Je est-il un autre ?

Dominique Labbé[1], Denis Monière[2]

[1]Institut d'Etudes Politiques – 38 040 Grenoble Cedex 9
dominique.labbe@iep-grenoble.fr

[2]Université de Montréal
denis.moniere@umontreal.ca

Abstract

How can we compare the frequencies of a single word within two different corpora? We present a method in order to decide whether these two frequencies are statistically significant or not. A Parametric bilateral test is presented with the example of the first person pronouns used by the four major candidates who were running for the 2007 French presidency (F. Bayrou, J.-M. Le Pen, S. Royal, N. Sarkozy). Theoretical and empirical values are compared. It appears that the density of first person pronouns is not dependant on the individual personality or style. It varies with the candidate strategy, the topics and the circumstances of their speeches.

Résumé

Comparaison des fréquences d'un même mot dans deux corpus. On présente une méthode pour décider si les écarts constatés entre ces deux fréquences sont statistiquement significatifs. Les tests issus de la loi normale sont comparés avec les variations constatées dans l'utilisation de la première personne par les 4 principaux candidats à l'élection présidentielle française de 2007 (F. Bayrou, J.-M. Le Pen, S. Royal, N. Sarkozy). Il apparaît que la densité de la première personne n'est pas déterminée par des personnalités ou des styles propres aux individus mais par les stratégies de communication, le sujet traité et la situation d'énonciation.

Mots-clés : discours politique, pronom personnel, test paramétrique bilatéral, énonciation de la subjectivité dans le discours.

1. Introduction

L'énonciation est l'acte individuel d'un sujet qui utilise la langue pour produire un discours (Benveniste, 1970). Elle peut être analysée grâce aux marques que le sujet imprime sur ce discours. Ces marques signalent, d'une part, la *distance* que le sujet établit avec l'objet dont il parle et avec le monde qui l'entoure et, d'autre part, la *tension* que ce sujet entretient avec le ou les destinataires de son discours. L'analyse de l'énonciation s'appuie essentiellement sur les pronoms, les déictiques, les marques temporelles, l'aspect des verbes et leur modalisation (Kerbrat-Orrecchioni 1980, Maingueneau, 1994). La statistique lexicale peut intégrer certaines de ces notions (voir notamment : Labbé 1981 et 1990b).

La présente étude porte sur les pronoms et plus spécifiquement, les pronoms de la première personne. La théorie de l'énonciation considère les pronoms comme des « formes vides » dont le sens provient essentiellement de l'acte d'énonciation d'un locuteur dans une situation donnée. Nous proposons de tester cette théorie à l'aide d'un corpus politique original : les 132 discours de F. Bayrou, J.-M. Le Pen, S. Royal et N. Sarkozy - lors de la campagne présidentielle de 2007 – placés en ligne sur les sites officiels de ces candidats. Le tableau 1 récapitule les principales caractéristiques de ce corpus.

	Nombre de discours (T)	Nombre de mots (N)	Formes différentes	Vocables (V)
Bayrou	28	227 446	11 200	6 396
Le Pen	21	87 114	10 961	7 298
Royal (1er tour)	31	146 687	10 206	6 103
Royal (2e tour)	8	48 269	5 093	3 255
Royal (total)	39	194 956	11 632	6 738
Sarkozy (1er tour)	36	231 024	12 289	7 215
Sarkozy (2e tour)	8	68 843	6 414	4 059
Sarkozy (Total)	44	299 867	13 817	7 879
Total présidentielles	132	809 383	25 419	13 658

Tableau 1. Le corpus de la campagne présidentielle de 2007

Textes déchargés par D. Monière sur les sites des candidats à partir de leur déclaration de candidature. Les chiffres des trois dernières colonnes sont susceptibles de changer de quelques unités (corpus en cours de révision).

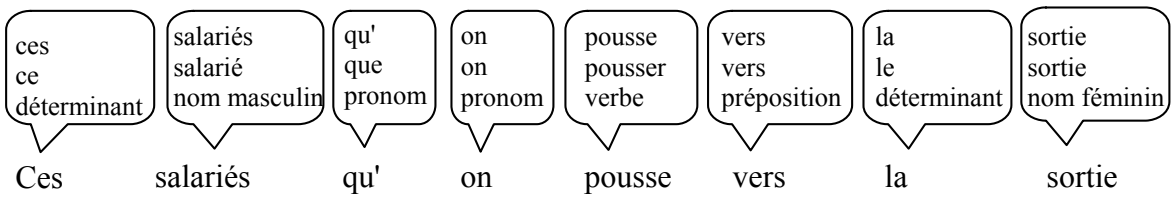
2. Traitements préalables

Ces textes ont fait l'objet d'une série de traitements préalables (pour le détail de ces traitements : Labbé 1990a).

En premier lieu, l'orthographe a été soigneusement corrigée et les graphies multiples ont été standardisées (événement et évènement ; puis et peux, etc.), tout particulièrement les sigles, les abréviations, les noms propres, les chiffres et les dates dont la transcription est d'une infinie variété… Ces tâches sont partiellement effectuées par des automates, mais les interventions manuelles sont nécessairement nombreuses et suivent des règles bien précises.

En second lieu, des balises indiquent les sources du texte puis délimitent les séquences (début et fin des propos des orateurs, interruptions, questions et réponses pour les interviews).

Enfin, la lemmatisation attache à chacun des mots du texte une étiquette contenant la forme graphique normalisée et l'entrée sous laquelle le mot peut être retrouvé dans un dictionnaire. L'exemple ci-dessous est tiré du début de la première déclaration de S. Royal après son investiture par le PS (17 novembre 2006).



Dans le bref exemple ci-dessus, 6 des 8 mots sont ambigus (« homographes » : une seule graphie mais plusieurs entrées possibles dans le dictionnaire) :

- salariés : verbe « salarier » au participe passé ou nom masculin

- que : pronom ou conjonction de subordination

- pousse : verbe « pousser » ou nom féminin

- vers : noms masculins - ver (de terre) ou vers (du poète) - et préposition

- la : pronom, article ou substantif masculin (note de musique)

- sortie : verbe « sortir » au participe passé, adjectif ou nom féminin.

Un tel cas est banal : dans tout texte écrit en français, plus du tiers des mots peuvent être rattachés à plus d'une entrée de dictionnaire.

L'étiquette comporte trois informations : la forme standardisée – majuscule initiale des mots communs ramenée en minuscule, réduction des formes multiples à une graphie standard, correction des fautes d'orthographe… - puis le vocable – c'est-à-dire l'entrée où se trouve la forme dans le dictionnaire et enfin la catégorie grammaticale, telle qu'elle figure en seconde position de l'entrée de dictionnaire. Ainsi, les conjugaisons d'un même verbe sont groupées sous son infinitif ou les pluriels du substantif sous le singulier ou encore les féminins et pluriels de l'adjectif sous le masculin singulier. Par exemple, « être v. » regroupe toutes les formes conjuguées de ce verbe, tandis que « être n. m. » ne se rencontre qu'avec le singulier et le pluriel. La nomenclature des mots, apprise à l'ordinateur, est systématique (par exemple, en français, les substantifs se distinguent par le genre, donc tous les substantifs doivent se voir affecter le masculin ou le féminin), elle est exhaustive (tous les mots doivent y trouver leur place), elle est univoque (une seule entrée par mot) elle exclut tout double compte, elle ne comporte pas de catégorie ad hoc, ou fourre-tout, etc. Enfin la lemmatisation est réversible : on peut retrouver le texte original, sans altération, à partir du fichier étiqueté.

Par rapport aux traitements sur les formes graphiques brutes, la normalisation et la lemmatisation donnent une existence aux verbes (en rassemblant leurs flexions sous une étiquette commune), ce qui permet de retrouver certains mots - comme le point cardinal « est », les substantifs « être », « avoir », « avions »… - dont les occurrences sont habituellement noyées dans l'océan des formes verbales homographes.

Les utilisations sont multiples. En premier lieu, le vocabulaire d'un corpus est aisément établi avec, sous les lemmes, les formes graphiques sous lesquelles les vocables sont attestés dans ce corpus (le tableau 2 donne un extrait de l'index du corpus). En second lieu, la normalisation et la lemmatisation rendent possibles de nombreux calculs comme, par exemple, celui des fréquences relatives des pronoms personnels (tableau 2). Les conventions de regroupement sont celles définies par Muller (1967) et Bernet (1983).

Vocables et formes	Total	Bayrou	Le Pen	Royal 1er tour	Sarkozy 1er tour	Royal 2e tour	Sarkozy 2e tour
je	20,0	18,5	4,6	19,4	18,7	25,8	26,1
j'	*3,0*	*3,3*	*0,8*	*2,6*	*2,0*	*4,2*	*4,7*
je	*14,8*	*13,2*	*3,1*	*14,5*	*14,7*	*19,3*	*17,9*
m'	*0,9*	*0,9*	*0,2*	*1,2*	*0,7*	*1,0*	*1,5*
me	*1,3*	*1,1*	*0,4*	*1,1*	*1,3*	*1,2*	*2,0*
il	13,1	13,9	8,8	9,3	12,4	11,8	13,8
elle	*3,0*	*1,9*	*2,2*	*2,7*	*4,0*	*1,7*	*3,4*
il	*10,0*	*12,0*	*6,6*	*6,6*	*8,5*	*10,2*	*10,4*
on	7,6	9,7	3,4	3,2	7,4	4,4	10,3
nous	8,4	11,0	5,0	9,2	5,4	7,1	4,8
vous	6,3	6,7	3,1	8,3	3,3	9,4	6,1
ils	4,9	6,1	4,0	3,7	3,6	4,4	4,1
elles	*0,8*	*0,8*	*0,4*	*0,9*	*0,5*	*1,3*	*0,4*
eux	*0,7*	*1,0*	*0,7*	*0,4*	*0,5*	*0,6*	*0,5*
ils	*3,4*	*4,3*	*2,9*	*2,3*	*2,6*	*2,5*	*3,1*

Tableau 2. Densité relative des principaux pronoms personnels (pour 1 000 mots)

La domination du « je » est le premier constat, ce qui justifie le choix de commencer l'analyse par ce pronom. Si l'on prend comme point de référence la moyenne du corpus total (deuxième colonne du tableau) on voit que J.-M. Le Pen sous-emploie tous les pronoms ; F. Bayrou privilégie le « nous » et le « on » ; les deux finalistes se situent dans la moyenne pour le premier tour mais augmentent considérablement le « je » au second tour. N. Sarkozy préfère le « on » au « nous » et S. Royal utilise beaucoup plus le « nous » et le « vous ».

Cependant, ces constats soulèvent un certain nombre de questions. Comment s'assurer que les différences entre les locuteurs ne sont pas simplement des fluctuations dues au hasard et, si ce n'est pas le cas, quelles sont les différences les plus significatives ?

3. Qui est différent ?

Il s'agit donc de porter un jugement sur une caractéristique inconnue - la propension à utiliser « je » chez deux individus différents – à partir d'un nombre limité d'observations réalisées avec les mêmes instruments (cf. la première partie de cette communication). La question revient à comparer le nombre d'apparitions d'un même mot *i* (n_{iA} et n_{iB}) dans deux corpus A et B, distincts et longs respectivement de N_A et N_B mots. Deux hypothèses sont à tester :

- H_0 : les deux fréquences relatives ne diffèrent pas significativement ($f_{iA} \approx f_{iB}$). Les deux locuteurs présentent une propension semblable à utiliser la première personne ;

- H_1 : les deux fréquences relatives diffèrent significativement ($f_{1A} \neq f_{iB}$). Les deux locuteurs n'ont pas la même propension à utiliser la première personne.

Les deux corpus sont considérés comme deux échantillons de N_A et N_B prélèvements indépendants opérés dans deux vastes populations indépendantes dont les paramètres (moyenne et variance) du caractère recherché (la première personne) sont inconnus. On définit un seuil α en dessous duquel on accepte un risque de se tromper (par exemple ici α = 0,05) en acceptant l'une des deux hypothèses alors que c'est l'autre qui est vraie. Naturellement, dire qu'une hypothèse est acceptée (avec un certain risque d'erreur) ne signifie pas qu'elle est « vraie » mais seulement que les observations disponibles ne sont pas incompatibles avec elle et que l'on n'a pas de raison de lui préférer l'hypothèse contraire (Desrosières 1988). Par exemple, accepter H_0 signifie qu'il y a tout lieu de penser que le phénomène est régi par les mêmes lois dans les deux corpus et que les différences constatées entre les deux fréquences sont dues aux fluctuations propres à tout phénomène naturel.

Ce *test paramétrique bilatéral* est organisé de la manière suivante (nous suivons les procédures usuelles en sciences de l'ingénieur, voir par exemple : CISIA-CERESTA (1995) ou Harris et Stocker (1998)).

En considérant H_0 (une seule population parente), l'espérance mathématique du « je » (f_{i0}) et l'écart type théorique ($\sigma theo_{i0}$) de cette variable seront estimés par :

$$(1)\ f_{i0} = \frac{n_{iA} + n_{iB}}{N_A + N_B} \text{ et } \sigma theo_{i0} = \sqrt{f_{i0}(1 - f_{i0})\left(\frac{1}{N_A} + \frac{1}{N_B}\right)}$$

Une région critique est définie autour de f_{i0} dont les bornes sont fixées à ± 1,96 $\sigma theo_{i0}$ autour de f_{i0} (schéma 1) (1,96 est tirée de la table de la loi normale pour α = 0,05).

A gauche, $|f_{i1} - f_{i2}| < 3{,}92\ \sigma theo_{i0}$: H_0 est acceptée ; H_1 est rejetée (avec un risque d'erreur de 5%).

A droite, $|f_{i1} - f_{i2}| > 3{,}92\ \sigma theo_{i0}$: H_0 est rejetée ; H_1 est acceptée (avec un risque d'erreur de 5%).

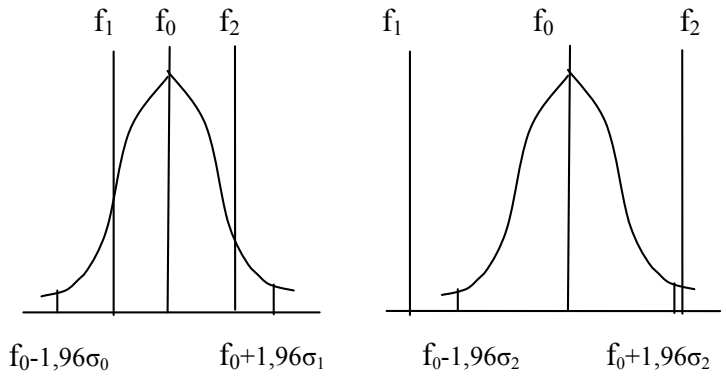


Schéma 1. Régions d'acceptation ou de rejet des hypothèses H_0 et H_1 (test paramétrique bilatéral)

Remarque : ce test postule implicitement que f_{i0} est sensiblement à mi-chemin entre f_{iA} et f_{iB}, c'est-à-dire que N_A et N_B ne sont pas trop différents. Quand ce n'est pas le cas et que $|f_{i1} - f_{i2}|$ est proche de $3{,}92\ \sigma theo_{i0}$, un autre test sera utilisé (cf. plus bas).

Voici le calcul réalisé sur les corpus Bayrou et Le Pen :

$$f_{i0} = \frac{4217 + 397}{227446 + 87114} = 0{,}014668$$

$$\sigma theo_{i0} = \sqrt{0{,}014668(1 - 0{,}014668)\left(\frac{1}{227446} + \frac{1}{87114}\right)} = 0{,}000479$$

$$|f_{iA} - f_{iB}| = \frac{4217}{227446} - \frac{397}{87114} = 0{,}01398$$

$$3{,}92\sigma theo_{i0} = 0{,}00188 < |f_{iA} - f_{iB}| \Rightarrow f_{iA} \neq f_{iB}$$

H_0 est rejetée ; H_1 est acceptée. On conclut avec moins de 5% de chances de se tromper que, lors de la campagne présidentielle de 2007, F. Bayrou a utilisé significativement plus la première personne que J.-M Le Pen.

Comparons maintenant F. Bayrou et S. Royal :

$$f_{i0} = \frac{4217 + 2846}{227446 + 146687} = 0{,}018888$$

$$\sigma theo_{i0} = \sqrt{0{,}018888(1 - 0{,}018888)\left(\frac{1}{227446} + \frac{1}{146687}\right)} = 0{,}00456$$

$$|f_{iA} - f_{iB}| = \frac{4217}{227446} - \frac{2846}{146687} = 0{,}00086$$

$$3{,}92\ \sigma theo_{i0} = 0{,}00179 > |f_{iA} - f_{iB}| \Rightarrow |f_{iA} = f_{iB}|$$

H_0 est acceptée ; H_1 est rejetée. Lors de la campagne présidentielle de 2007, F. Bayrou et S. Royal ont utilisé la première personne avec une densité *probablement* semblable (« probablement » signifie : « en acceptant un risque d'erreur de 5% »).

Le même calcul est répété pour tous les couples possibles. Le tableau 3 résume les conclusions de ces tests. Le signe ≈ signifie que les fréquences d'emploi du « je » ne diffèrent pas significativement dans les deux corpus et le signe ≠ la situation inverse.

	Sarkozy (2e tour)	Royal (2e tour)	Sarkozy (1e tour)	Royal (1e tour)	Le Pen
Bayrou	≠	≠	≈	≈	≠
Le Pen	≠	≠	≠	≠	
Royal (1er tour)	≠	≠	≈		
Sarkozy (1er tour)	≠	≠			
Royal (2e tour)	≈				

Tableau 3 Comparaison des fréquences d'emploi de la première personne. Test paramétrique bilatéral avec les formules (1)

Avant de commenter ces résultats, il faut se demander si ce test est adapté à la situation analysée et si ses résultats « collent » avec les observations disponibles.

4. Contrôle sur les données empiriques

L'interrogation porte sur l'estimation de la variance : dans la réalité, les fluctuations de densité de la première personne, au sein d'un corpus homogène à auteur unique, correspondent-elles au modèle théorique de la loi normale ? La campagne présidentielle de 2007 donne la possibilité de répondre à cette question. Pour cela, on cherche si les fluctuations de la fréquence du vocable *i* dans le corpus *A* ($\sigma\, obs_{iA}$) correspondent au modèle théorique utilisé ci-dessus et dont les résultats sont résumés dans le tableau 3 ($\sigma\, theo_{iA}$). Soit :

N_A : longueur du corpus *A* (nombre de mots : deuxième colonne du tableau 1)

T_A : le nombre de textes contenus dans le corpus A : première colonne du tableau 1

$N_{j\in A}$: longueur du texte de rang *j* dans le corpus A : première colonne du tableau 4

f_{iA} : fréquence relative du vocable *i* dans l'ensemble du corpus A (tableau 2)

$f_{ij\in A}$: fréquence relative du vocable *i* dans le texte de rang *j* dans le corpus A

$n_{ij\in A}$: nombre d'occurrences du vocable *i* dans le texte de rang *j* dans le corpus A : tableau 4

$E_{ij\in A}$: espérance mathématique du nombre d'occurrences du vocable *i* dans le texte *j*, en fonction de sa fréquence dans l'ensemble du sous-corpus A :

$$E_{ij\in A} = n_{iA}\,\frac{N_{j\in A}}{N_A}$$

$$(2)\ \sigma_{theo_{iA}} = \sqrt{\frac{f_{iA}(1-f_{iA})}{N_A}}\ \ et\ \ \sigma_{obs_{iA}} = \sqrt{\frac{\sum_{j=1}^{J=T_A}(n_{ij\in A} - E_{ij\in A})^2}{N_A}}$$

Le tableau 4 détaille le calcul de l'écart-type empirique (σobs_{iA}) sur les 28 interventions de F. Bayrou et compare ces résultats avec l'écart-type attendu ($\sigma theo_{iA}$) en utilisant les formules (2).

Il apparaît que les fluctuations empiriques sont nettement plus importantes que ce que laisse attendre le modèle de la loi normale (NB : nous avons choisi F. Bayrou car il est connu pour écrire lui-même ses discours). Autrement dit, dans le corpus Bayrou, la fréquence d'emploi du « je » est soumise à d'autres facteurs perturbateurs que le simple hasard.

Le même phénomène est observable dans les 5 autres corpus comme l'indiquent les résultats présentés dans le tableau 5 : l'écart-type empirique excède toujours les valeurs attendues.

Texte *j* (année mois jour)	N_j (nombre de mots dans le texte j)	n_{ij}(fréquence absolue de « je » dans le texte j)	$(n_{ij}-E_{ij})^2$
2006 12 02	1966	22	208,83
2006 12 14	7916	135	138,48
2006 12 19	6419	166	2207,82
2007 01 09	3141	41	297,09
2007 01 10	3357	42	409,70
2007 01 20	2656	49	0,06
2007 01 25	10113	234	2162,09
2007 02 07	11643	171	2013,22
2007 02 12	11977	186	1300,43
2007 02 16	8894	195	905,97
2007 02 17	3748	71	2,28
2007 02 27	9463	170	29,71
2007 03 01	8755	210	2273,05
2007 03 05	8680	127	1151,45
2007 03 12	12104	240	242,85
2007 03 16	12034	275	2691,70
2007 03 17	13307	221	661,55
2007 03 23	10293	203	147,89
2007 03 25	4848	59	953,89
2007 03 26	7464	128	107,90
2007 03 30	8939	174	68,31
2007 04 04	8923	150	238,34
2007 04 05	10574	181	226,47
2007 04 10	8752	146	264,64
2007 04 17	11381	192	361,43
2007 04 20A	9432	216	1691,22
2007 04 20B	10176	197	69,39
2007 04 22	491	16	47,56
Totaux	227446	4 217	20 873,34

Tableau 4 Fréquence du « je » dans les discours de F. Bayrou. Ecart type empirique et comparaison avec les valeurs obtenues grâce aux formules (2)

	Valeurs observées		Valeurs attendues
	absolues	relatives	relatives
Moyenne		0,018541	0,018541
écart-type	144,48	0,000635	0,000332
Borne inférieure	3 933,83	0,017296	0,017892
Borne supérieure	4 500,17	0,019786	0,019190

NB : dans ces tableaux, les deux derniers chiffres ne sont pas significatifs mais sont conservés pour le contrôle des arrondis.

Trois conclusions peuvent être tirées de cette expérience.

En premier lieu, les écarts autour de la moyenne sont toujours nettement supérieurs à ceux que laisse attendre la loi normale alors même que l'on se trouve dans la situation la plus favorable : corpus homogènes émis en même temps par des locuteurs uniques placés dans les mêmes situations d'énonciation. L'alternative est la suivante.

D'une part, examinons l'hypothèse selon laquelle la propension à utiliser « je » est stable, mais que d'autres facteurs perturbateurs s'ajoutent aux fluctuations d'échantillonnage. Par exemple, il est de notoriété publique que, à part F. Bayrou, les principaux candidats ont utilisé plusieurs plumes de l'ombre et il est probable que ces collaborateurs avaient des propensions différentes à mettre du « je » dans la bouche de leur patron. Le degré d'improvisation peut également jouer. En effet, par rapport à l'écrit, le français oral présente une plus forte densité de pronoms personnels et de verbes (Labbé 2003). Le corpus « Présidentielles 2007 » offre toutefois un relatif démenti à cette première hypothèse d'une propension stable à dire « je » propre à chaque individu. En effet, tant pour S. Royal que N. Sarkozy, le test départage nettement leurs prestations d'avant le premier tour d'avec celles de l'entre-deux tours.

	F. Bayrou	Le Pen	Royal (1er tour)	Sarkozy (1er tour)	Royal (2e tour)	Sarkozy (2e tour)
Fréquence	18,54	4,56	19,40	18,72	25,79	26,09
Valeurs attendues :						
Borne inférieure	17,89	4,11	18,69	18,17	24,38	24,90
Borne supérieure	19,19	5,01	20,11	19,27	27,20	27,28
Valeurs observées :						
Borne inférieure	17,30	3,58	17,60	16,56	22,74	20,91
Borne supérieure	19,79	5,53	21,20	20,87	28,85	31,27

Tableau 5. Comparaison des valeurs attendues et des valeurs empiriques (pour mille mots)

Dès lors, il faut examiner la seconde hypothèse selon laquelle la propension à dire « je », chez un même locuteur, dépend surtout des circonstances (la situation d'énonciation) et de l'objet qu'il traite (thème). Dans une situation ou devant un thème qu'il sait peu favorable, l'orateur aura tendance à se faire plus discret. Par exemple, dans son discours sur la politique agricole (14 janvier 2007), S. Royal n'utilise pas une seule fois la première personne alors que sa performance moyenne du premier tour en laisserait attendre 17 dans ce discours…

Deuxièmement, le tableau 5 aboutit au même résultat que celui obtenu avec le test paramétrique bilatéral (tableau 3) qui isole trois groupes d'orateurs. Dans le schéma 2 ci-dessous, les trois zones de fluctuation normale, correspondant à ces trois groupes, sont délimitées par des traits pointillés. Seul J.-M. Le Pen se distingue des 3 autres par un emploi

significativement faible du « je ». Pour les trois autres, les circonstances – et non les personnalités - semblent être la principale explication des différences constatées.

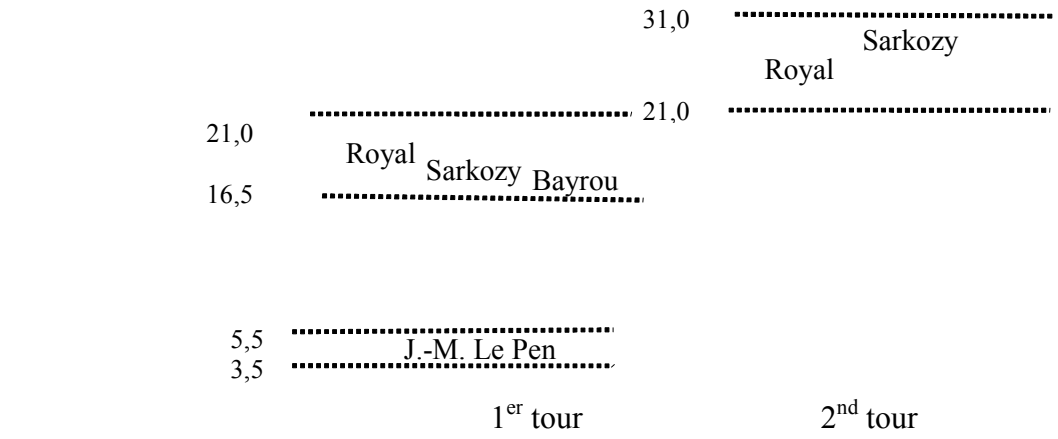


Schéma 2. Trois groupes d'orateurs en fonction de leur fréquence d'emploi du « je »

Troisièmement, l'écart type théorique étant systématiquement inférieur à l'écart type empirique, les deux hypothèses (H_0 et H_1) n'ont pas exactement le même statut. La seconde ($f_{iA} \neq f_{iB}$) doit être considérée avec prudence. A chaque fois que les deux corpus sont divisibles en un nombre suffisant de textes, il faut calculer les écarts types empiriques avant d'accepter H_1.

5. Conclusions

Cette expérience ne permet pas d'écarter l'idée selon laquelle chaque individu présente une certaine propension à utiliser la première personne, propension qui serait une dimension caractéristique de sa personnalité. Elle démontre que les discours publics ne permettent pas l'observation de cette dimension. La densité plus ou moins forte de la première personne dans un discours *public* semble plutôt déterminée par la stratégie de communication et par des effets de contexte. Ainsi, en 2007, la campagne pour le premier tour de l'élection présidentielle a été un peu moins personnalisée que celle du second tour. Autrement dit, le phénomène ne dépend pas de la variable « auteur » mais de l'interaction entre cet auteur, sa stratégie de communication et la situation dans laquelle il est placé lorsqu'il émet son message.

Pour autant, si l'on se souvient que la théorie standard considère que les pronoms sont des « formes vides », des fréquences semblables ne signifient pas que les orateurs disent la même chose quand ils prononcent « je ». Pour le vérifier, il faut étudier les « univers lexicaux » de ces pronoms, c'est-à-dire les vocables qui leur sont associés (Labbé 1998a). La comparaison de ces univers, deux à deux, permet de savoir si – au-delà d'une ressemblance formelle – il est possible d'affirmer que deux auteurs disent plutôt la même chose ou, au contraire, s'ils diffèrent plus ou moins fondamentalement (Labbé 1998b).

La méthode n'est applicable que pour les grands corpus (plusieurs dizaines de milliers de mots) et sur des vocables de fortes fréquences dans ces corpus (au moins une trentaine d'occurrences dans chacun). Il reste donc à imaginer un test qui puisse s'appliquer aux vocables de moyenne et basse fréquences qui sont les plus nombreux et qui forment la « chair » du discours. Ce test devrait présenter les deux caractéristiques suivantes : insensibilité aux différences de longueur entre les corpus comparés et non dépendance par

rapport à la fréquence des mots, ce qui n'est pas le cas du calcul dit des « spécificités » (Labbé & Labbé 1997). De plus, l'hypothèse d'une différence significative entre les fréquences d'emploi dans deux corpus ne devrait être acceptée qu'après avoir été validée par un certain nombre d'observations concordantes.

Remerciements

Les auteurs remercient Cyril Labbé (Université de Grenoble I) - qui a écrit avec D. Labbé les programmes utilisés pour cette communication – ainsi que l'un des relecteurs anonymes dont les observations ont permis la correction de ce texte.

Références

Benveniste E. (1966 & 1970). *Problèmes de linguistique générale*. Paris, Gallimard (rééd. 1980).

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

CISIA-CERESTA (1995). *Aide-mémoire statistique*. Saint-Mandé, CISIA-CERESTA.

Desrosières A. (1988). *La partie pour le tout : comment généraliser ? Cinq contributions à l'histoire de la statistique*. Paris, Economica.

Harris J.-W. & Stocker H. (1998). *Handbook of Mathematics and Computational Science*. New York-Berlin, Springer.

Hubert P. & Labbé D. (1995). La structure du vocabulaire du général de Gaulle. In Bolasco Sergio et al. *IIIe Giornate internazionali di analisi statistica dei dati testuali*. Rome, CISU, II, p 165-176.

Kerbrat-Orecchioni C. (1981). *L'énonciation de la subjectivité dans le langage*. Paris, A. Colin.

Labbé C. & Labbé D. (1994). *Que mesure la spécificité du vocabulaire ?* Grenoble, CERAT. Repris dans : *Lexicometrica*, 3, 2001.

Labbé D. (1981). Moi et l'autre. Le débat Giscard-Mitterrand. *Revue française de science politique*, décembre 1981, p. 951-981.

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

Labbé D. (1990b). *Le vocabulaire de François Mitterrand*. Paris, Presses de la Fondation nationale des sciences politiques.

Labbé D. (1998a). Le « nous » du général de Gaulle. *Quaderni di studi linguistici*, 4/5, p. 331-354.

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

Labbé D. (2003). Coordination et subordination en français oral. *IVe journées de l'ERLA*, Brest 14-15 novembre 2003. In Banks D. (éd). *Coordination/subordination dans le texte de spécialité*. Paris, l'Harmattan, 2006.

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

Maingueneau D. (1994). *L'énonciation en linguistique française*. Paris, Hachette.

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