CIDE.7 : Développer la méthode expérimentale en analyse textuelle

Jean-Marc Leblanc[1]

[1]*CEDITEC, Université de Paris 12 Val de Marne, 94 Créteil Cedex*
leblanc.jeanmarc@free. fr

Résumé :
À partir d'un corpus de discours institutionnels fortement ritualisés (messages de vœux des présidents de la 5e république, 1959-2001), diverses fonctions cooccurrentielles de Lexico3, Weblex, Hyperbase, Alceste sont successivement utilisées et comparées. Elles mettent en évidence, par les calculs, les tris, et les présentations graphiques des faits de cooccurrences différents, susceptibles d'interprétations complémentaires concernant l'ethos discursif de chaque président. La comparaison permet par ailleurs d'insister sur la nécessité d'une expérimentation approfondie dans le traitement des données textuelles avant toute entreprise herméneutique.

MOTS-CLES : cooccurrences, lexicogrammes, discours institutionnel.

Abstract :
The functional comparison of various approaches to cooccurrence in some robust standard textual analyzers makes it possible to develop an experimental method in the courses of textual data processing intended for researchers in social sciences, which is too often absent from the field. Using a corpus of highly ritualized institutional speeches (New year addresses of the presidents of the French 5th Republic, 1958-2003), various functions to measure cooccurrence with Lexico3, Weblex, Hyperbase, Alceste are used in turn and compared. Different cooccurrence data are thus highlighted through figures, sorting, and graphic display, each leading to specific interpretations. The comparison shows the absolute necessity of in-depth experimentation in the analysis of textual data before any hermeneutic venture.

KEYWORDS: cooccurrences, lexicogrammes, political discourse.

1. Méthodologie expérimentale et corpus des vœux présidentiels

Le calcul des cooccurrences, ou mesure probabilisée des attirances entre formes dans un contexte donné, qui repose sur une notion bien décrite dans le domaine, a connu plusieurs types d'implémentations et de nombreuses exploitations ces dernières années. Mesure des cooccurrences binaires [Lafon, 1984], recherche des lexicogrammes, [Tournier, 2003 Heiden, 2003], voisinages [Labbé, 1990], cooccurrences des formes spécifiques [Salem, Martinez, 2003]. Le site Textopol[1] qui offre un accès ergonomique et initie à divers logiciels facilite une démarche expérimentale, cumulative et comparée visant à mettre en évidence les propriétés de chaque méthodologie.

2. Le corpus des vœux présidentiels

Nous examinons ici le statut du *je* présidentiel réduit à sa forme graphique en explorant ses espaces cooccurrentiels dans les messages de vœux aux Français sous la Cinquième République (1959-2001). Nous mobilisons différents outils statistiques pour mettre en lumière les réseaux lexicaux mais aussi sémantiques et thématiques qui gravitent autour d'une marque personnelle structurante dans ce genre institutionnel. Les réseaux de cooccurrence, saisis par ces approches différentes apportent un éclairage sur la construction de l'ethos présidentiel [Amossy, 1999, Maingueneau, 2003] qui se manifeste au sein d'un genre de discours politique fortement codifié.

2.1. Les outils

Deux types d'outils sont mis en œuvre et éprouvés :
- Des Logiciels lexicométriques dits « classiques » (Lexico 3, Hyperbase). Travaillant sur la base d'un tableau lexical, après réorganisation de la séquence textuelle et segmentation en unités minimales (ici la forme graphique), ces outils introduisent la notion de partition, sur laquelle portent des analyses contrastives et des mesures de ventilation du stock lexical dans les sous parties du corpus (bornes chronologiques, locuteurs…). Les fonctions documentaires (concordances, contextes), statistiques (spécificités), analyses multidimensionnelles (Analyse factorielle des correspondances, arborées), constituent les fonctionnalités essentielles de ces outils.
- Des « Cooccurrenceurs » (Weblex, Alceste) où la mesure des voisinages joue un rôle essentiel.

Weblex présente des caractéristiques similaires aux logiciels de type lexicométrique mais intègre des fonctionnalités évoluées de recherche de cooccurrences reposant sur un modèle probabiliste. (Cooccurrences, lexicogrammes

[1] Le site Textopol est réalisé dans le cadre du Céditec (Ea 3119), Université de Paris 12 Val de Marne. (http://textopol.free.fr)

simples et récursifs, associés ou non à une forme pôle dont le principe sera détaillé ultérieurement).

Dans la méthodologie Alceste dont la perspective diffère sensiblement, l'algorithme ne repose pas sur une segmentation pré-établie mais constitue des classes d'énoncés indépendamment des grandes divisions du corpus. Cette démarche inductive, fondée sur une analyse statistique distributionnelle met en évidence les grandes articulations du corpus, ses « mondes lexicaux », en classant les énoncés du texte en fonction de la distribution de leur vocabulaire. Le texte, considéré comme un ensemble d'énoncés est découpé en unités des contexte (plus ou moins la phrase). Le logiciel effectue un repérage des unités lexicales, identifiées au moyen d'un dictionnaire, puis procède à une lemmatisation. Les énoncés sont alors triés en fonction de la présence / absence des formes qui les composent puis classés selon la méthode de classification descendante hiérarchique. On obtient des classes de mots les plus représentatifs de ces énoncés, triés selon leur coefficient d'association à la classe par la méthode du khi2.

2.2. Les paramètres du corpus

Nous nous intéresserons aux interventions produites de décembre 1959 à décembre 2001 soit 43 discours représentant un volume textuel de 41125 occurrences pour 5203 formes. Ces discours, qui forment par ailleurs une série textuelle chronologique[2] seront abordés dans leur dimension synchronique. Bien qu'étant attentifs, lorsqu'il s'agit de retourner au texte, aux emplois individuels, nous mènerons l'expérience en considérant le corpus dans son ensemble.

Locuteur	**Nb Discours**	**Nb occurrences**	**Nb formes**	**Longueur Moyenne**
De Gaulle	10	11498	2407	1150
Pompidou	5	2850	890	570
Giscard	7	6066	1360	866
Mitterrand	14	11991	2521	856
Chirac	7	8720	1799	1245

Tableau 1 : *Principales caractéristiques de la partition locuteur*

Les représentations graphiques produites ci-dessus mettent l'accent sur le relatif déséquilibre des sous parties. Le matériau est beaucoup plus riche pour de Gaulle et Mitterrand que pour les autres locuteurs, les cinq présidents n'ayant pas assumé leur charge sur les mêmes durées.

[2] Nous reprenons ici la définition qu'en donnent [Lebart, Salem 1997] : « Corpus homogènes constitués par des textes produits en des situations d'énonciation similaires, si possible par un même locuteur, individuel ou collectif, et présentant des caractéristiques lexicométriques comparables. » Autre caractéristique essentielle, l'étalement dans le temps permettant de mettre en évidence des variations chronologiques.

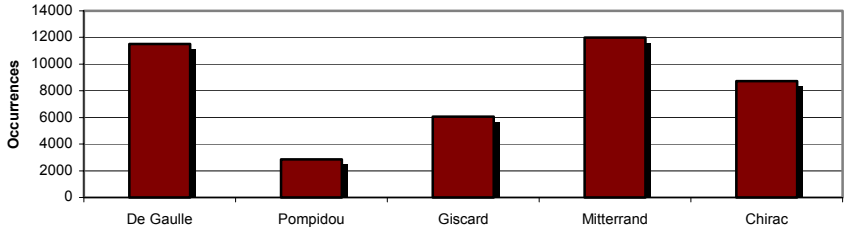

Tableau 2: *Distribution des fréquences*

En outre la longueur des messages est très variable selon les locuteurs. De Gaulle et Chirac consacrent en moyenne 1150 et 1245 occurrences à leurs discours tandis que Mitterrand et Chirac ne leur accordent que 856 et 866 occurrences. Les messages de Pompidou sont les plus brefs avec 570 occurrences. Ces données restent cependant comparables et autorisent des conclusions fiables.

2.3. Les marques énonciatives

Les spécificités3 par partie des pronoms personnels et adjectifs possessifs mettent en lumière des profils énonciatifs très contrastés qui caractérisent ce discours pourtant très ritualisé. (Tableau 3).

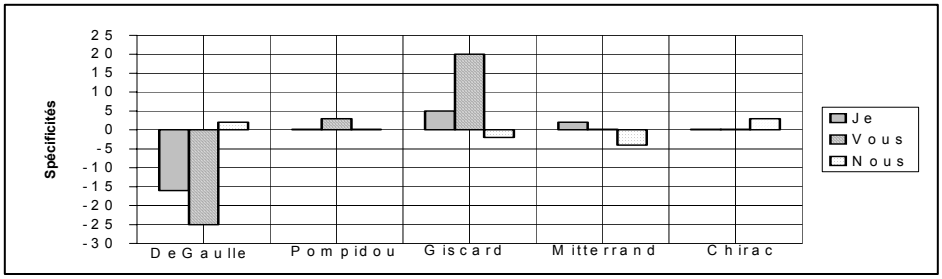

Tableau 3 : *Histogramme des spécificités de je, vous, nous sur la partition locuteur.*

[3] La méthode des spécificités permet de porter un jugement sur la répartition des formes dans les parties d'un corpus. Ce jugement s'exprime en termes de sur-emploi (spécificité positive) et de sous-emploi (spécificité négative). Selon le modèle hypergéométrique, une forme est notée spécifiquement positive si sa fréquence dans une partie est supérieure à la fréquence théorique attendue, et spécifiquement négative si cette fréquence est inférieure au seuil retenu. Ces fréquences probabilisées s'appuient sur la comparaison de quatre données : le nombre des occurrences du corpus, le nombre des occurrences dans la partie, la fréquence de chaque forme dans le corpus, et la fréquence de chaque forme dans la partie. Les indices indiquent le degré de spécificité de chaque forme et représentent la valeur absolue de l'exposant de probabilité. Un exposant de valeur 2 exprime une probabilité de l'ordre du centième, 3 du millième…L'absence d'exposant indique que l'usage ne présente pas de caractéristique remarquable. On dira que la forme est banale pour la partie considérée.

À l'énonciation fortement personnalisée, centrée sur le *je* de Giscard (*je*, +5), multipliant aussi les marques en direction des Français (*vous*, +20), on oppose la prise de distance chez de Gaulle par le rejet des marques de la première personne du singulier (*je*, -16) et de la seconde du pluriel (*vous*, -25), la prise en charge de l'énoncé étant assurée par un *nous* dont le référent est la France. (*nous*, +2).

	De Gaulle	Pompidou	V.G.E	Mitterrand	Chirac
nous	+E02	-	-E02	-E04	+E03
je	-E16	-	+E05	+E02	-
j'	-E06	-	-	+E05	-
vous	-E25	+E03	+E20	-	-
on	-E02	-E03	-	+E10	-E03
notre	+E03	-	-E02	-E05	+E02
nos	+E02	-E03	-E04	-	+E03
mes	-E08	-	+E03	+E03	-
votre	-E09	+E03	+E11	-E02	-E02
vos	-E06	-	+E11	-E03	-
moi		-	-	+E02	-E02
me	-E02	-	+E02	+E02	-E03
m'	-E03	-	-	+E02	-

Tableau 4 : *Spécificités des principaux pronoms personnels et adjectifs possessifs.*

3. Approches des cooccurrences de JE

3.1. Topographie textuelle et cooccurrents spécifiques (Lexico 3)

L'outil carte des sections établit la distribution de la forme personnelle dans la linéarité du texte, délimité en paragraphes.

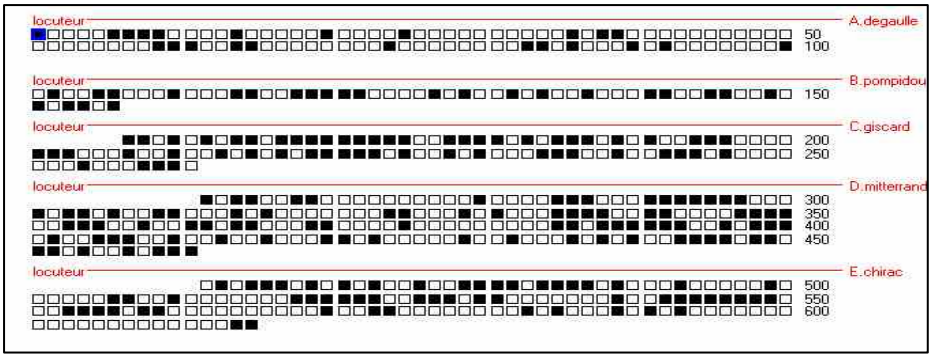

Tableau 5 : *Carte des sections (paragraphe) de la forme JE sur la partition locuteur*

Dans cette « topographie textuelle », [Lamalle, Salem, 2002] chaque rectangle du tableau 5 matérialise une section de texte, les unités colorées indiquant les paragraphes attestant au moins une fois la forme recherchée. Au moyen de cette cartographie, on appréhende des usages, des positionnements énonciatifs en termes de fréquences mais aussi de rythme, de cadence. Entre de Gaulle et Giscard par exemple, on note deux configurations : de longues successions de paragraphes contenant *je* chez Giscard, de brèves interruption, ou rares îlots chez De Gaulle.

Le calcul des cooccurrents spécifiques met en évidence à partir des sections délimitées par cette cartographie les formes spécifiques des paragraphes attestant le *je*. La liste produite porte sur l'ensemble du corpus et ne présente que les formes dont la valeur absolue de l'indice de spécificité est supérieure à 2. Les seuils sont de 5%, la fréquence minimale des formes considérées est de deux occurrences. Ce calcul ne diffère pas du modèle de spécificité évoqué précédemment si ce n'est que les parties sur lesquelles porte la comparaison ne sont plus matérialisées par une partition en locuteurs mais constituées selon la présence ou l'absence du pronom personnel. Le diagnostic de spécificité est alors établi sur la base d'une partition binaire : l'ensemble des sections dans lesquelles la forme analysée est présente par rapport à l'ensemble du corpus. Les coefficients indiqués au tableau qui suit correspondent donc à des indices de spécificité. Une spécificité positive signifie qu'une forme considérée a tendance à apparaître de façon plus importante que le modèle théorique ne le laissait prévoir dans les contextes du pôle analysé, par rapport aux autres sections du corpus, une spécificité négative indiquera un rejet ou un sous-emploi. En d'autres termes, ce calcul appliqué aux sections permet de repérer les fréquences remarquables au voisinage de la forme pôle. Le calcul des cooccurrents spécifiques ne fournit pas d'information quant aux cooccurrents droits ou gauches du pôle choisi et ne comporte pas d'autre notion de distance entre les termes co-présents que celle fixée par la longueur des sections.

Forme	Frq. Tot.	Fréquence	Coeff.	Forme	Frq. Tot.	Fréquence	Coeff.
je	344	344	51	j	88	51	3
souhaite	65	64	23	grandeur	7	7	3
vous	326	227	23	soir	42	28	3
voeux	80	62	10	adresser	7	7	3
mes	102	75	10	ma	20	15	3
sais	19	19	8	nom	30	21	3
voudrais	19	19	8	mon	29	19	3
pense	22	21	7	amis	11	10	3
suis	26	24	7	fraternité	19	14	3
forme	17	16	5	seuls	14	12	3
vive	60	42	5	france	302	150	3
heureuse	22	19	5	vivent	10	9	3
bonne	76	51	5	m	23	17	3
veux	11	11	5	vois	6	6	3
crois	11	11	5	ministre	6	6	3
que	677	336	5	famille	25	17	3
dire	48	35	5	très	27	18	3
dis	12	12	5	fais	6	6	3
compatriotes	62	43	5	françaises	41	27	3
ai	41	31	5	doivent	14	1	-3
année	205	110	4	quel	12	0	-3
vos	39	27	4	la	1397	546	-3
chers	55	37	4	algérie	21	3	-3
me	22	17	4	économique	46	10	-3
français	142	80	4	qu	313	108	-3
votre	59	38	4	peut	50	11	-3
chacune	25	20	4	europe	99	28	-3
mer	19	14	3	part	32	3	-5
				nous	655	217	-7

Tableau 6 *: Cooccurrents spécifiques de « je » sur la totalité du corpus « vœux »*

Les spécificités positives montrent la forte proportion de verbes, gravitant autour du référent du locuteur. (Tableau 6). Verbes marquant la volition (*souhaite, voudrais, forme* [le vœux], *veux),* le jugement (*pense crois),* factifs (*fais* +3), verbes d'état et auxiliaires (*Suis*, *ai*), énonciatifs (*dis*), verbes marquant la connaissance (*sais*, *vois*), quelques infinitifs (*Dire, adresser*), constituent l'essentiel du système verbal restitué par la recherche des cooccurrents spécifiques. On note aussi de façon plus inattendue la présence d'un verbe à la troisième personne du pluriel : *vivent (+3)*, dont on trouve la réalisation dans de fréquentes adresses aux Français *qui vivent à l'étranger* (De Gaulle, 1967), *qui vivent dans la solitude* (V.G.E, 1978)*, qui vivent dans la peine* (Mitterrand, 1986)*, qui vivent dans la difficulté quotidienne* (Mitterrand, 1988).

Ces messages sont donc particulièrement marqués par des verbes de « circonstance », (*souhaiter*, *adresser former*), par des volitifs et des verbes exprimant la connaissance. Cependant, cette interprétation sémantique a priori doit être corrigée par l'examen des contextes.

Une analyse approfondie indique que la forme *voudrais* est intimement liée au référent de l'interlocuteur, sur employée chez les locuteurs qui précisément multiplient les marques énonciatives en direction des Français. La valeur n'est donc que rarement purement volitive, les emplois étant essentiellement métadiscursifs, modalisateurs, intervenant dans des annonces de plan où bien souvent le locuteur s'adresse à une certaine catégorie de Français (*Je voudrais d'abord exprimer ma sympathie à toutes celles et à tous ceux qui vivent ces derniers jours de 1999 dans l'épreuve.* [Chirac, 1999]). Giscard et Pompidou qui entretiennent un lien plus étroit avec les Français emploient cette forme dans une modalité directive qui intensifie la relation (*Je voudrais que vous sentiez, que vous compreniez*…). Les contextes de *veux* montrent également une tendance vers des emplois métadiscursifs ou explicatifs (*Je veux dire*), même si la volition apparaît parfois chez Chirac et Mitterrand dans une faible mesure.

Quant aux verbes exprimant le jugement, on remarque que *penser* intervient essentiellement dans des énoncés énumératifs (*Je pense aux artisans, je pense aux agriculteurs, je pense à certaines petites entreprises*) mais bien souvent affectifs et empathiques, liés à l'évènementiel (*Et je pense aussi à nos compatriotes de Toulouse*…[Chirac, 2001]) ou plus généralement destinés à adresser un geste en direction des Français les plus démunis, évocation qui devient systématique à partir de Pompidou. (*Je pense spécialement à ceux de nos aînés qui vont franchir seuls le cap du nouvel an.* [Chirac, 2000], *Je pense à celles et à ceux d'entre vous qui connaissent le deuil, les chagrins, le poids de la maladie et de la solitude, qui souffrent du chômage*. [Mitterrand, 1981]).

Parmi les verbes exprimant la conscience et la connaissance, l'examen des contextes montre que la forme *sais* entre essentiellement dans des modalités allocutives. Les emplois sont avant tout des renforçateurs d'empathie, plus particulièrement chez Chirac, parfois constitutifs d'un procédé argumentatif. Cette marque d'empathie introduit dans de nombreux cas chez Chirac une relance incitative et mobilisatrice, que l'on peut synthétiser dans le tableau 7.

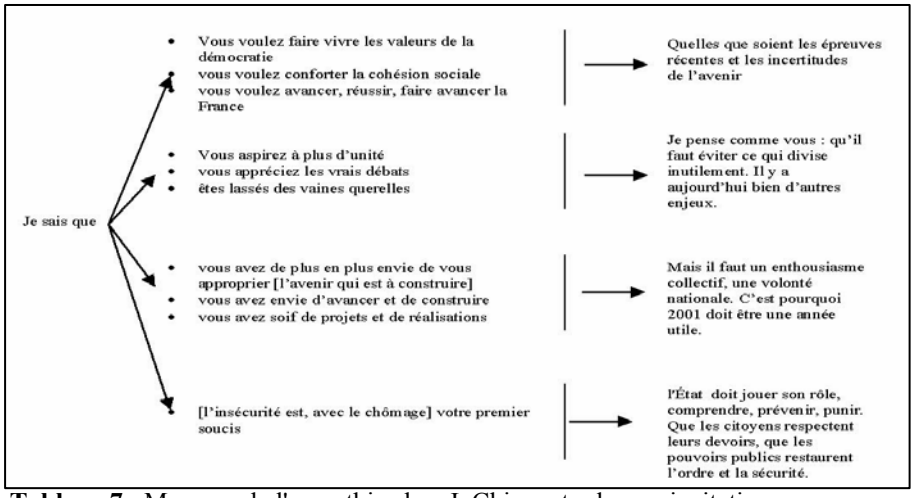

Tableau 7 : Marques de l'empathie chez J. Chirac et relances incitatives.

Le champ lexical du rituel, verbes, substantifs, formules d'adresse est largement représenté parmi les formes qui gravitent autour du *je* présidentiel : *compatriotes*, *chers*, *mer* (*compatriotes d'outre mer*), *Français*, *Françaises*, *vive*, *année*, *bonne*, *heureuse*, *vœux*.
Parmi les spécificités négatives - assez minoritaires et dont les indices se révèlent moins élevés que pour les sur-emplois - nous retiendrons le diagnostic porté sur la première personne du pluriel. (*nous*, -7). Ainsi, le schéma *je/vous*, dont on relève les traces parmi les pronoms personnels et adjectifs possessifs se construit par le rejet de la première personne du pluriel.

3.2. La notion d'environnement thématique (Hyperbase)

C'est à partir de la recherche des contextes qu'il nous est offert d'étudier l'environnement thématique sous Hyperbase. Cette recherche procède d'un calcul de spécificité particulier. Plutôt que de porter un jugement sur la fréquence d'apparition d'une forme dans une sous-partie du corpus par rapport aux autres et à l'ensemble, on calcule les spécificités des mots qui se trouvent au voisinage d'une forme pôle. En d'autres termes - nous empruntons cette formulation à Etienne Brunet - : « Ce programme de repérage thématique fait le décompte de tous les mots situés dans le même paragraphe que le ou les mots-pôles et mesure leur spécificité, c'est-à-dire la plus ou moins grand attirance que le mot-pôle exerce sur eux. ». Le choix des partitions n'est pas sans incidence sur la recherche thématique. Aussi adopterons-nous la partition locuteur qui présente l'avantage de ne pas morceler le corpus par trop ainsi qu'un découpage en paragraphes qui constitue un échantillon contextuel suffisamment large. Les formes relevées comme constitutives de « l'environnement thématique » de *je* sont classées selon leur rang de significativité, c'est-à-dire dans l'ordre décroissant de l'écart réduit, qui s'étend ici de la valeur 25,79 à 2. L'écart réduit peut être considéré comme une approximation – satisfaisante pour des corpus de grande ampleur – du modèle hypergéométrique utilisé dans le calcul des

spécificités. La loi normale en usage dans les versions antérieures d'Hyperbase, appliquée aux spécificités et chaque fois que la distribution d'une forme devait être pondérée pour tenir compte des étendues inégales des textes comparés. Toute fréquence était alors traduite en écart réduit. La puissance de calcul des ordinateurs actuels autorise désormais l'application de la loi hypergéométrique (utilisée également sous Lexico et Weblex), plus appropriée car s'appliquant à des données discrètes (la loi normale traitant des valeurs continues) et d'une plus grand exactitude sur des corpus de faible étendue. Les résultats sont toutefois convertis afin de conserver l'échelle de l'écart réduit et la représentation adoptée dans les précédentes versions d'Hyperbase. L'écart réduit s'interprète en termes de déficits ou d'excédents, sa valeur est donc positive ou négative. Une valeur avoisinant 2 ou –2 est considérée comme négligeable. Le calcul de l'environnement thématique ne tient pas compte des déficits. Ainsi, contrairement aux cooccurrents spécifiques de Lexico, aucun diagnostic n'est porté sur les sous-emplois. La seconde colonne du tableau 8 présente l'effectif de la forme considérée sur l'ensemble du corpus, la dernière colonne indique le nombre d'occurrences dans la sous partie, constituée à partir des contextes de *je*. On note que la fréquence de l'extrait est parfois supérieure à l'effectif total. Le cas le plus flagrant est celui de la forme que nous avons choisie comme pôle de référence : 344 occurrences au total pour 347 formes dans l'extrait. Cet artefact s'explique par la comptabilisation des graphies qui apparaissent plusieurs fois au sein d'un même paragraphe. Il conviendrait donc de revoir la segmentation du corpus afin de neutraliser ce comptage. Toutefois de multiples vérifications, ainsi que l'utilisation comparée des plusieurs outils nous autorise à considérer ces résultats comme fiables. Enfin, pour faciliter la lecture, et la comparaison avec les listes produites par Lexico nous avons porté les indices de spécificité relevés par cet outil. (Colonne *Sp*). L'essentiel des faits que nous avions abordés par l'entrée des cooccurrents spécifiques est restitué. Les premiers rangs sont identiques à ceux observés au moyen de la méthode de cooccurrents spécifiques, ce qui conforte nos premières conclusions sur le rituel et la thématique des vœux. Parmi les verbes : *souhaite* dont on a déjà examiné les contextes et établi qu'il était essentiellement employé dans la formulation de vœux, de même *forme*, *adresser*, mais aussi *limiter* (*je ne voudrais pas me limiter à vous présenter mes vœux…).* Les usages de la première personne sont essentiellement conditionnés par le genre discursif y compris chez les locuteurs dont l'énonciation est peu personnalisée. Le *Je* est à la fois dialogique et familier, rituel et circonstanciel. On s'interrogera sur la procédure d'élagage appliquée sous Hyperbase - dont on maîtrise peu les paramètres mais dont la présentation synthétique est propre à faire émerger les faits saillants du corpus mais aussi plus anecdotiques, qui apparaissaient sous Lexico parmi de nombreuses autres formes, dotées généralement d'indices très faibles de spécificité. On soulignera cependant l'absence remarquable des formes *chers* (indice positif de 4 sous Lexico), *compatriotes* (+5), *France* (+3), *Vive* (+5) *bonne* (+5) qui entrent dans la réalisation de formules d'adresse ou qui appartiennent à la dimension rituelle du discours (*Vive la France, bonne année…*). Ces outils fournissent des représentations des faits de cooccurrences, propres à orienter l'analyse dans des directions parfois différentes, d'où la nécessité de les interroger systématiquement.

Ecart	Corpus	Extrait	Mot	Sp.	Ecart	Corpus	Extrait	Mot	Sp.
25.79	344	347	JE	51	2.74	4	4	LÉGISLATIVES	2
11.80	323	213	VOUS	23	2.74	4	4	AUGMENTÉ	2
11.06	65	65	SOUHAITE	23	2.71	29	17	MON	3
7.23	79	58	VŒUX	10	2.65	11	8	RAISONS	2
6.46	19	20	VOUDRAIS	8	2.61	19	12	FRATERNITÉ	3
6.17	26	24	SUIS	7	2.60	127	58	AUX	1
5.99	22	21	PENSE	7	2.60	122	56	AVEC	2
5.98	19	19	SAIS	8	2.60	17	11	ADRESSE	2
5.56	48	35	DIRE	5	2.50	6	5	UNS	2
5.50	41	31	AI	5	2.50	6	5	MINISTRE	3
5.44	674	301	QUE	5	2.50	6	5	EXPRIME	1
5.15	17	16	FORME	5	2.50	6	5	DÉPARTEMENTS	2
4.75	12	12	DIS	5	2.50	6	5	AVAIS	2
4.55	11	11	VEUX	5	2.40	178	77	PAS	2
4.55	11	11	CROIS	5	2.40	33	18	GOUVERNEMENT	1
4.34	25	19	CHACUNE	4	2.40	8	6	COMPRENDRE	2
4.25	59	36	VOTRE	4	2.35	202	86	ANNÉE	4
4.20	22	17	ME	4	2.35	18	11	AVEZ	2
4.17	101	55	MES	10	2.33	14	9	SEULS	3
4.03	42	27	SOIR	3	2.33	14	9	CHOSES	2
3.68	30	20	NOM	3	2.22	41	21	FRANÇAISES	3
3.63	7	7	ADRESSER	3	2.17	90	41	AUSSI	2
3.59	3	4	LIMITER	2	2.16	21	12	POSSIBLE	1
3.55	101	52	CEUX	2	2.14	30	16	ABORD	2
3.52	39	24	VOS	4	2.13	37	19	AN	2
3.49	27	18	CŒUR	2	2.13	5	4	VRAIMENT	1
3.36	6	6	VOIS	3	2.13	5	4	VENU	1
3.36	6	6	FAIS	3	2.13	5	4	STRASBOURG	1
3.30	22	15	BONHEUR	2	2.13	5	4	RÉPONDRE	1
3.28	11	9	AMIS	3	2.13	5	4	PROGRESSER	1
3.08	25	16	FAMILLE	3	2.13	5	4	OFFRE	1
3.08	23	15	M'	3	2.13	5	4	KOWEÏT	2
3.07	5	5	MESSAGE	2	2.13	5	4	GARANT	1
3.07	5	5	MARS	2	2.13	5	4	FRAPPÉ	1
3.07	5	5	FIER	2	2.13	5	4	FIDÈLE	1
3.07	5	5	FAIBLES	1	2.13	5	4	ADRESSENT	1
3.07	5	5	DISAIS	2	2.12	19	11	MER	3
3.05	83	42	J'	3	2.09	17	10	ÉTRANGER	2
3.01	10	8	VIVENT	3	2.04	7	5	PRÉSIDENT	1
2.88	26	16	TRÈS	3	2.04	7	5	PARLER	1
2.85	22	14	HEUREUSE	5	2.04	7	5	MAJORITÉ	2
2.85	20	13	MA	3	2.04	7	5	FRATERNELLE	1
2.84	7	6	GRANDEUR	3	2.02	11	7	SOUFFRENT	1
2.84	7	6	DEMANDE	2	2.02	11	7	PROFESSIONNEL	2
2.77	142	65	FRANÇAIS	4	2.02	9	6	SOLITUDE	1
2.74	4	4	VAIS	2	2.02	9	6	RÉUSSIR	1
2.74	4	4	TIRE	2	2.02	9	6	DÉPIT	1
2.74	4	4	TIENS	2	2.02	9	6	DÉBUT	2
2.74	4	4	REÇU	2	2.00	24	13	TROP	2
2.74	4	4	QUICONQUE	2	2.00	24	13	BESOIN	1

Tableau 8 : *Environnement thématique de la forme « je » - Hyperbase. Partition locuteur, contexte paragraphe*

3.3. Lexicogrammes simples, lexicogrammes récursifs (Weblex)

Le calcul de cooccurrences implémenté dans Weblex repose sur le modèle développé par Pierre Lafon [Lafon, 1984]. On se reportera à [Heiden 2004] et [Tournier, 2003] pour des approfondissements méthodologiques.

Cette méthode permet de porter sur le lexique présidentiel deux éclairages complémentaires.

- Le lexicogramme simple, associé à une forme affiche les formes « les plus cooccurentes » d'une forme pôle. « Le lexicogramme d'un mot s'interprète comme une synthèse des cooccurrents gauches et droits d'un mot, à l'intérieur de toutes les phrases où il apparaît ». [Heiden, 2004]. Le tableau 9 fournit ainsi une dimension supplémentaire qui ne pouvait être appréhendée avec les outils précédents. Il

présente les principaux cooccurrents gauches et droits de la forme pivot *je* dans l'ensemble du corpus au seuils Fréquence et co-fréquence minimale de 3 occurrences, probabilité de 5% (soit 5.0^{e}-2), distance moyenne 1000 occurrences.
Pour chaque cooccurrent la Fréquence totale de la forme dans le corpus (**F**), sa cofréquence avec la forme pôle, c'est-à-dire le nombre de rencontres attestées (**CF**), le diagnostic de probabilité de la rencontre (**P**) et la distance moyenne ($\mathbf{d_m}$) fournissent les indications quantitatives et statistiques de ces rencontres.
Nous constatons au premier regard un déséquilibre entre le cooccurents gauches et les cooccurents droits qui dominent très largement. Ce phénomène ne surprendra pas en raison de la nature de la forme pivot qui implique par essence de plus fréquentes relations sur sa droite que sur sa gauche. Les cooccurrents gauches restituent la dimension rituelle du discours. Cinq d'entre eux sont constitutifs de formules d'adresse (*mes chers compatriotes de métropole et d'outre mer*. La position du substantif « vœux » ne surprendra pas d'avantage : (*les vœux que je forme*, *que je vous adresse*…) ni la présence du substantif *cœur* dont les contextes attestent la réalisation (*C'est de tout cœur que je…*).
Les verbes, généralement post-posés au pronom personnel figurent tout naturellement à droite du pivot. On y retrouve ceux que nous avions recensés au moyen des autres outils mais la hiérarchie est quelque peu différente. *Souhaite* se trouve au premier rang (tri par probabilités) comme c'était le cas avec les autres analyses. Sa distance moyenne avec la forme pôle indique qu'il s'agit sans doute souvent de collocations ou qu'un terme peut s'interposer, probablement une marque de l'interlocuteur. (*je vous souhaite*). Les onze premières formes à droite sont des verbes dont la morphologie indique qu'ils sont fléchis à la première personne du singulier, et qui sont peu distants de la forme pivot. On note que le temps est très majoritairement le présent de l'indicatif. Après la catégorie verbale viennent des adjectifs et des substantifs qui ancrent le discours dans le présent, ou sont issus de la thématique des vœux : (*heureuse*, *vœux*, *soir*, *bonne*, *année*, *adresse*)…Peu de termes politiques émergent de ces lexicogrammes, ainsi que nous l'avions déjà observé au moyen des autres outils cooccurentiels. Ceci confirme nos premières observations : le *je* est essentiellement mobilisé par le genre discursif. Une dernière remarque concerne la distance moyenne des cooccurrents, plus importante sur la gauche de la forme pivot que sur sa droite où l'on observe quelques collocations. Qu'en est il de la présence forte de l'interlocuteur que nous avions cru déceler dans les analyses précédentes ? Parmi les verbes, aucun ne semble être conjugué à la deuxième personne du pluriel, ainsi que nous l'avions déjà noté. On ne s'étonnera pas de l'absence des marques personnelles renvoyant à l'interlocuteur dont on a noté la forte spécificité sous Lexico, confirmée par Hyperbase. Nous avons en effet choisi de conserver l'élagage des formes outils du vocabulaire afin de porter un regard différent sur le corpus, en ne considérant que les seuls mots pleins.
Une expérience menée sur ces mêmes lexicogrammes sans suppression des formes outils a cependant confirmé l'attirance importante des marques de la première du singulier et de la seconde du pluriel.

je (346)

cooccurrents gauches				
	f	**cf**	**p**	**d_m**
chers	55	21	4e-06	7.3
métropole	18	10	3e-05	11.1
compatriotes	62	20	1e-04	7.2
françaises	41	15	2e-04	10.7
outre-mer	14	7	1e-03	8.6
voeux	81	21	2e-03	4.9
coeur	27	9	8e-03	4.3
vois	6	3	4e-02	17.0

cooccurrents droits				
	f	**cf**	**p**	**d_m**
souhaite	65	64	6e-57	0.2
pense	22	21	9e-18	0.0
voudrais	19	19	3e-17	0.1
sais	19	18	4e-15	0.6
forme	19	16	8e-12	0.0
crois	11	11	3e-10	0.1
veux	11	11	3e-10	0.1
dis	12	11	3e-09	0.7
dire	39	19	1e-07	3.3
adresser	7	7	9e-07	6.9
fais	6	6	7e-06	0.8
heureuse	22	12	7e-06	13.5
voeux	81	26	9e-06	7.3
soir	42	17	1e-05	6.8
bonne	75	24	2e-05	8.5
demande	7	6	4e-05	2.2
disais	5	5	5e-05	0.8
vois	6	5	3e-04	0.0
vais	4	4	4e-04	0.0
année	202	42	7e-04	10.1
adresse	17	8	9e-04	1.0
fier	5	4	2e-03	6.5
faibles	5	4	2e-03	20.8
salue	3	3	3e-03	0.7
dirai	3	3	3e-03	1.0
répète	3	3	3e-03	1.3
souviens	3	3	3e-03	1.3
assure	3	3	3e-03	14.0
sûr	12	6	3e-03	1.0
rendre	9	5	4e-03	7.6
promis	6	4	4e-03	2.0
espère	6	4	4e-03	5.8
propose	6	4	4e-03	6.5
dit	14	6	8e-03	3.8
fraternelle	7	4	9e-03	17.8
demandé	4	3	9e-03	2.0
satisfaction	4	3	9e-03	12.0
tire	4	3	9e-03	19.3
nom	30	9	2e-02	6.4
vivre	30	9	2e-02	13.1
parle	8	4	2e-02	2.2
constate	5	3	2e-02	0.0
ardents	5	3	2e-02	3.0

Tableau 9 : *Lexicogramme de la forme je dans le corpus vœux. Seuils : f 3, cf 3, p 5.0E-2, dm 1000.0*

- Le lexicogramme récursif associé à une forme révèle des réseaux de cooccurences sur le principe d'un enchaînement de cooccurrents de cooccurents. En partant de la source du lexicogramme (*je*) il en établit les principales attirances. Partant de ces nouvelles formes, d'autres cooccurents sont mis en évidence selon le principe que chaque nouveau cooccurrent est à son tour pris comme source. La représentation qui en résulte synthétise l'ensemble des connexions qui prennent naissance autour de la

forme pôle. Le lexicogramme récursif de la première personne du singulier, a été établi avec élagage des formes outils aux seuils précédemment adoptés.

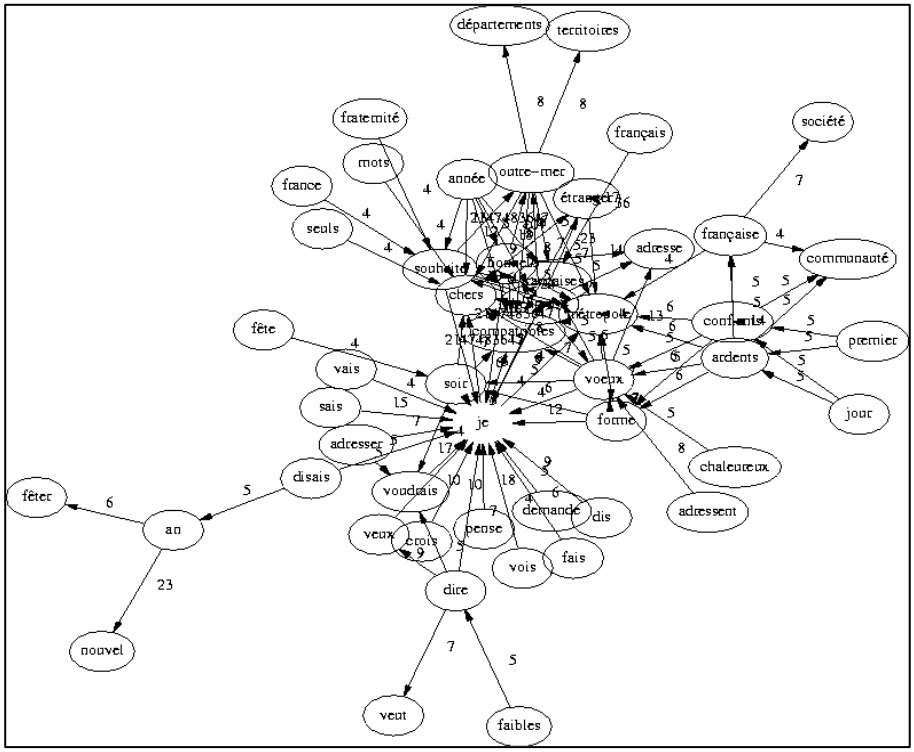

Tableau 10 : *Lexicogramme récursif de la forme je sur le corpus voeux2. Élagage des formes outils, numéraires et hapax. Seuils : f 3, cf 3, p 5.0E-2, d_m 1000.0*

La thématique des vœux ressort nettement de cette représentation. La connexion de verbes gravitant autour du *je* est manifeste : *forme*, *adresser*, *sais*, *demande*, *dis*…, verbes qui nous sont désormais familiers. Dans cette configuration, l'ordre des termes n'est pas restitué bien qu'il soit possible de le faire. Cette perspective spatiale permet de saisir des faits qui ne seraient pas nécessairement apparus aussi clairement sur des listes hiérarchisées. Ainsi les formules d'adresse sont-elles matérialisées par les nœuds *départements*, *territoires*, *outre-mer*. Les qualifiants de *vœux*, *ardents* et *confiants* que le général de Gaulle *adresse* à la France sont clairement identifiés, de même que les emplois métadiscursifs (*je veux dire*, *je voudrais dire*, *je voudrais adresser*). On se gardera cependant d'interpréter trop hâtivement ces connexions. Suivons par exemple la « branche » qui relie le pronom personnel à *disais*. Le verbe est lui-même relié à *an* qui forme un nœud à partir duquel deux nouveaux arcs partent vers *fêter* et *nouvel*. Nul doute que le substantif *an* entre en cooccurrence avec le verbe et l'adjectif. Des réalisation comme « *fêter le nouvel an* » ne seront pas difficiles à attester. Pas de doute non plus sur la relation *je-disais* vérifiée en contexte. Il en va tout autrement de réalisations poly-cooccurrentes qui attesteraient

je, *disais*, *fêter*, *nouvel an* dans un même contexte. Ceci tient au caractère de récursivité des lexicogrammes. Le lexicogramme récursif établit bien les cooccurents de *je* parmi lesquels le verbe *disais.* Cependant considérant ce verbe à son tour comme source une nouvelle cooccurrence sera établie comme ici *an*. *Disais* est bien cooccurrent de *an*, et donc de *je*. Mais *an* ne renvoie pas nécessairement dans les mêmes contextes à *fêter* et *nouvel*. Il s'agit donc de ne pas considérer le chemin que nous venons de retracer comme un « squelette de phrase » [Martinez, 2003]. Illustrons notre propos à l'aide de quelques contextes.

Il y a 3 occurrences de *"an"[]*"disais"%c* dans le corpus voeux

Pompidou, 1971 une fête . Ils oublient la présence de l' hiver pour ne pressentir que le prochain printemps . Ils veulent croire que vieillir est un moyen de marcher vers le meilleur . Cela s' appelle l' espérance . Avons - nous , en tant que Français , des raisons d' espérer ? Eh bien , oui , n' en déplaise à tous les spécialistes de la triste figure . Il y a un **an , jour pour je vous disais** : " Nous ne sommes pas les plus forts , mais nous comptons et nous sommes respectés . " L' année 1971 n' en a-t - elle pas apporté quelques preuves ? Les visites amicales que nous ont faites tant de chefs d' État et de gouvernement étrangers , une délégation chinoise , le premier responsable soviétique , l' entrevue que j' ai eue en terre européen

Pompidou, 1971 l' élargissement de la Communauté européenne et la crise monétaire internationale . A Berlin , aux Nations unies , son action a été visible et utile . Il n' y a pas lieu d' en tirer vanité . Mais , pourquoi le dissimuler , notre pays , indépendant , pacifique et sûr de lui , n' a pas déchu du rang où l' avait placé le général de Gaulle . Il y a un **an , je vous disais** encore : " Nous ne sommes pas les plus riches , mais nous sommes parmi les plus heureux . Il suffit de regarder autour de nous . " Or , aujourd'hui , il suffit d' écouter la voix des commentateurs étrangers , qu' ils soient Anglais , Américains ou Russes , pour apprendre que la situation de la France est appréciée par tous et enviée par beaucoup

Pompidou, 1973 onde dans ses biens , dans sa situation , dans ses libertés . Je suis convaincu que vous en avez conscience et c' est pourquoi c' est en pleine confiance et de tout coeur que je vous dis ce soir : Bonne année ! Que 1973 apporte à vous tous un peu plus de bonheur . De mon mieux , soyez - en certains , j' y aiderai . Françaises , Français , il y a un **an , en vous offrant mes voeux pour l' année 1973 , je vous disais** que ce serait une année d' expansion exceptionnelle et de grands progrès dans divers domaines . Eh bien , c' est ce qui s' est passé . Les chiffres le prouvent et les observateurs sérieux , les plus rigoureux , le reconnaissent . Et pourtant , il faut admettre que l' année se termine dans une atmosphère moins sereine et que les perspectives sont p

3.4. Distributions statistiques et distributions linguistiques (Alceste)

L'expérimentation menée ici repose sur une utilisation particulière d'Alceste. Il ne s'agit pas de faire émerger les structures saillantes du corpus, d'en identifier les classes d'énoncés ou « mondes lexicaux » [Reinert, 1993, 1998] comme le fait la procédure par défaut. On utilise ici le tri croisé sur une forme pour mettre en évidence les cooccurrents du *je*, sur l'ensemble du corpus, afin de recouper cette analyse aux examens précédemment réalisés.

L'analyse en tri croisé, qui peut porter sur une forme ou sur une variable consiste à croiser forme ou variable avec l'ensemble du corpus ce qui aura pour effet de scinder le corpus en deux parties, celle où la forme est attestée et l'autre où elle n'apparaît pas. Contrairement aux analyses usuelles qui reposent sur le principe de la classification descendante, 100% des U.C.E (unités de contexte élémentaire) sont classées. Ici, l'ensemble des U.C.E est pris en compte, sur la base d'une classification ascendante.

Nous produisons dans le tableau 11 qui suivent les formes les plus caractéristiques des deux classes obtenues, ordonnées selon le Khi2 décroissant[4].

Classe 1: 255 U.C.E (27%)

Forme réduite	Khi2
je	969
vous	148,97
souhait<	148,06
suis	60
dire.	45,77
mes	44,38
voeu+	38,95
dire+	38,14
forme+	37,03
me	27,54
annee+	27,44
vouloir.	27,02
heur+eux	23,87
compatriote+	23,61
francais+	23,31
adress+er	22,45
cher+	19,82
ai	18,98
nom+	17,59
outre	15,11
qu+	14,12
enjeu+	14,07
adresse+	13,15
*loc_giscard	12,76
propos+er	12,36

Forme réduite	Khi2
votre	12,07
soir+	12,07
bonne+	11,34
redire.	11,25
tiens	11,25
present+er	11,25
echang+er	11,25
metropole+	10,98
savoir.	10,9
m	10,52
soiree+	10,12
avais	10,12
coeur+	9,97
promettre.	9,86
parl+er	9,43
année_1974	9,34
mon	9,34
pas	9,32
ce	9,27
ma	8,67
mer+	8,67
recu+	8,43
aim+er	8,43
demand+er	8,09
rappel+er	7,63

Classe 2 : 714 U.C.E (73%)

Forme réduite	Khi2
*loc_dg	52,38
nous	14,84
econom+	10,6
année_1963	10,31
devoir.	8,67
année_1965	7,29
année_1961	7,27
année_1968	6,24
année_1960	6,16
moyen+	5,56
developpement+	5,5
*année_1966	5,12
arme+	5,07
cooperati+f	5,07
techn+	5,07
peuple+	5,06
face+	4,71
organisat+ion	4,71
*année_2001	4,57
avons	4,52
elle	4,46
jusqu+	4,34
plan+	4,34
conflit+	4,34
inflation<	4,34

Forme réduite	Khi2
trouv+er	4,1
ailleurs	3,97
leur	3,97
guerre+	3,93
moderne+	3,73
solid+e	3,61
rapport+	3,61
acquerir	3,61
soi	3,61
quel	3,61
troubl+er	3,61
scientifi<	3,61
ensemble+	3,55
*année_2000	3,55
*année_1962	3,44
mondia+	3,44
moment+	3,38
cas	3,24
dehors	3,24
telle	3,24
multipli+er	3,24
monetaire+	3,24
but+	3,24
route+	3,24
aid+er	3,01

Tableau 11 : *Croisement de la forme je avec le corpus. Les 50 premières formes caractéristiques*

Ces classes nous donnent confirmation de certains faits observés précédemment.
- Sur l'aspect pronominal, la dimension dialogique du discours est confortée, ainsi que le rejet de la première personne du pluriel: *vous* figure au second rang des formes significatives de la classe 1, les marques de la première du pluriel à un rang identique dans la classe opposée.
- Sur l'aspect énonciatif: Parmi les énoncés qui s'opposent au *Je* la variable loc_dg est la première forme significative de la deuxième classe. C'est-à-dire que la plupart des énoncés du locuteur de Gaulle se trouve classé dans la catégorie la plus éloignée du *je*.
- Sur la chronologie : Les années 63, 65, 61, 68, 60, 66, 2001 et 2000 sont dotées d'indices de Khi2 relativement importants et sont donc fortement constitutives de la classe 2. Ainsi cette classification binaire et somme toute assez brutale porte un éclairage chronologique sur les emplois pronominaux puisque, recoupant cette information par les spécificités par partie, il ressort que la forme *je* est sous employée dans les messages correspondants par rapport à l'ensemble du corpus. Ainsi résumons nous notre démarche par ce va et vient entre les différents outils, les résultats de l'un nous engageant à interroger l'autre.

[4] La distance du Khi^2 est une pondération de la distance euclidienne. Dans le cas présent cet indice est révélateur du degré d'appartenance d'une forme réduite à une classe. La forme la plus constitutive des énoncés où figure le *je* est la seconde personne du pluriel. La forme la plus représentative de la classe opposée est la variable *De Gaulle* puis la première personne du singulier.

Dans la première classe en revanche, c'est la variable Giscard qui est la plus représentative. Nous avions constaté qu'il était celui chez qui la personnalisation du discours était la plus sensible. Ses messages par ailleurs sont ceux qui laissent la part la plus importante à la thématique des vœux. Ce fait se vérifie, là encore par la méthode des spécificités. Cette thématique des vœux est précisément très représentée dans la première classe, à travers des substantifs : *vœux*, *année*, *soirée,* des adjectifs : *heureux*, *bonne* mais aussi des verbes : *souhait*<, *forme*, *adresser*, *présenter*, *échanger*.
Le rituel et les formules d'adresse sont par ailleurs constitutifs de cette première classe : *Compatriote* +, *cher* +, *metropole* +,. On notera que le lexique est plus intimiste, plus affectif dans cette même classe, la classe 2 recensant un lexique plutôt politique, économique et social: *economie*, *développement*, *organisation*, *conflit*…
Ceci nous amène à nous interroger sur la nature des relations entretenues entre les marques de la première personne et les différentes thématiques des messages : le *je* rejette-t-il dans ces messages la dimension politique au profit d'un lexique plus circonstanciel ou rituel, cette dimension politique serait-elle alors développée autour d'autres référents ?

4. Conclusion

Le croisement des approches et des outils statistiques dont la finalité première n'est pas nécessairement la recherche localisée de phénomènes cooccurentiels révèle des résultats convergents pour les faits les plus saillants. Cette démarche expérimentaliste et comparative a permis de réunir un faisceau sur lequel l'analyse interprétative peut s'appuyer plus fortement, que sur les faits révélés par un outil unique.

Le croisement des perspectives valide par ailleurs la démarche lexicométrique sur un corpus de faible étendue tel que celui des vœux présidentiels. L'application de paramètres et de seuils, plus ou moins intuitifs mais souvent différents, le choix des fenêtres contextuelles, l'élagage ou non de certaines formes de vocabulaire, sont autant de facteurs qui incitent à confronter ces outils avant toute phase interprétative. Spécificités positives et négatives, topographie textuelle (Lexico 3), présentation synthétique (Hyperbase), ou ordonnée (Weblex lexicogrammes simples), réseaux d'affinités et récursivité (Lexicogrammes récursifs), partition binaire et classification ascendante (Alceste), fournissent des perceptions différentes et complémentaires des faits de cooccurrences.

5. Bibliographie

Amossy R. (1999). *Images de soi dans le discours, la construction de l'ethos*. Lausanne, Delachaux et Niestlé.

Heiden S. (2004). *Interface hypertextuelle à un espace de cooccurrences: implémentation dans Weblex*, JADT 2004: 7èmes.

Labbé D. (1990). *Le vocabulaire de François Mitterrand*, Paris, Presses de la Fondation Nationale des Sciences Politiques.

Lafon P. (1984). *Dépouillements et Statistiques en Lexicométrie*. Slatkine-Champion.

Lamalle C., Salem A. (2002). *Types généralisés et topographie textuelle dans l'analyse quantitative des corpus textuels*, JADT 2002 : 6émes

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

Maingueneau D. (2003). « Discours éphémères et non-éphémères : deux gestions de l'ethos ? » In J. Härmä dir., *Le langage des médias: discours éphémères*? Paris, l'Harmattan, p. 67-82.

Martinez W. (2003). *Contribution à une méthodologie de l'analyse des cooccurrences lexicales multiples dans les corpus textuels*, Thèse de troisième cycle, Sorbonne Nouvelle Paris 3, décembre 2003.

Reinert M. (1993). Les « mondes lexicaux » et leur logique. *Langage et société* 66

Reinert M. (1998). Quelques interrogations à propos de l'« objet » d'une analyse de discours de type statistique et de la réponse « Alceste ». *Langage et société* (1998).

Reinert M. (1990). Système Alceste: Une méthodologie d'analyse des données textuelles. JADT 1990.

Tournier M. (2003). *De France à Je. La traversée des emplois Cooccurrences et connexions*. In Des sources du sens, École Normale Supérieure Lettres Sciences Humaines Lyon, Collection Langages.