou sur l'énonciation chez de Gaulle (Cabasino, Cotteret-Moreau). En définitive, cette convergence est plutôt rassurante : la statistique textuelle peut apporter des instruments puissants sans remettre en cause les démarches traditionnelles.

III. DISCUSSION

Les calculs, encore expérimentaux, méritent une discussion.

En premier lieu, le procédé ne s'applique qu'aux textes ponctués d'une manière traditionnelle, ce qui est évidemment le cas du général de Gaulle ou de F. Mitterrand. Il sera plus difficilement applicable à des transcriptions de l'oral et impossible à utiliser sur des corpus de type "nouveau roman".

En second lieu, on peut se demander pourquoi avoir recours à la lemmatisation et aux catégories grammaticales ? L'opération est nécessaire pour l'étude des pronoms et des verbes : le calcul effectué sur les "formes" apprendra seulement que les verbes associés aux pronoms personnels sont accordés avec eux ! Il en est de même pour les principaux déterminants avec les substantifs, etc. Au-delà de ces évidences, les catégories grammaticales jouent un rôle clef dans le calculs eux-mêmes. Par exemple, dans une phrase où figure un pronom personnel sujet, il y aura une certaine densité de verbes qui ne sera probablement pas celle du corpus entier. Par exemple, si les phrases avec "je" contiennent 20% de verbes — contre seulement 15% dans le corpus entier — il faut augmenter d'un quart l'espérance

mathématique des verbes et pondérer en conséquence celle des substantifs, des adjectifs, des pronoms, etc.

Cette intuition est-elle vérifiée dans la pratique ? Le tableau 1 ci-dessous propose une vérification empirique sur les phrases de de Gaulle contenant les pronoms "je" et "nous".

Les phrases avec "je" représentent 19,2% du total (coefficient de proportionnalité) mais elles contiennent 24,6% des pronoms, 21,9% des verbes et, à l'opposé, seulement 15,3% des adjectifs. Autrement dit, en donnant aux autres pronoms une probabilité de co-occurence avec "je" de 19,2%, on sous-estimerait leur espérance mathématique de près de 30%. Cette sous-estimation serait encore de 14% pour les verbes. En revanche, on surestimerait de 20% la densité probable des adjectifs et celle des substantifs de 11%. Etant donné la sensibilité du test proposé ci-dessus, le biais serait considérable : un grand nombre de verbes et de pronoms se retrouveraient dans les liaisons positives et les liaisons négatives seraient emplies d'adjectifs et de substantifs…

Tableau I. Les principales catégories grammaticales dans les phrases avec "je" et "nous" dans le vocabulaire du général de Gaulle

	"Je"	"nous"
Nombre de phrases dans le corpus	6 592	6 592
Nombre de phrases contenant le pronom (P)	1 149	1152
Nombre de mots dans P (N_p)	38 688	44 384
Coefficient de proportionnalité (Np/N)	*0.192*	*0.220*
Proportion des verbes dans P	0.219	0.220
Proportion des substantifs dans P	0.171	0.211
Proportion des adjectifs dans P	0.153	0.220
Proportion des pronoms dans P	0.246	0.243
Proportion des adverbes dans P	0.198	0.225
Proportion des déterminants dans P	0.176	0.206
Proportion des conjonctions dans P	0.184	0.224

De plus, ces biais n'ont rien de systématique comme le suggère la comparaison entre les colonnes du tableau 1. Dans les phrases contenant "nous", les densités des diverses catégories s'écartent peu du coefficient moyen (22%) et pas forcément dans le même sens qu'avec "je" : les conjonctions et les adverbes sont légèrement sur-employés avec "nous" alors qu'ils sont sous-employés avec "je". La densité des différentes catégories grammaticales varie donc de manière imprévisible. Par conséquent, dans la formule (2) ci-dessus, N et Np seront remplacés par l'effectif de la catégorie grammaticale C à laquelle appartient le vocable considéré dans le corpus entier (N_c) et dans P (N_{cp}).

Une troisième question mérite d'être signalée, même si sa discussion dépasse la taille de cette contribution : quelle est la fréquence minimale à partir de laquelle le calcul peut être fait ? La solution que nous avons adoptée est la suivante : le calcul ne porte que sur les vocables dont la fréquence totale est au moins égale à une

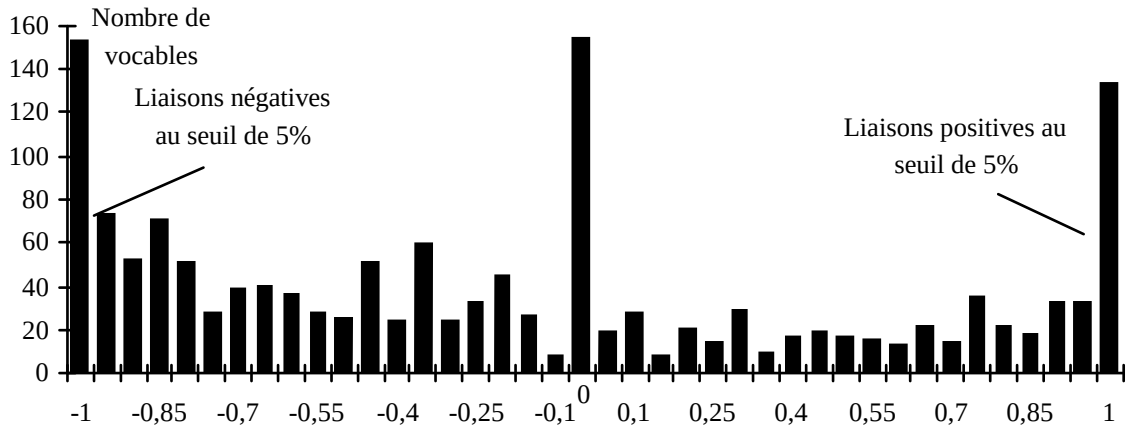
valeur telle qu'une absence dans P aboutisse à un indice négatif significatif au seuil choisi. En effet, en dessous de cette fréquence, seule la liaison positive devient concevable ce qui introduirait une asymétrie fâcheuse dans la structure étudiée. Ce seuil de liaison négative nous a été inspiré par André Salem (Salem, 1987). En général, pour le corpus de Gaulle, cela représentait une espérance mathématique dans P au moins égale à 3. Naturellement, ce choix limite le calcul aux vocables les plus fréquents qui ne sont pas forcément considérés comme les plus caractéristiques du point de vue rhétorique et stylistique.

Une quatrième difficulté tient à l'influence de la fréquence sur l'indice. En effet, plus on s'élève dans l'échelle des fréquences, plus la probabilité est forte de rencontrer des indices élevés. Plusieurs explications sont possibles. En premier lieu, l'influence de la loi hypergéométrique : au fur et à mesure qu'augmente le nombre des épreuves, la loi prédit un rapprochement rapide de la moyenne observée avec l'espérance mathématique. Cette "convergence en probabilité" s'explique par le fait que la loi hypergéométrique décrit les résultats d'un tirage exhaustif (ou "sans remise"). Deuxièmement, plus la fréquence augmente, moins le hasard a de place dans l'utilisation des mots. Ceci peut être exprimé ainsi : pour les basses fréquences, le nombre des combinaisons possibles entre les mots — calculé à l'aide de la loi hypergéométrique — ne s'éloigne pas trop du nombre de combinaisons réalisables en respectant les règles du français. En revanche, pour les hautes fréquences, le nombre de combinaisons statistiquement possibles devient immense et la plupart de ces combinaisons ne sont pas réalisables à cause du projet de l'auteur et de la syntaxe du français.

L'explication réside aussi dans le principe même du calcul : plus un mot est employé, plus il a de chances de se trouver combiné avec d'autres et, par là-même, plus ses relations avec le reste du vocabulaire seront repérables.

Enfin, les valeurs de l'indice et le nombre des liaisons posent également problème. En effet, dans une population soumise au seul hasard, 1% à 5% des individus — selon le seuil choisi — devraient sortir de la distribution "normale" par le haut et par le bas. Par exemple, pour le "je", le calcul a porté sur 1 553 vocables dépassant le seuil de spécificité négative. Au lieu des 16 ou 17 liaisons caractéristiques que l'on attendrait dans une distribution "normale", le calcul en fait apparaître 75 au seuil de 1%. Il est vrai que, chez de Gaulle, le pronom "je" est, de loin, le vocable qui suscite le plus grand nombre de liens mais les autres vocables usuels dépassent aussi nettement ce que donnerait une répartition aléatoire. La distribution des vocables en fonction de l'indice de liaison n'a que peu de ressemblance avec la célèbre "courbe en cloche" (tableau 2).

Tableau 2. Distribution des vocables en fonction de leur indice de liaison dans l'univers du "je" chez le général de Gaulle.



Au total, au seuil de 5%, 18% des vocables soumis au calcul, sont liés significativement avec le pronom "je". Mais le graphique signale aussi l'importance des vocables dont la fréquence observée ne s'écarte pas ou très peu de la fréquence calculée (ce sont généralement des mots de fréquence faible).

En définitive, ce que révèlent les calculs d'inspiration probabiliste, comme celui qui vient d'être présenté, c'est le caractère fondamentalement non-aléatoire du discours : les mots ne viennent pas au hasard, leurs associations sont fortement contraintes à la fois par les choix de l'auteur — que nous voulons mettre au jour —, et les règles du français que nous cherchons à neutraliser dans la mesure du possible puisque nous ne sommes pas lexicologues !

En conclusion, nous voudrions évoquer l'importance des normes de dépouillement. La norme utilisée pour dépouiller le corpus doit être clairement précisée et stable tout au long des opérations si l'on veut que les résultats obtenus aient un sens. Cela est d'autant plus important que, étant donné la sensibilité de l'indice proposé, des différences assez minimes peuvent être significatives. Or la statistique textuelle est confrontée à une pluralité de normes de dépouillement et à un grand laxisme dans leur utilisation : à quoi sert de développer des modèles et des calculs sophistiqués s'ils sont ensuite appliqués à des matériaux trop peu sûrs ?

BIBLIOGRAPHIE

Brunet, E. (1980). Loi hypergéométrique et loi normale. Comparaison dans les grands corpus. *Actes du second colloque de lexicologie politique* (Saint-Cloud, septembre 1980). Paris. Klincksieck, tome III, p 699-716.

Cabasino, F. (1983). *Malraux e de Gaulle*. Rome, Bulzoni.

Cotteret, J.-M., Moreau R. (1969). *Recherches sur le vocabulaire du général de Gaulle*. Paris, Presses de la Fondation nationale des sciences politiques.

Labbé, D. (1990a). *Normes de dépouillement et procédures d'analyse des textes politiques*. Grenoble, CERAT.

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

Labbé, C., Labbé, D. (1994). *Que mesure la spécificité du vocabulaire ?*. Grenoble, CERAT.

Lafon, P. (1981). Analyse lexicométrique et recherche des coocurrences. *Mots*, 3, octobre 1981, p 95-148.

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

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

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

Salem, A. (1987). *Pratique des segments répétés. Essai de statistique textuelle*. Paris, Klincksieck.

Salem, A. (1993). *Méthodes de la statistique textuelle*. Thèse pour le doctorat d'Etat. Paris, Université de Paris III, 3 tomes.

Tournier, M. (1985). Textes propagandistes et coocurences. Hypothèses et méthodes pour l'étude de la sloganisation. *Mots*, 11, octobre 1985, p 155-187.